EARTHALIGN: LEARNING LANGUAGE-ALIGNED REPRESENTATIONS OF EARTH SYSTEM DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Interpreting Earth system conditions is central to understanding environmental risks and informing decisions across weather- and climate-sensitive sectors. Existing multimodal weather systems primarily target task-specific prediction, generation, or question answering, while general-purpose alignment between Earth system data and language remains underexplored due to the limited availability of diverse language supervision grounded in meteorological conditions. To fill this gap, we introduce EarthAlign, a large-scale dataset with broad geographic coverage spanning 2000–2024 that combines 10 expert-curated natural event and disaster data sources with climatology-derived normal-weather samples. Each sample includes temporal and geospatial information and four grounded caption styles aligned with 63-variable ERA5 meteorological data. Building on EarthAlign, we introduce MeteoLIP, a dual-encoder framework for learning aligned meteorology-language representations. MeteoLIP incorporates MetTok, which uses a geolocation- and time-conditioned query to aggregate variable-specific patch embeddings into a compact sequence of meteorology tokens. To comprehensively assess Earth system multimodal representations, we introduce three downstream tasks: cross-modal retrieval, weather-condition classification, and physical variable estimation. Our experiments show that MeteoLIP improves upon the corresponding CLIP and SigLIP baselines across these tasks. Further analyses highlight the importance of meteorological input, temporal and geospatial conditioning, and multi-style supervision for meteorology-language alignment. Thus, EarthAlign and MeteoLIP provide a foundation for learning and evaluating language-aligned representations of Earth system data. We will publicly release the dataset & code.

1 INTRODUCTION

Earth system forecasts and analyses play a critical role in supporting decision-making across sectors such as agriculture, water, energy, and disaster management (Brasseur & Gallardo, 2016; White et al., 2017). Recent data-driven weather models and Earth system foundation models, including FourCastNet (Pathak et al., 2022), Pangu-Weather (Bi et al., 2023), ClimaX (Nguyen et al., 2023), and Aurora (Bodnar et al., 2025), have achieved strong performance in forecasting atmospheric and broader Earth system states. Their representations are primarily learned through objectives defined over physical variables, such as forecasting future atmospheric states (Nguyen et al., 2023; Bodnar et al., 2025) and reconstructing masked meteorological fields (Lessig et al., 2023; Schmude et al., 2024). While these objectives effectively model relationships within numerical Earth system data, they do not explicitly align the learned representations with natural-language descriptions of weather conditions and natural phenomena.

Natural language provides a valuable source of supervision that can align multivariate meteorological data with descriptions of weather conditions and natural phenomena. In natural-image domains, CLIP (Radford et al., 2021) has demonstrated that language supervision can support shared representations for cross-modal retrieval and text-based classification, and such visual representations have subsequently become key components of multimodal models across different domains (Liu et al., 2023; Shabbir et al., 2025). Despite this progress, general-purpose alignment between Earth system data and language remains underexplored, partly due to the lack of diverse language super-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

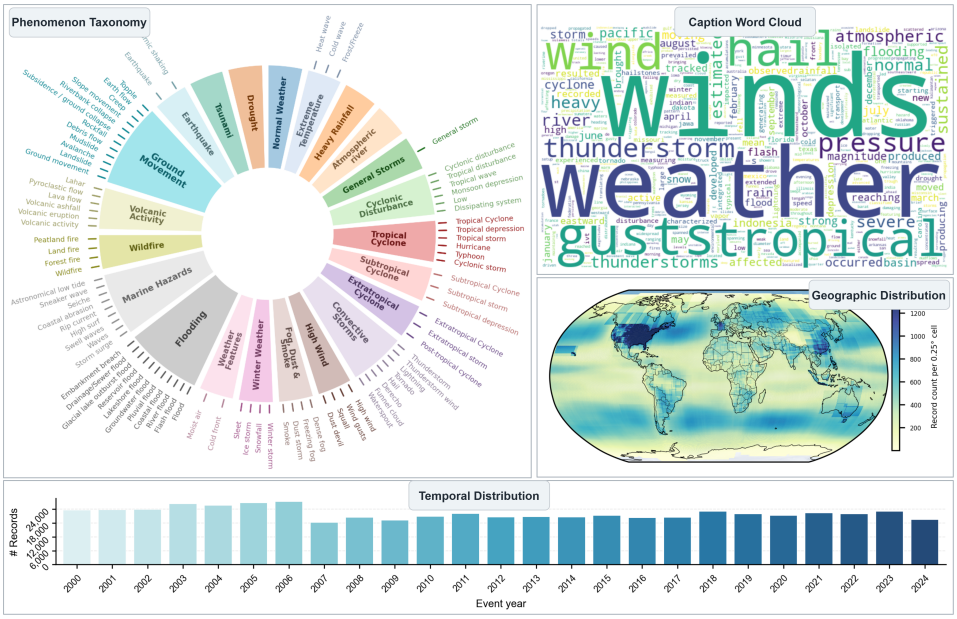


Figure 1: **EarthAlign Overview.** EarthAlign is a vision-language alignment dataset for the Earth system, comprising 553,738 samples that span 2000–2024, with a broad taxonomy of natural phenomena, broad geographic coverage, and diverse caption content to enable multimodal alignment.

vision grounded in meteorological conditions and natural phenomena. Existing multimodal datasets such as WeatherQA (Ma et al., 2024), CLLMate (Li et al., 2025), and MeteorPred (Tang et al., 2026) have enabled progress in severe-weather question answering, event prediction, and warning generation, but are primarily designed around task-specific prediction, generation, or question answering. Learning robust meteorology-language representations requires rich language supervision covering diverse weather conditions and natural phenomena, with descriptions from different perspectives and temporal and geographic context to associate them with the corresponding Earth system states.

To address this, we construct EarthAlign (Figure 1), which brings heterogeneous natural event, disaster, and normal-weather records into an Earth system–language setting. EarthAlign comprises 553,738 samples spanning 2000–2024, with broad geographic coverage and diverse natural phenomena. We compile natural event and disaster records from governmental, institutional, and published sources, and complement them with normal-weather samples identified relative to local climatology. For each record, we generate four metadata-grounded caption styles to provide different semantic perspectives on the underlying Earth system condition. Each record is also paired with a 63-variable ERA5 (Hersbach et al., 2020) snapshot at the corresponding time and spatial extent to enable meteorology-language alignment. To learn semantically aligned representations of meteorological data and language, we introduce MeteoLIP, a dual-encoder framework that maps both modalities into a shared representation space. As meteorological inputs comprise 63 surface and atmospheric variables, MeteoLIP uses MetTok to aggregate variable-specific patch embeddings into compact meteorology tokens conditioned on temporal and geospatial information. During alignment training, MeteoLIP samples one of the four caption styles for each meteorological input, exposing the model to complementary descriptions of the same Earth system condition.

Our contributions can be summarized as follows: **(1)** We introduce EarthAlign, a large-scale Earth system–language dataset with broad geographic coverage, diverse natural phenomena and normal-weather conditions, metadata-grounded multi-style captions, and alignment with multivariate ERA5 data; **(2)** We introduce MeteoLIP for meteorology-language representation learning, and train Meteo-CLIP and Meteo-SigLIP variants that improve over their corresponding CLIP and SigLIP baselines; **(3)** We establish three evaluation tasks spanning cross-modal alignment, semantic understanding, and quantitative inference through retrieval, weather-condition classification, and physical variable estimation. We will make the dataset, code, and model weights publicly available.

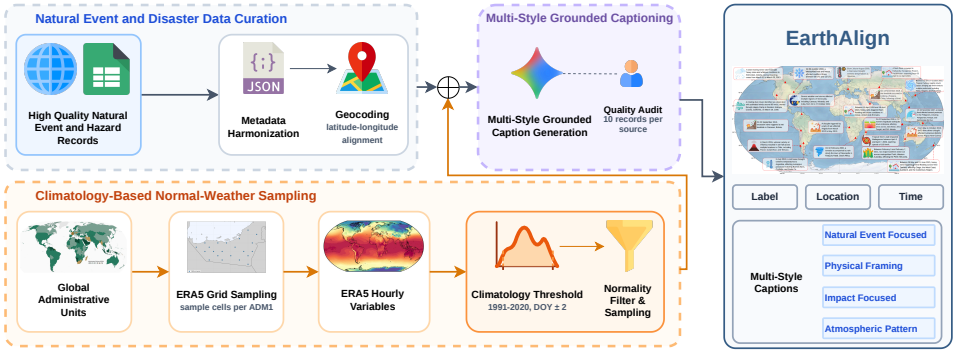


Figure 2: **Data curation pipeline.** EarthAlign combines expert-curated natural-event records with climatology-based normal-weather samples. The resulting records contain standardized spatial and temporal information, labels, and four grounded caption styles for alignment with Earth system data.

2 Related Work

Vision-language alignment. Vision-language pretraining learns shared representations by aligning visual data with rich natural language descriptions. CLIP (Radford et al., 2021) demonstrated the effectiveness of large-scale image-text contrastive learning, while SigLIP (Zhai et al., 2023) and InternVL (Chen et al., 2024) further improved its training objective and model scaling. Beyond natural images, this paradigm has been adapted to specialized domains such as medical imaging (Zhang et al., 2025; Khattak et al., 2024) and remote sensing (Liu et al., 2024; Xiong et al., 2025). Due to their transferable representations, these visual encoders are widely adopted in multimodal language models across domains (Liu et al., 2023; Wang et al., 2025; Sellergren et al., 2026; Soni et al., 2025). Despite these developments, vision-language alignment for Earth system data remains underexplored, primarily due to the lack of large-scale language supervision paired with geolocation and time information. EarthAlign addresses this gap by introducing geocoded natural language captions describing diverse historical weather conditions and natural phenomena from around the world, enabling their alignment with Earth system data, such as gridded meteorological data.

Multimodal weather and climate datasets and benchmarks. As summarized in Table 8, recent work has increasingly connected meteorological data with natural language for weather- and climate-related tasks. WeatherQA (Ma et al., 2024) pairs expert forecast analyses with mesoscale analysis images for severe-weather question answering. ClimateIQA (Chen et al., 2025) generates meteorological heatmaps from four ERA5 variables (Hersbach et al., 2020) to detect three types of severe weather events. Aquilon (Varambally et al., 2025) defines seven question types and constructs samples from ERA5 and an expert-curated event record. CLLMate (Li et al., 2025) curates expert-verified weather and climate events from environmental news and pairs them with four ERA5 variables for event prediction. MeteorPred (Tang et al., 2026) uses meteorological agency warnings as supervision and ERA5 variables as inputs for warning generation in China. While these datasets vary in scope and task formulation, they are primarily designed for instruction following or question answering. In contrast, EarthAlign provides large-scale dataset spanning 2000–2024, with broad geographic coverage and natural phenomena, with four grounded caption styles per record to enable vision-language alignment with Earth system data. We further introduce retrieval, classification, and physical variable estimation tasks to evaluate multimodal representations in this domain.

3 Dataset and Evaluation Tasks

3.1 Dataset Construction

To address the limitations of existing multimodal climate datasets, we present EarthAlign, a large-scale dataset designed to facilitate alignment of Earth-system data with language. We build the dataset based on two types of sources: records of natural events and disasters, and climatological data for normal weather samples. EarthAlign comprises 553,738 samples with broad geographic

coverage spanning 2000–2024 and encompasses a diverse set of keywords describing a wide range of natural phenomena (Fig. 1). Each sample contains temporal information, a latitude-longitude bounding box, and four captions in different styles. Our data curation pipeline is shown in Fig. 2.

Natural event and disaster records. Natural events and disasters constitute the majority of samples in EarthAlign, providing large-scale and diverse observations of natural phenomena. These records were collected from 10 expert-curated sources (Appendix Table 9): 1) EM-DAT (Delforge et al., 2025), 2) NOAA Storm Event Database (NOAA NCEI), 3) IBTrACS (Knapp et al., 2010; Gahtan et al., 2024), 4) COOLR (Juang et al., 2019), 5) PIKART (Vallejo-Bernal et al., 2025), 6) AusDE (Australian Government, Attorney-General's Department, 2023), 7) MexDE (Centro Nacional de Prevención de Desastres (CENAPRED), 2023), 8) CanDE (Public Safety Canada, 2026), 9) IndDE (Badan Nasional Penanggulangan Bencana, 2026), and 10) FraDE (Ministère de la Transition écologique, 2026). Since these sources provide different types of information in different formats, we apply source-specific processing steps (Appendix B.1). In general, we begin by manually filtering the records to retain only natural phenomena. We then extract all available metadata to support the grounded generation of detailed descriptions. For sources that provide only textual location names, we apply geocoding to obtain coordinates and latitude-longitude bounding boxes. After processing, we obtain 489,185 records with standardized phenomenon categories, associated metadata, as well as temporal and geospatial information.

Climatology-based normal-weather sampling. To complement the natural event and disaster records, we identify normal-weather conditions relative to local climatology. We construct local climatologies from ERA5 (Hersbach et al., 2020) hourly variables over the 1991–2020 WMO reference period (WMO, 2017). Adapting the centered 5-day window proposed in Zhang et al. (2005), we compute each variable-specific threshold for each day of year and hour using a centered $\pm$2-day window. For scalability, we sample 20% of the ERA5 grid cells within each first-level administrative unit (ADM1), based on administrative boundaries from geoBoundaries (Runfola et al., 2020), with at least 10 cells whenever possible, and retrieve hourly variables from 1991 to 2025. Depending on the variable, we define normality by P05–P95, $\leq$P90, or $\geq$P10 thresholds. A grid-cell hour is considered normal only if all variables satisfy their respective thresholds, and an ADM1-hour is retained when at least 80% of its sampled cells are normal. Using this conservative filter, we retain 1.6% and 11% of candidate hours in California and Abu Dhabi, respectively. We then sample eligible normal-weather hours globally over 2000–2024 with broad country coverage and area-proportional sampling, yielding 64,553 samples. Further details on the meteorological variables and their variable-specific normality criteria are provided in Appendix B.2.

Grounded caption generation. From the metadata, we generate four stylistically different captions (i.e. *natural-event-focused*, *physical-framing*, *impact-focused*, and *atmospheric-pattern*) to provide complementary semantic perspectives of each record and support alignment with Earth system data. We use Gemini 3.1 Pro with a prompt that requires every caption to be directly supported by the input metadata and excludes unsupported causes, impacts, forecasts, and descriptive modifiers. To assess caption grounding and adherence to the target styles, a trained graduate annotator audited a stratified sample of 440 captions across all data sources and caption styles. The overall pass rate was 95.68% (421/440), with style-specific pass rates of 100%, 98%, 89%, and 95%. We provide the generation prompt in Appendix C.1, and the audit rubric and error analysis in Appendix C.2.

Earth system data alignment. Earth system data are heterogeneous across data types and modalities, variables, and spatial and time resolutions. ERA5 (Hersbach et al., 2020) provides globally gridded hourly data at 0.25° resolution and is widely used in recent Earth system models and multimodal meteorological benchmarks (Rasp et al., 2024; Bodnar et al., 2025; Li et al., 2025; Tang et al., 2026). Motivated by this, we adopt ERA5 as a common baseline while preserving compatibility with other Earth system products. For each record, we extract a single multivariate ERA5 patch at the corresponding timestamp and spatial extent. Each patch contains 28 single-level variables and 7 pressure-level variables at 5 pressure levels, resulting in 63 input variables as listed in Appendix D.

3.2 Evaluation Tasks

To comprehensively evaluate Earth system multimodal representations, we devise three downstream tasks assessing their cross-modal alignment, semantic understanding, and quantitative inference capability: **(1) Retrieval** evaluates the alignment between Earth system (e.g. ERA5) representa-

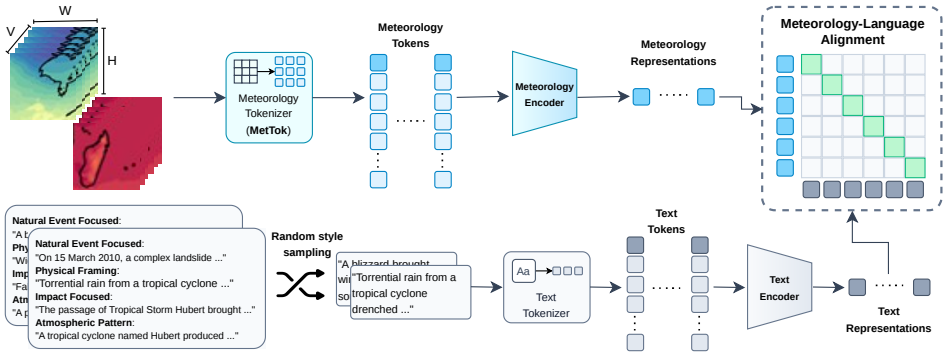


Figure 3: **MeteoLIP Overview.** MeteoLIP consists of transformer-based meteorological and text encoders. MetTok converts multivariate meteorological input into a compact token sequence. During training, we randomly sample a caption style for each meteorological input and optimize a contrastive objective to align meteorological and textual representations in a shared embedding space.

tions and their corresponding textual descriptions using bidirectional text-to-meteorology (T2M) and meteorology-to-text (M2T) retrieval; **(2) Classification** evaluates whether Earth system representations can capture category-level semantics of meteorological conditions. Given an Earth system input (e.g. ERA5), the model predicts one of 11 weather conditions: normal weather, drought, heat wave, cold wave, frost, atmospheric river, cyclone, convective storm, high wind, fog, and winter weather; and **(3) Physical variable estimation** evaluates quantitative inference by estimating cyclone maximum sustained wind speed and minimum central pressure from an Earth system input.

4 METEOLIP: ALIGNING METEOROLOGICAL DATA WITH LANGUAGE

EarthAlign comprises paired ERA5 data and complementary textual descriptions. To learn aligned representations of these descriptions and their meteorological conditions, we introduce MeteoLIP, a dual-encoder framework for aligning meteorological data with natural language. As illustrated in Fig. 3, MeteoLIP utilizes separate transformer-based meteorology and text encoders and maps their outputs into a shared representation space. To process the 63-variable ERA5 input, MeteoLIP uses MetTok, which provides a mechanism for incorporating geospatial and temporal information while producing a compact token sequence whose length depends on the spatial resolution. Through this alignment, MeteoLIP learns cross-modal representations that enable a range of downstream tasks.

4.1 MODEL ARCHITECTURE AND TRAINING OBJECTIVE

Dual-encoder formulation. Let $\mathbf{x}_T$ denote a textual input and $\mathbf{x}_M \in \mathbb{R}^{V\times H\times W}$ denote meteorological data, where V is the number of variables and $H \times W$ are the spatial dimensions. Each meteorological input is supplemented with temporal information $\mathbf{t} \in \mathbb{R}^3$, corresponding to year, day of year, and hour of day, and latitude–longitude geolocation information $\boldsymbol{\ell} \in \mathbb{R}^{2\times H\times W}$. MeteoLIP employs dual encoders to map the two modalities into a shared D-dimensional representation space, with $\mathbf{z}_M = \mathcal{F}_M(\mathcal{T}_M(\mathbf{x}_M, \mathbf{t}, \boldsymbol{\ell}))$ and $\mathbf{z}_T = \mathcal{F}_T(\mathcal{T}_T(\mathbf{x}_T))$, where $\mathbf{z}_M, \mathbf{z}_T \in \mathbb{R}^D$ denote the meteorological and textual representations, respectively. $\mathcal{T}_M$ and $\mathcal{T}_T$ denote the meteorology and text tokenizers, while $\mathcal{F}_M$ and $\mathcal{F}_T$ denote the transformer-based meteorology and text encoders.

Meteorology tokenization (MetTok). Unlike natural images, meteorological inputs contain many physical variables. Let ps be the patch size and $P = HW/ps^2$ be the number of patches. Independently patchifying each variable would produce VP tokens, which would increase the self-attention complexity in the meteorology encoder from $\mathcal{O}(P^2D)$ to $\mathcal{O}(V^2P^2D)$. MetTok instead aggregates the variables within each patch, thereby retaining a sequence of P tokens. Specifically, we independently patchify each variable in $\mathbf{x}_M$ using a variable-specific linear projection and subsequently add a learnable variable-identity embedding, producing variable-patch embeddings $\mathbf{E}_M \in \mathbb{R}^{V\times P\times d}$, where $d = D/8$ is the MetTok embedding dimension. We encode the temporal information $\mathbf{t}$ by linearly normalizing the year and applying sine-cosine transformations to the

day of year and hour of day, resulting in $\widetilde{\mathbf{t}} \in \mathbb{R}^5$, which is spatially expanded to $\mathbb{R}^{5\times H\times W}$. Similarly, we apply sine-cosine transformations to latitude and longitude, resulting in $\widetilde{\boldsymbol{\ell}} \in \mathbb{R}^{4\times H\times W}$. To obtain temporal-geospatial patch embeddings $\mathbf{E}_{t\ell} \in \mathbb{R}^{P\times d}$, we concatenate the spatially expanded $\widetilde{\mathbf{t}}$ and $\widetilde{\boldsymbol{\ell}}$ along the channel dimension and jointly patchify them using a linear–GELU–linear projection, allowing temporal and geospatial features to interact nonlinearly. Inspired by Bodnar et al. (2025); Jaegle et al. (2021), we introduce a learnable query embedding $\mathbf{e}_q \in \mathbb{R}^d$ to aggregate the variable embeddings in $\mathbf{E}_M$ through cross-attention. Since meteorological conditions depend on geographic location and temporal context, we condition the query on $\mathbf{E}_{t\ell}$ to obtain time- and location-aware queries $\mathbf{E}_q = [\mathbf{e}_q + \alpha\,\mathbf{e}_{t\ell,1}; \ldots; \mathbf{e}_q + \alpha\,\mathbf{e}_{t\ell,P}] \in \mathbb{R}^{P\times d}$, where $\alpha \in \mathbb{R}$ is a learnable scalar. We then apply cross-attention independently within each patch, using the time- and location-aware query as the query and the corresponding variable embeddings as keys and values, yielding $\mathbf{e}_{o,p} = \mathbf{e}_{q,p} + \mathrm{CrossAttn}(\mathbf{e}_{q,p}, \mathbf{e}_{M,p}, \mathbf{e}_{M,p})$ for $p = 1, \ldots, P$, where $\mathbf{e}_{q,p} \in \mathbb{R}^d$ and $\mathbf{e}_{M,p} \in \mathbb{R}^{V\times d}$ are the query and variable embeddings for patch p, respectively. We then project each aggregated representation to the meteorology encoder dimension using a linear layer, $\widehat{\mathbf{e}}_{o,p} = \mathbf{W}_o\mathbf{e}_{o,p} + \mathbf{b}_o \in \mathbb{R}^D$. The resulting meteorology tokens $[\widehat{\mathbf{e}}_{o,1}; \ldots; \widehat{\mathbf{e}}_{o,P}]$ are prepended with a class token and passed to the meteorology encoder, thereby avoiding the additional quadratic self-attention cost induced by variable-wise tokenization.

Learning objective. During training, we sample one caption style for each meteorological input in a minibatch to expose the model to diverse linguistic descriptions from different perspectives and improve robustness. Given a minibatch of B paired meteorological and textual inputs, we compute their pairwise cosine similarities as $s_{ij} = \mathrm{sim}\left(\mathbf{z}_M^{(i)}, \mathbf{z}_T^{(j)}\right)$. Following Radford et al. (2021); Zhai et al. (2023), we optimize a cross-modal contrastive objective that increases s_{ii} for paired meteorological and textual representations while decreasing s_{ij} for unmatched pairs $i \neq j$.

4.2 Task-Specific Inference

Cross-modal retrieval. Given a retrieval set of N paired samples $\{(\mathbf{x}_M^{(i)}, \mathbf{x}_T^{(i)})\}_{i=1}^N$, MeteoLIP encodes each meteorological and textual input into shared representations $\mathbf{z}_M^{(i)}$ and $\mathbf{z}_T^{(i)}$, respectively. We then compute the cosine similarity between the two modalities as $s_{ij} = \mathrm{sim}\left(\mathbf{z}_M^{(i)}, \mathbf{z}_T^{(j)}\right)$. For T2M retrieval, we rank meteorological candidates by their similarity to a textual query, and reverse the direction for M2T retrieval.

Classification. We construct K textual prompts $\{p_k(c)\}_{k=1}^K$ for each class $c \in \mathcal{C}$ and apply prompt ensembling to obtain a class prototype $\overline{\mathbf{z}}_T^{(c)} = \frac{1}{K}\sum_{k=1}^K \mathcal{F}_T(\mathcal{T}_T(p_k(c)))$. Given a meteorological representation $\mathbf{z}_M$, we predict its class as $\hat{c} = \arg\max_{c\in\mathcal{C}} \mathrm{sim}\left(\mathbf{z}_M, \overline{\mathbf{z}}_T^{(c)}\right)$.

Physical variable estimation. Inspired by Christensen et al. (2024); Maani et al. (2026), we estimate cyclone maximum sustained wind speed and minimum central pressure from $\mathbf{x}_M$ using value-conditioned text prototypes. We begin by constructing $K = 80$ textual templates $\{p_k(v)\}_{k=1}^K$, such as *"a cyclone with sustained wind speeds of v knots"* for each candidate value $v \in \mathcal{V}$, where $\mathcal{V} = \{v_{\min}, v_{\min}+1, \ldots, v_{\max}\}$, and obtain a value prototype $\overline{\mathbf{z}}_T^{(v)} = \frac{1}{K}\sum_{k=1}^K \mathcal{F}_T(\mathcal{T}_T(p_k(v)))$. We then compute the similarity score for each v as $s_v(\mathbf{z}_M) = \mathrm{sim}\left(\mathbf{z}_M, \overline{\mathbf{z}}_T^{(v)}\right)$. Let $\mathcal{S}_\rho(\mathbf{x}_M)$ denote the top-ρ candidate values by score, where ρ is the nearest odd integer to $0.05|\mathcal{V}|$. For robust estimation, we take the median of $\mathcal{S}_\rho(\mathbf{x}_M)$ as the final estimate.

5 Experiments and Results

5.1 Experimental Setup

Meteorological data processing. We use the 63 ERA5 meteorological variables listed in Appendix D. We use the specified time for records with an exact timestamp, while we sample a timestamp within the reported interval for records spanning a time range. We standardize each meteorological input to 56×56 grid cells. For samples spanning less than $14^\circ \times 14^\circ$, we expand the bounding box to this extent. For larger samples, we construct a square region using the larger

latitudinal or longitudinal extent and linearly interpolate it to 56×56 to avoid introducing extremes. We also normalize each variable using its mean and standard deviation computed across samples.

Dataset split. We employ a year-based train-test split to assess whether the learned representations remain effective across years, using 2000–2010 and 2014–2024 for training and withholding 2011–2013 for evaluation on unseen intermediate years. This results in 490,395 training samples, with 63,343, 48,645, 7,353, and 8,114 test samples for retrieval, classification, cyclone maximum sustained wind speed estimation, and cyclone minimum central pressure estimation, respectively.

MeteoLIP and training configuration. We investigate two MeteoLIP variants based on the CLIP (Radford et al., 2021) and SigLIP (Zhai et al., 2023) architectures and their respective learning objectives, denoted as Meteo-CLIP and Meteo-SigLIP. We initialize the encoders from the corresponding OpenAI CLIP and SigLIP WebLI checkpoints and use a text context length of 77 tokens. For the meteorology encoder, we use a patch size of 4×4, which results in $(56/4)^2 = 196$ spatial patches. Unless otherwise specified, we train all models for 30 epochs on 4×A100 (40GB) GPUs with a batch size of 256 per GPU, using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2}. We also apply a cosine learning-rate schedule with 2,000 warmup steps. We report the mean and standard deviation over three training runs using the final epoch checkpoint.

Baselines. For the CLIP and SigLIP baselines, following the multichannel patchification used in FourCastNet (Pathak et al., 2022), we concatenate all meteorological variables along the channel dimension and jointly project each spatial patch using a linear layer, producing one token per patch. To follow the original image-text pair formulation of CLIP and SigLIP, we train the baselines using only the *natural-event-focused* caption, which identifies the natural phenomenon together with its temporal and geospatial context when available.

5.2 Main Results

Retrieval. To evaluate meteorology-language alignment, we first assess bidirectional cross-modal retrieval. As shown in Table 1a, Meteo-CLIP and Meteo-SigLIP improve over their corresponding CLIP and SigLIP baselines across all retrieval metrics. Meteo-CLIP improves M2T and T2M R@1 from 20.25 and 20.03 to 58.79 and 60.10 respectively, while Meteo-SigLIP achieves the best performance with 66.07 and 66.79. Similar improvements are observed at R@5 and R@10, demonstrating strong meteorology-language alignment. Since the default baselines use a single caption style, we further examine the effects of multi-style supervision and MetTok conditioning in Section 5.3.

Classification. Table 1b evaluates category-level meteorological representations under two protocols: *text-prototype inference*, which applies the text-based inference described in Section 4.2 without task-specific post-training, and *linear probing*, which trains a linear prediction head on top of the frozen meteorology encoder. Under text-prototype inference, Meteo-CLIP improves F1 from 30.15 to 47.91 and Top-1 accuracy from 33.69 to 57.69, while Meteo-SigLIP improves from 21.56 to 50.19 and Top-1 accuracy from 22.13 to 55.59. The improvements can also be observed under linear probing, where Meteo-SigLIP achieves the highest F1 of 80.75 and Top-1 accuracy of 92.35. These results indicate that MeteoLIP improves both language-based inference and the category-level semantic information encoded in the frozen meteorology representations. We provide detailed category-wise results in Appendix F.1, where drought, heat wave, and cold wave are identified as the most challenging categories.

Physical variable estimation. As provided in Table 1c, under text-prototype inference, Meteo-CLIP improves R^2 for cyclone maximum sustained wind speed from 5.29 to 62.27 and for minimum central pressure from 26.81 to 67.64, while Meteo-SigLIP improves them from -11.59 to 60.17 and from 45.12 to 73.39, respectively. Under linear probing, Meteo-SigLIP achieves the best performance for both variables, with R^2 values of 71.85 for maximum sustained wind speed and 77.36 for minimum central pressure. We further compare against physical baselines derived directly from ERA5 in Appendix F.2. These results demonstrate that the learned meteorology-language representations are capable of capturing physically meaningful quantitative information.

5.3 Ablation Studies

Caption supervision. We study the effects of caption style for retrieval and multi-style supervision. Table 2 shows that the natural-event-focused caption results in the highest retrieval performance,

Table 1: **Comparison of CLIP and SigLIP baselines with their MeteoLIP counterparts.** (a) Cross-modal retrieval. (b) Classification under *text-prototype inference* and *linear probing*. (c) Physical variable estimation of cyclone maximum sustained wind speed and minimum central pressure under the same two protocols. **Bold** and underline indicate the best and second-best results.

(a)

	Meteorology-to-Text			Text-to-Meteorology		
Model	**R@1** ↑	**R@5** ↑	**R@10** ↑	**R@1** ↑	**R@5** ↑	**R@10** ↑
CLIP	20.25 ± 0.13	45.67 ± 0.19	58.25 ± 0.27	20.03 ± 0.14	45.05 ± 0.39	57.43 ± 0.47
Meteo-CLIP	58.79 ± 1.43	88.13 ± 0.60	94.09 ± 0.32	60.10 ± 0.23	87.68 ± 0.13	93.50 ± 0.12
SigLIP	32.06 ± 0.56	61.11 ± 0.75	72.07 ± 0.69	30.23 ± 0.22	58.12 ± 0.25	68.98 ± 0.25
Meteo-SigLIP	$\mathbf{66.07 \pm 0.64}$	$\mathbf{91.77 \pm 0.29}$	$\mathbf{96.14 \pm 0.15}$	$\mathbf{66.79 \pm 0.50}$	$\mathbf{91.27 \pm 0.31}$	$\mathbf{95.62 \pm 0.18}$

(b)

	Text-Prototype Inference			Linear Probing		
Model	**F1** ↑	**Top-1 Acc.** ↑	**Top-3 Acc.** ↑	**F1** ↑	**Top-1 Acc.** ↑	**Top-3 Acc.** ↑
CLIP	30.15 ± 2.83	33.69 ± 3.10	66.95 ± 2.53	75.04 ± 0.20	90.61 ± 0.05	98.62 ± 0.06
Meteo-CLIP	47.91 ± 2.10	$\mathbf{57.69 \pm 1.43}$	$\mathbf{85.98 \pm 0.97}$	79.98 ± 0.26	92.16 ± 0.11	$\mathbf{98.82 \pm 0.05}$
SigLIP	21.56 ± 1.48	22.13 ± 2.95	50.15 ± 3.52	76.30 ± 0.54	91.06 ± 0.08	98.61 ± 0.05
Meteo-SigLIP	$\mathbf{50.19 \pm 2.95}$	55.59 ± 5.31	82.29 ± 3.63	$\mathbf{80.75 \pm 0.14}$	$\mathbf{92.35 \pm 0.04}$	98.78 ± 0.04

(c)

		Text-Prototype Inference			Linear Probing		
Variable	**Model**	R^2 ↑	**MAE**↓	**RMSE**↓	R^2 ↑	**MAE**↓	**RMSE**↓
Wind	CLIP	5.29 ± 19.85	14.62 ± 1.38	21.38 ± 2.24	68.09 ± 1.04	9.42 ± 0.17	12.46 ± 0.20
	Meteo-CLIP	$\mathbf{62.27 \pm 4.75}$	9.35 ± 0.42	$\mathbf{13.53 \pm 0.84}$	71.34 ± 0.51	8.65 ± 0.08	11.80 ± 0.10
	SigLIP	-11.59 ± 16.49	14.46 ± 0.88	23.25 ± 1.72	70.52 ± 0.48	8.91 ± 0.11	11.97 ± 0.10
	Meteo-SigLIP	60.17 ± 1.73	$\mathbf{9.32 \pm 0.21}$	13.91 ± 0.30	$\mathbf{71.85 \pm 0.37}$	$\mathbf{8.50 \pm 0.04}$	$\mathbf{11.70 \pm 0.08}$
Pressure	CLIP	26.81 ± 14.46	9.70 ± 0.80	16.54 ± 1.60	74.08 ± 0.82	7.04 ± 0.11	9.87 ± 0.16
	Meteo-CLIP	67.64 ± 4.23	6.37 ± 0.22	11.01 ± 0.72	76.36 ± 0.46	6.41 ± 0.09	9.43 ± 0.09
	SigLIP	45.12 ± 3.00	7.96 ± 0.20	14.36 ± 0.39	75.98 ± 0.35	6.60 ± 0.06	9.50 ± 0.07
	Meteo-SigLIP	$\mathbf{73.39 \pm 1.61}$	$\mathbf{5.96 \pm 0.09}$	$\mathbf{10.00 \pm 0.30}$	$\mathbf{77.36 \pm 0.45}$	$\mathbf{6.16 \pm 0.09}$	$\mathbf{9.23 \pm 0.09}$

achieving 58.79 M2T and 60.12 T2M R@1. In comparison, the other caption styles yield substantially lower retrieval performance, suggesting that explicitly describing the natural phenomenon together with its temporal and geospatial context provides more instance-discriminative information for retrieval. Table 3 further demonstrates the benefit of multi-style supervision during training. Increasing from one to four caption styles improves both model architectures: CLIP increases from 20.25/20.03 to 31.64/31.10 M2T/T2M R@1, while Meteo-CLIP improves from 34.24/55.55 to 58.79/60.10. Moreover, Meteo-CLIP consistently outperforms CLIP under both one-style and four-style training, demonstrating its effectiveness across different levels of caption-style supervision.

Model initialization. As shown in Table 4, initializing from the OpenAI CLIP checkpoint improves M2T R@1 from 34.46 to 58.79 and T2M R@1 from 41.43 to 60.10 compared with training from scratch. This suggests that, despite the modality gap between meteorological data and RGB images, pretrained visual representations provide a useful initialization for learning spatial structure in meteorological fields. One possible explanation is that some meteorological variables can reflect coherent spatial patterns associated with natural phenomena such as cyclones and atmospheric rivers, which can benefit from the spatial pattern recognition capabilities of pretrained image encoders.

MetTok conditioning. Table 5 highlights the contribution of meteorological input and MetTok's conditioning components across retrieval, weather condition classification, and maximum sustained wind estimation. For retrieval, variable-identity embedding and geolocation provide relatively modest improvements, whereas temporal conditioning contributes substantially. Notably, using geolocation and time alone without meteorological input achieves 50.41/58.49 M2T/T2M R@1, showing that temporal and geographic metadata provide strong cues for matching natural event captions. However, this metadata-only configuration achieves substantially lower performance on weather condition classification and fails to provide useful sustained wind estimates, achieving 32.64 F1 and an R^2 of -155.14, compared with 47.91 F1 and 62.27 R^2 for the full MetTok configuration. To further examine the role of metadata in retrieval, Table 6 compares the full MetTok configuration with metadata only using original and time-removed captions under different candidate pools. With the original captions, the relative gap between the two configurations is 14.3% under both candidate pools. When temporal information is removed from the captions, the gap increases substantially: the full configuration achieves 15.41 M2T R@1 compared with 6.57 using metadata only over all test samples, and 38.65 compared with 23.26 when candidates are restricted to the same calendar month.

Table 2: Meteo-CLIP retrieval performance by caption style.

Caption Style	M2T R@1↑	T2M R@1↑
Natural Event Focused	**58.79 ± 1.44**	**60.12 ± 0.25**
Physical Framing	21.18 ± 0.18	21.67 ± 0.08
Impact Focused	19.09 ± 0.17	18.55 ± 0.08
Atmospheric Pattern	22.24 ± 0.41	21.52 ± 0.08

Table 3: Ablation on multi-style caption supervision during training. #: # Styles.

Model	#	M2T R@1↑	T2M R@1↑
CLIP	1	20.25 ± 0.13	20.03 ± 0.14
	4	**31.64 ± 0.24**	**31.10 ± 0.05**
Meteo-CLIP	1	34.24 ± 8.52	55.55 ± 4.57
	4	**58.79 ± 1.43**	**60.10 ± 0.23**

Table 4: Ablation on model initialization for Meteo-CLIP training.

Init.	M2T R@1↑	T2M R@1↑
Scratch	34.46 ± 1.78	41.43 ± 2.21
OpenAI	**58.79 ± 1.43**	**60.10 ± 0.23**

Table 5: Ablation of MetTok components. We train Meteo-CLIP under different MetTok configurations to assess the contribution of meteorological input, variable-identity embedding, geolocation, and temporal conditioning. **Bold** and underlined values indicate the best and second-best results, respectively.

Input & Encoding				Retrieval		Weather Condition		Sustained Wind	
Met.	Var. ID	Geo.	Time	M2T R@1↑	T2M R@1↑	F1↑	Acc.↑	R^2↑	MAE↓
✓	–	–	–	32.02 ± 0.33	31.74 ± 0.19	47.04 ± 3.67	56.01 ± 6.49	52.66 ± 5.20	10.33 ± 0.40
✓	✓	–	–	32.44 ± 0.34	32.71 ± 0.21	**48.17 ± 1.56**	57.23 ± 2.32	46.74 ± 13.37	10.44 ± 0.67
–	–	✓	✓	50.41 ± 2.43	58.49 ± 1.22	32.64 ± 4.30	45.12 ± 11.06	−155.14 ± 34.45	25.75 ± 1.93
✓	✓	✓	–	33.21 ± 0.69	32.83 ± 0.44	47.25 ± 1.41	54.93 ± 3.20	**62.51 ± 3.87**	9.43 ± 0.15
✓	✓	–	✓	57.30 ± 1.65	59.17 ± 0.44	41.27 ± 1.19	47.47 ± 4.88	59.82 ± 4.56	9.79 ± 0.47
✓	✓	✓	✓	**58.79 ± 1.43**	**60.10 ± 0.23**	47.91 ± 2.10	**57.69 ± 1.43**	62.27 ± 4.75	**9.35 ± 0.42**

Table 6: Retrieval cue analysis. M2T R@1 of Full MetTok and metadata only using original and time-removed captions under all test and same month candidate pools.

Candidate Pool	Caption	Full MetTok	Metadata only	Rel. Gap
All test samples	Original	58.79 ± 1.43	50.41 ± 2.43	14.3%
	Time removed	15.41 ± 0.02	6.57 ± 0.18	57.4%
Same month	Original	61.84 ± 2.48	53.02 ± 3.41	14.3%
	Time removed	38.65 ± 4.06	23.26 ± 2.99	39.8%

Qualitative examples in Figure 7 further show that the full configuration retrieves meteorologically related events, whereas using metadata alone can retrieve unrelated phenomena. These results and analysis highlight the importance of meteorological input for semantic and physical inference and show that it also contributes substantially to retrieval when direct temporal cues are removed.

Training data scaling. We study the effect of training data scale using 10%, 25%, 50%, and 100% of the training set. Figure 4 demonstrates that performance generally improves across all tasks with more training data, with the full training set achieving the best performance. The improvements are particularly notable for retrieval and maximum sustained wind speed estimation, while classification and minimum central pressure estimation also benefit from increased data scale. These results suggest that the model can further benefit from additional meteorology-language supervision.

Source-held-out generalization. We study source-level generalization by training Meteo-CLIP while simultaneously removing AusDE, CanDE, FraDE, IndDE, and MexDE from training, and then evaluating retrieval on the test sets of the five held-out record sources. Table 7 shows that the decrease in retrieval performance varies substantially across sources, which may reflect differences in source-specific reporting criteria, local terminology, and spatial granularity. These results highlight the importance of local record sources for meteorology-language alignment, while showing that cross-source generalization remains challenging.

Table 7: Source-held-out M2T R1; mean over 3 runs.

	Training Data	
Test Data	Full	Held-Out
AusDE *(n=27)*	80.2	72.8 (-7.4)
CanDE *(n=47)*	64.5	63.1 (-1.4)
FraDE *(n=966)*	24.0	10.6 (-13.4)
IndDE *(n=3,835)*	69.6	13.1 (-56.5)
MexDE *(n=336)*	48.1	37.5 (-10.6)

6 Conclusion

We present EarthAlign, a large-scale Earth system–language dataset that pairs meteorological data with multi-style natural language descriptions of diverse natural phenomena and normal-weather conditions. Alongside EarthAlign, we develop MeteoLIP to learn language-aligned representations of high-dimensional meteorological data while incorporating variable-specific, temporal, and geospatial information. Across cross-modal retrieval, weather-condition classification, and physical variable estimation, MeteoLIP improves over the corresponding CLIP and SigLIP baselines. Our analyses further demonstrate the importance of meteorological input, temporal and geospatial conditioning, multi-style language supervision, and diverse training data for effective meteorology-language alignment. We believe EarthAlign can provide a useful foundation for advancing language-aligned multimodal representation learning of Earth system data.

AI USE STATEMENT

In this work, we used generative AI tools to generate captions from available metadata, assist with coding for experiments and visualizations, help identify record data sources and relevant literature, assist with translation when working with source-specific records in Mexico, France, and Indonesia, and provide writing assistance to improve clarity and readability. We have reviewed the AI-assisted work. For example, a subset of the generated captions was manually audited for consistency with the underlying metadata across the record data sources; AI-assisted code was manually inspected and executed; identified record data sources and literature were independently verified; and AI-assisted writing refinements were manually reviewed and further revised. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The preprocessed data, source code, and model weights will be made publicly available. The distribution of EarthAlign is reported in Figure 1 and Appendix E, with representative examples illustrated in Figure 5. The overall dataset construction is described in Section 3.1, while Appendices B, C, and D provide further details on source-specific record processing, grounded multi-style caption generation, and ERA5 meteorological data alignment and preprocessing, respectively. MeteoLIP model configurations and training hyperparameters are provided in Section 5.1. We report experimental results as mean and standard deviation across independent runs to account for variation across training runs.

REFERENCES

Australian Government, Attorney-General's Department. Disaster events with category impact and location. data.gov.au, 2023. URL https://data.gov.au/data/dataset/disaster-events-with-category-impact-and-location. Accessed May 2026.

Badan Nasional Penanggulangan Bencana. Data informasi bencana indonesia (DIBI). Badan Nasional Penanggulangan Bencana, 2026. URL https://dibi.bnpb.go.id/superset/dashboard/2/. Accessed May 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the earth system. *Nature*, 641(8065): 1180–1187, May 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

Guy P. Brasseur and Laura Gallardo. Climate services: Lessons learned and future prospects. *Earth's Future*, 4(3):79–89, 2016. doi: https://doi.org/10.1002/2015EF000338. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015EF000338.

Centro Nacional de Prevención de Desastres (CENAPRED). Declaratorias sobre emergencia, desastre y contingencia climatológica a nivel municipal entre 2000 y 2023. Datos Abiertos de México,

Gobierno de México, 2023. URL https://historico.datos.gob.mx/busca/dataset/declaratorias-sobre-emergencia-desastre-y-contingencia-climatologica/resource/946ad8a6-abe4-4ba8-b96b-7a45bfe687ea. Accessed May 2026.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Wei Chen, Bing Zhu, Junwei Liang, and Zixuan Yuan. Climateiqa: A new dataset and benchmark to advance vision-language models in meteorology anomalies analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5322–5333, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737406. URL https://doi.org/10.1145/3711896.3737406.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, May 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02959-y. URL https://doi.org/10.1038/s41591-024-02959-y.

Damien Delforge, Valentin Wathelet, Regina Below, Cinzia Lanfredi Sofia, Margo Tonnelier, Joris A.F. van Loenhout, and Niko Speybroeck. Em-dat: the emergency events database. *International Journal of Disaster Risk Reduction*, 124:105509, 2025. ISSN 2212-4209. doi: https://doi.org/10.1016/j.ijdrr.2025.105509. URL https://www.sciencedirect.com/science/article/pii/S2212420925003334.

J. Gahtan, K. R. Knapp, C. J. Schreck, H. J. Diamond, J. P. Kossin, and M. C. Kruk. International best track archive for climate stewardship (ibtracs) project, version 4r01, 2024. URL https://www.ncei.noaa.gov/products/international-best-track-archive. since1980; accessed March 2026.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision (ECCV)*, 2024.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

Caroline S. Juang, Thomas A. Stanley, and Dalia B. Kirschbaum. Using citizen science to expand the global map of landslides: Introducing the cooperative open online landslide repository (coolr). *PLOS ONE*, 14(7):1–28, 07 2019. doi: 10.1371/journal.pone.0218657. URL https://doi.org/10.1371/journal.pone.0218657.

Muhammad Uzair Khattak, Shahina Kunhimon, Muzammal Naseer, Salman Khan, and Fahad Shahbaz Khan. Unimed-clip: Towards a unified image-text pretraining paradigm for diverse medical imaging modalities, 2024. URL https://arxiv.org/abs/2412.10372.

Kenneth R. Knapp, Michael C. Kruk, David H. Levinson, Howard J. Diamond, and Charles J. Neumann. The international best track archive for climate stewardship (ibtracs): Unifying tropical cyclone best track data. *Bulletin of the American Meteorological Society*, 91(3):363–376, 2010. doi: 10.1175/2009BAMS2755.1.

Christian Lessig, Ilaria Luise, Bing Gong, Michael Langguth, Scarlet Stadtler, and Martin Schultz. Atmorep: A stochastic model of atmosphere dynamics using large scale representation learning, 2023. URL https://arxiv.org/abs/2308.13280.

Haobo Li, Zhaowei Wang, Jiachen Wang, Yueya Wang, Alexis Kai Hon Lau, and Huamin Qu. CLLMate: A multimodal benchmark for weather and climate events forecasting. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17536–17562, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.886. URL https://aclanthology.org/2025.emnlp-main.886/.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. doi: 10.1109/TGRS.2024.3390838. URL https://doi.org/10.1109/TGRS.2024.3390838.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather?, 2024. URL https://arxiv.org/abs/2406.11217.

Fadillah Maani, Numan Saeed, Tausifa Jan Saleem, Zaid Farooq, Hussain Alasmawi, Werner Diehl, Ameera Mohammad, Gareth Waring, Saudabi Valappil, Leanne Bricker, and Mohammad Yaqub. Fetalclip: a visual-language foundation model for fetal ultrasound image analysis. *npj Digital Medicine*, Jun 2026. ISSN 2398-6352. doi: 10.1038/s41746-026-02907-9. URL https://doi.org/10.1038/s41746-026-02907-9.

Ministère de la Transition écologique. Base nationale de gestion ASsistée des procédures administratives relatives aux risques (GASPAR). data.gouv.fr, 2026. URL https://www.data.gouv.fr/datasets/base-nationale-de-gestion-assistee-des-procedures-administratives-relatives-aux-risques-gaspar. Accessed May 2026.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

NOAA NCEI. Storm events database. https://www.ncei.noaa.gov/stormevents/. Accessed March 2026.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022. URL https://arxiv.org/abs/2202.11214.

Public Safety Canada. Canadian disaster database (cdd). Government of Canada Open Government Portal, 2026. URL https://open.canada.ca/data/en/dataset/1c3d15f9-9cfa-4010-8462-0d67e493d9b9. Accessed May 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: https://doi.org/10.1029/2023MS004019. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2023MS004019.

Daniel Runfola, Austin Anderson, Heather Baier, Matt Crittenden, Elizabeth Dowker, Sydney Fuhrig, Seth Goodman, Grace Grimsley, Rachel Layko, Graham Melville, Maddy Mulder, Rachel Oberman, Joshua Panganiban, Andrew Peck, Leigh Seitz, Sylvia Shea, Hannah Slevin, Rebecca Youngerman, and Lauren Hobbs. geoboundaries: A global database of political administrative boundaries. *PLOS ONE*, 15(4):1–9, 04 2020. doi: 10.1371/journal.pone.0231866. URL https://doi.org/10.1371/journal.pone.0231866.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik Hamann, Campbell Watson, Manil Maskey, Tsengdar J Lee, Juan Bernabe Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate, 2024. URL https://arxiv.org/abs/2409.13598.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2026. URL https://arxiv.org/abs/2507.05201.

Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. In *Forty-second International Conference on Machine Learning*, ICML'25, 2025.

Sagar Soni, Akshay Dudhane, Hiyam Debary, Mustansar Fiaz, Muhammad Akhtar Munir, Muhammad Sohail Danish, Paolo Fraccaro, Campbell D Watson, Levente J Klein, Fahad Shahbaz Khan, and Salman Khan. Earthdial: Turning multi-sensory earth observations to interactive dialogues. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14313, 2025. doi: 10.1109/CVPR52734.2025.01334.

Shuo Tang, Jian Xu, Jiadong Zhang, Yi Chen, Qizhao Jin, Lingdong Shen, Chenglin Liu, and Shiming Xiang. Meteorpred: A meteorological multimodal large model and dataset for severe weather event prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22910–22919, June 2026.

S. M. Vallejo-Bernal, T. Braun, N. Marwan, and J. Kurths. Pikart: A comprehensive global catalog of atmospheric rivers. *Journal of Geophysical Research: Atmospheres*, 130(15):e2024JD041869, 2025. doi: https://doi.org/10.1029/2024JD041869. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2024JD041869.

Sumanth Varambally, Veeramakali Vignesh Manivannan, Yasaman Jafari, Luyu Han, Zachary Novack, Zhirui Xia, Salva Rühling Cachay, Srikar Eranky, Ruijia Niu, Taylor Berg-Kirkpatrick,

Duncan Watson-Parris, Yian Ma, and Rose Yu. Aquilon: Towards building multimodal weather LLMs. In *ICML 2025 Workshop on Assessing World Models*, 2025. URL https://openreview.net/forum?id=KVxOwEYAF4.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Christopher J. White, Henrik Carlsen, Andrew W. Robertson, Richard J.T. Klein, Jeffrey K. Lazo, Arun Kumar, Frederic Vitart, Erin Coughlan de Perez, Andrea J. Ray, Virginia Murray, Sukaina Bharwani, Dave MacLeod, Rachel James, Lora Fleming, Andrew P. Morse, Bernd Eggen, Richard Graham, Erik Kjellström, Emily Becker, Kathleen V. Pegion, Neil J. Holbrook, Darryn McEvoy, Michael Depledge, Sarah Perkins-Kirkpatrick, Timothy J. Brown, Roger Street, Lindsey Jones, Tomas A. Remenyi, Indi Hodgson-Johnston, Carlo Buontempo, Rob Lamb, Holger Meinke, Berit Arheimer, and Stephen E. Zebiak. Potential applications of subseasonal-to-seasonal (s2s) predictions. *Meteorological Applications*, 24(3):315–325, 2017. doi: https://doi.org/10.1002/met.1654. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.1654.

WMO. *WMO Guidelines on the Calculation of Climate Normals*. World Meteorological Organization, Geneva, Switzerland, 2017. URL https://library.wmo.int/idurl/4/55797. WMO-No. 1203.

Zhitong Xiong, Yi Wang, Weikang Yu, Adam J Stewart, Jie Zhao, Nils Lehmann, Thomas Dujardin, Zhenghang Yuan, Pedram Ghamisi, and Xiao Xiang Zhu. Dofa-clip: Multimodal vision-language foundation models for earth observation, 2025. URL https://arxiv.org/abs/2503.06312.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640, 2025. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.

Xuebin Zhang, Gabriele Hegerl, Francis W. Zwiers, and Jesse Kenyon. Avoiding inhomogeneity in percentile-based indices of temperature extremes. *Journal of Climate*, 18:1641—-1651, 2005. doi: 10.1175/JCLI3366.1. URL https://doi.org/10.1175/JCLI3366.1.

APPENDIX

Table 8: Comparison of recent existing multimodal weather and climate datasets by temporal and geographic coverage, evidence source, natural-phenomenon types, dataset size, and evaluation tasks.

Dataset	Years	Geo. Scope	Evidence Source	# NP Types	Size	Tasks
WeatherQA	2014–2020	CONUS	Expert forecast analyses	7	8,511	Severe weather MC-VQA and mesoscale case study generation
ClimateIQA	2023	Global	Meteorological heatmaps	3	26.3k img. (762.1k QA)	Meteorological anomaly VQA and description generation
Aquilon	1979–2022	Global	Expert-curated records & ERA5	N/R	444,244	Weather analysis, forecasting, and extreme event VQA
CLLMate	2015–2023	Global	Expert-verified news	13	7,747	Weather and climate phenomena and consequences MC-VQA
MeteorPred	2023–2024	China	Agency warnings	7	421,363	Severe weather VQA, MC-VQA, and structured generation tasks
EarthAlign	2000–2024	Global	Expert-curated records & ERA5	93	553,738	VL alignment: retrieval, classification, and physical variable estimation

NP: Natural phenomena. **N/R**: not reported or unknown. **VQA**: visual question answering. **MC**: multiple-choice. **VL**: vision-language.

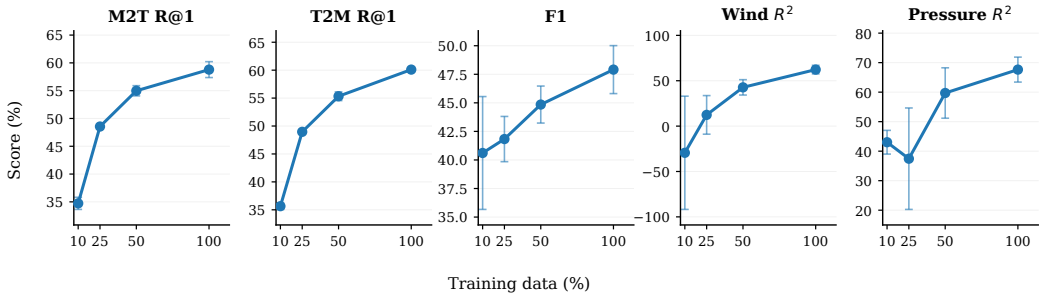


Figure 4: Effect of training data scale. We train Meteo-CLIP with 10%, 25%, 50%, and 100% of the training data and evaluate model performance across retrieval, classification, and physical variable estimation. Error bars denote standard deviation over three training runs.

A QUALITATIVE EXAMPLES



Figure 5: Examples from EarthAlign. The global map illustrates geographically diverse natural-event records, while the bottom panels provide detailed examples of a winter storm and a landslide with their spatial and temporal metadata along with four grounded caption styles.

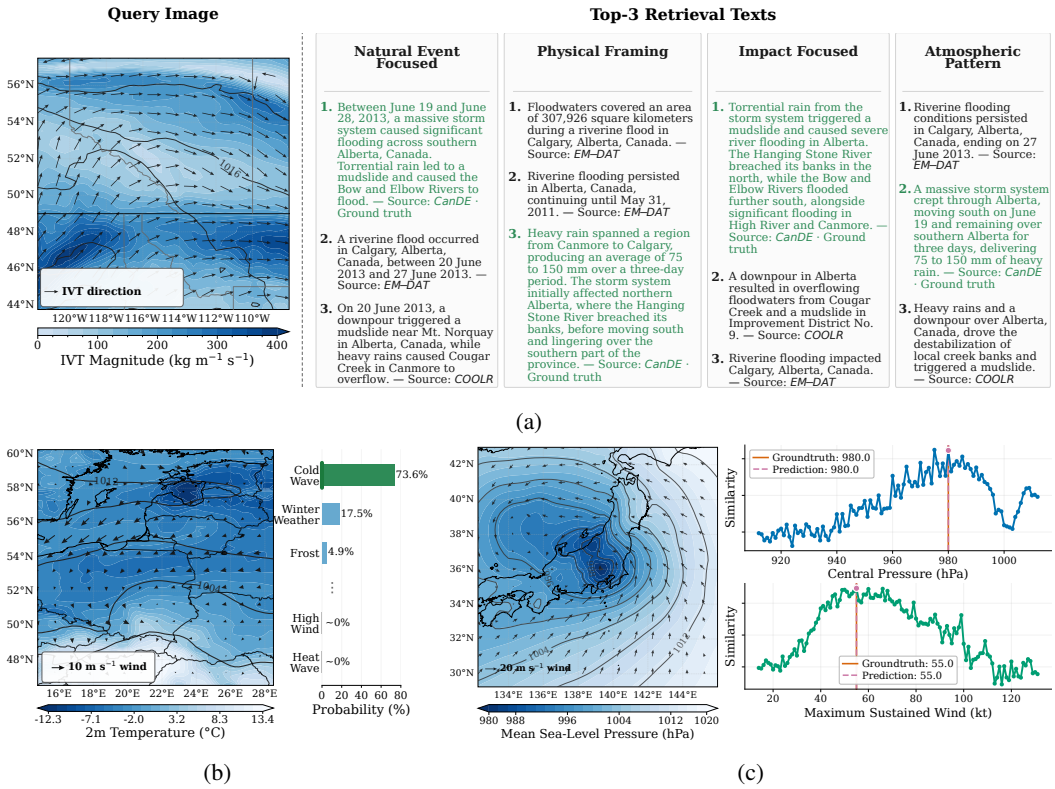


Figure 6: **Qualitative results using the MeteoLIP meteorology and text encoders.** (a) M2T retrieval retrieves captions describing the same or closely related event across different caption styles. (b) The cold-wave example assigns the highest probability to cold wave, followed by winter weather and frost. (c) The similarity profiles over candidate physical values show peaks around the predicted minimum central pressure and maximum sustained wind speed.

Ground Truth	Full MetTok	Geo. + Time Only
Example 1 A **bushfire** broke containment lines from a **prescribed burn** in Leeuwin-Naturaliste National Park in Western Australia. A subsequent **fire** started at Nannup, contributing to a broader **outbreak of fires** across the region.	**1.** A **bushfire** broke containment lines from a **prescribed burn** in Leeuwin-Naturaliste National Park in Western Australia. A subsequent **fire** started at Nannup, contributing to a broader **outbreak of fires** across the region. **2.** **Land fires** involving brush and pasture burned across the Tasmanian locations of Sommers Bay, Dunalley, and Bicheno in Australia. **3.** Two major **bushfires** broke out across metropolitan Perth, Western Australia, affecting the Perth Hills area. **4.** A **land fire** burned through brush and pasture in the Margaret River region of Western Australia. **5.** An **extreme heatwave** affected South Australia and Victoria. Temperatures exceeded 40 degrees Celsius in multiple locations, including Adelaide, Melbourne, and Dartmoor, which recorded a maximum of 43 degrees Celsius.	**1.** A series of **severe storms** struck Perth and Southern Western Australia. The events included a brief **tornado** with 125 km per hour winds in Dianella and Morley, followed by **heavy rain** , **thunder** , and **wind gusts** up to 140 km per hour. **2.** A **cyclone** , reported as either tropical or extratropical, was active in the South Indian basin off Western Australia, with maximum sustained winds of 20 knots and a central pressure of 1000 mbar. **3.** A **tropical cyclone** was active in the Western Australia sub-basin of the South Indian Ocean, with maximum sustained winds of 20 knots and a central pressure of 1000 mbar. **4.** A **tropical cyclone** was active in the Western Australia sub-basin of the South Indian Ocean, with maximum sustained winds of 20 knots and a central pressure of 1000 mbar. **5.** A **tropical cyclone** was active in the Western Australia sub-basin of the South Indian Ocean, with maximum sustained winds of 20 knots and a central pressure of 1000 mbar.
Example 2 **Post-Tropical Storm Irene** moved north into Canada, bringing **heavy rain** and **strong wind gusts** to Quebec and the Maritime provinces.	**1.** **Post-Tropical Storm Irene** moved north into Canada, bringing **heavy rain** and **strong wind gusts** to Quebec and the Maritime provinces. **2.** **Heavy rains** from a **tropical cyclone** triggered a **landslide** in L'Assomption, Quebec, Canada. **3.** **Extratropical cyclone conditions** were observed over Quebec, Canada, with maximum sustained **winds** of 40 knots and a central pressure of 987 millibars. **4.** **Wildfires** and associated **smoke** affected the communities of North West River and Sheshatshiu in Newfoundland and Labrador, Canada. **5.** An **extratropical cyclone** was located over New Brunswick, Canada, with maximum sustained **winds** of 40 knots and a central pressure of 1002 mbar.	**1.** **Normal weather conditions** were observed across parts of Canada, including Quebec, New Brunswick, Nova Scotia, Prince Edward Island, and Newfoundland and Labrador. **2.** **Normal weather conditions** were observed across parts of Canada, including New Brunswick, Ontario, Prince Edward Island, and Quebec. **3.** **Normal weather conditions** were observed across Newfoundland and Labrador and Quebec, Canada. **4.** **Normal weather conditions** were observed across New Brunswick and Quebec, Canada. **5.** **Normal weather conditions** were observed across New Brunswick and Quebec, Canada.
Example 3 A **downpour** triggered a **landslide** in Santa Cruz County, California, causing a combination of shale and mud to give way.	**1.** A **downpour** triggered a **landslide** in Santa Cruz County, California, causing a combination of shale and mud to give way. **2.** A **downpour** triggered at least two **mudslides** in the Santa Cruz mountains of California. **3.** **Downpours** triggered **landslides** in Marin County and along the Central California coast near Big Sur and Carmel. **4.** A storm brought **heavy rain** and winds to Central and Northern California, triggering a **mudslide** and downing trees in Monte Rio. **5.** A **downpour** triggered a complex **landslide** along the steep cliffs south of Carmel in Monterey County, California.	**1.** **Strong winds** with estimated gusts around 35 knots affected Northern Monterey Bay, California. **2.** A **strong wind event** brought estimated gusts of 33 to 61 knots to the San Francisco and Monterey Bay areas of California. **3.** A series of systems brought **heavy rain** and **strong winds** to the San Francisco and Monterey Bay areas of California, resulting in **floods** , **flash floods** , and **debris flows** . **4.** **Downpours** triggered **landslides** in Marin County and along the Central California coast near Big Sur and Carmel. **5.** A series of storm systems brought **strong winds** , **heavy rain** , and **debris flows** to the San Francisco and Monterey Bay areas of California.
Example 4 An **extreme heatwave** affected northeast Iowa, bringing maximum apparent temperatures between 100 and 110 each day.	**1.** **Excessive heat** settled over northeast Iowa, producing afternoon heat indices between 110 and 120. **2.** An **excessive heat event** affected Iowa, featuring daily temperatures in the 90s and overnight lows remaining in the 70s. **3.** An **extreme heatwave** affected northeast Iowa, bringing maximum apparent temperatures between 100 and 110 each day. **4.** An **excessive heat event** affected multiple counties in Iowa, with afternoon temperatures reaching the upper 90s to lower 100s. **5.** A significant **heat wave** affected multiple counties in Iowa, bringing the hottest weather with maximum temperatures soaring into the 90s and lower 100s.	**1.** A **mudslide** triggered by a **downpour** and **flash flooding** uprooted a tree in Lansing Township, Iowa. **2.** Several rounds of **thunderstorms** brought **heavy rain** , **hail** up to 1 inch, and **wind gusts** of 55 to 70 knots to northeast Iowa. **3.** **Thunderstorms** brought **heavy rain** , **lightning** , and **hail** up to one inch in diameter, leading to **flash flooding** across Floyd and Winneshiek counties in Iowa. **4.** A couple rounds of **thunderstorms** brought **heavy rain** , **hail** , and **strong winds** to northeast Iowa. **5.** **Severe thunderstorms** produced estimated **wind gusts** around 52 to 55 knots across parts of northeast Iowa.

Figure 7: Qualitative comparison of M2T retrieval using the full MetTok configuration and geolocation+time only, showing the top-5 retrieved texts for each ground-truth event, with green indicating meteorologically related events and red indicating unrelated phenomena. Full MetTok retrieves more meteorologically relevant events, while geolocation and time alone can retrieve unrelated phenomena.

Table 9: Expert-curated natural event and disaster sources used in EarthAlign. The table reports the geographic scope, provider type, natural phenomena type, years used, and number of records after source-specific processing.

Source	Geographic Scope	Provider Type	Natural Phenomena Type	Years Used	Num. of Records
EM-DAT	Global	Research institution	Natural hazards and disasters	2000–2024	25,172
NOAA	United States	Government agency	Severe weather events	2000–2024	286,292
IBTrACS	Global	Government agencies	Cyclones	2000–2024	84,764
COOLR	Global	Government agency	Landslides	2000–2024	12,511
PIKART	Global	Research institution	Atmospheric rivers	2000–2023	36,840
AusDE	Australia	Government agency	Natural hazards and disasters	2000–2014	95
MexDE	Mexico	Government agency	Natural hazards and disasters	2000–2016	1,628
CanDE	Canada	Government agency	Natural hazards and disasters	2000–2022	385
IndDE	Indonesia	Government agency	Natural hazards and disasters	2008–2024	34,623
FraDE	France	Government agency	Natural hazards and disasters	2000–2024	6,875

B Dataset Samples Record Construction

B.1 Expert-Curated Natural Event and Disaster Records

We collect natural event and disaster records from 10 expert-curated sources, summarized in Table 9. We apply source-specific processing to construct standardized metadata records. The resulting metadata is subsequently used to ground multi-style caption generation. We describe the processing for each source below.

NOAA Storm Events Database. The event record source (NOAA NCEI) provides records of severe weather events across the United States. We retain records from 2000–2024 and extract the available event metadata, including the episode and event narratives, event type, state and county, start and end times, event magnitude and magnitude type, flood cause, and tornado F/EF-scale rating. We further group event records using the combination of `EPISODE_ID` and `EPISODE_NARRATIVE`, and aggregate the associated event types, measurements, event narratives, locations, and start and end times. We additionally convert all reported local start and end times to UTC. To obtain a standardized geospatial extent for each episode, we map the reported state and county or weather-zone information to official U.S. geographic boundary data[1] using GeoPandas, and convert the matched geographic boundaries to their enclosing latitude–longitude bounding box. We then construct a standardized metadata record containing the episode start and end times, event types, location, available physical measurements, flood cause, tornado F/EF-scale rating, and narrative.

IBTrACS. IBTrACS (Knapp et al., 2010; Gahtan et al., 2024) provides the most comprehensive global archive of tropical cyclones. We retain records from 2000–2024 and extract the available metadata, including basin and sub-basin, cyclone nature, time, center location, WMO wind and pressure, storm translation speed and direction, and agency-provided cyclone classifications. We further retain only main-track records with agency-reported position and intensity values, i.e. those whose `TRACK_TYPE` is `main` and whose `IFLAG` contains `O` or `V`. We exclude records with an unknown or unreported cyclone nature. To obtain more detailed location information, we reverse geocode cyclone center coordinates using OpenStreetMap[2] for records whose `DIST2LAND` is less than 50 km. We additionally derive a more specific cyclone subcategory from the available agency-specific classifications while preserving differences among agency classifications. We then construct a standardized metadata record containing the snapshot time, basin and sub-basin, center location, cyclone nature and subcategory, WMO wind and pressure, and storm translation speed and direction.

EM-DAT. EM-DAT (Delforge et al., 2025) provides records of major disasters worldwide. We retain records from 2000–2024 whose disaster group is `Natural` and whose disaster subgroup

[1] `https://www2.census.gov/geo/tiger/GENZ2024/shp/cb_2024_us_county_500k.zip,https://www.weather.gov/source/gis/Shapefiles/WSOM/z_18mr25.zip,https://www.weather.gov/source/gis/Shapefiles/WSOM/mz18mr25.zip,https://www.weather.gov/source/gis/Shapefiles/WSOM/oz18mr25.zip,https://www.weather.gov/source/gis/Shapefiles/WSOM/hz20fe25.zip`

[2] `https://www.openstreetmap.org/`

is `Climatological`, `Meteorological`, `Hydrological`, or `Geophysical`, and exclude records with missing start or end years. We extract the available metadata, including disaster type and subtype, event name, country, location, origin, associated disaster types, magnitude and magnitude scale, and start and end dates. Each disaster record is uniquely identified by its `DisNo.`. To obtain a standardized geospatial extent for each record, we geocode each reported location using OpenStreetMap and aggregate the resulting geospatial boundaries to derive an enclosing latitude–longitude bounding box. For disasters spanning large geographic extents, we further perform spatial upsampling by generating localized bounding boxes for subregions within the original extent. Starting from 9,380 filtered records, our upsampling strategy generates 15,792 additional samples from 2,041 large-extent records. We then construct a standardized metadata record containing the disaster type and subtype, event name, origin, associated disaster types, magnitude and magnitude scale, start and end dates, and locations.

COOLR. COOLR (Juang et al., 2019) provides records of reported landslides and mass movements worldwide. We retain records from 2000–2024 with available latitude and longitude coordinates and extract the available metadata, including event date, event title and description, landslide category and trigger, and coordinates. To obtain a textual location for each event, we reverse geocode the reported latitude–longitude coordinates using OpenStreetMap and exclude records for which reverse geocoding does not return a location. We then construct a standardized metadata record containing the landslide category and trigger, location, event date, event title, and event description.

PIKART. PIKART (Vallejo-Bernal et al., 2025) provides a global high-resolution catalog of atmospheric rivers derived from ERA5 meteorological data based on integrated vapor transport (IVT). We retain records from 2000–2023 and select one timestamp from each atmospheric river track. For each selected timestamp, we extract the available metadata, including intersected continents, mean IVT, track centroid speed, and track centroid positions, and use the corresponding atmospheric river extent as the geospatial extent. To provide a qualitative characterization of IVT magnitude, we compute the distribution of mean IVT across all timestamps and tracks and categorize values below the 10th percentile as `low`, above the 90th percentile as `high`, and between the 40th and 60th percentiles as `moderate`. We additionally derive the atmospheric river propagation direction from the displacement of the track centroid between the beginning and end of each track. We then construct a standardized metadata record containing the snapshot time, intersected continents, mean IVT magnitude and category, and track-mean atmospheric river propagation speed and direction.

AusDE. AusDE (Australian Government, Attorney-General's Department, 2023) provides records of events and disasters across Australia from a government-curated source. We retain records from 2000–2014, exclude events occurring outside Australia, and remove non-natural events such as transport accidents, criminal acts, and industrial incidents. We extract the available metadata, including event title and description, start and end dates, and latitude–longitude coordinates. To obtain a standardized geospatial extent for each record, we use GPT-5.5 Pro with the event title, description, and provided coordinates to derive a latitude–longitude bounding box corresponding to the affected area. We then construct a standardized metadata record containing the disaster title, start and end dates, and event description.

MexDE. MexDE (Centro Nacional de Prevención de Desastres (CENAPRED), 2023) provides records of events and disasters across Mexico from a government-curated source. We retain records from 2000–2024 and exclude non-natural events such as chemical events. We extract the available metadata, including state, municipality, start and end dates, phenomenon type, classification, and observations, and translate the original Spanish-language metadata into English using GPT-5.5. We further group records sharing the same start date, end date, phenomenon type, classification, and observations. To obtain a standardized geospatial extent for each record, we geocode each reported municipality and state using OpenStreetMap, and aggregate the resulting bounding boxes to derive an enclosing latitude–longitude bounding box. We then construct a standardized metadata record containing the phenomenon type, observations, start and end dates, and location.

CanDE. CanDE (Public Safety Canada, 2026) provides records of events and disasters across Canada from a government-curated source. We retain records from 2000–2024 whose event group is `Natural` and exclude records whose event subgroup is `Biological`. We extract the available metadata, including event group and subgroup, disaster type, start and end dates, reported places, affected provinces, and event comments. To obtain a standardized geospatial extent for each record, we use GPT-5.5 Pro with the reported places, affected provinces, and event comments to

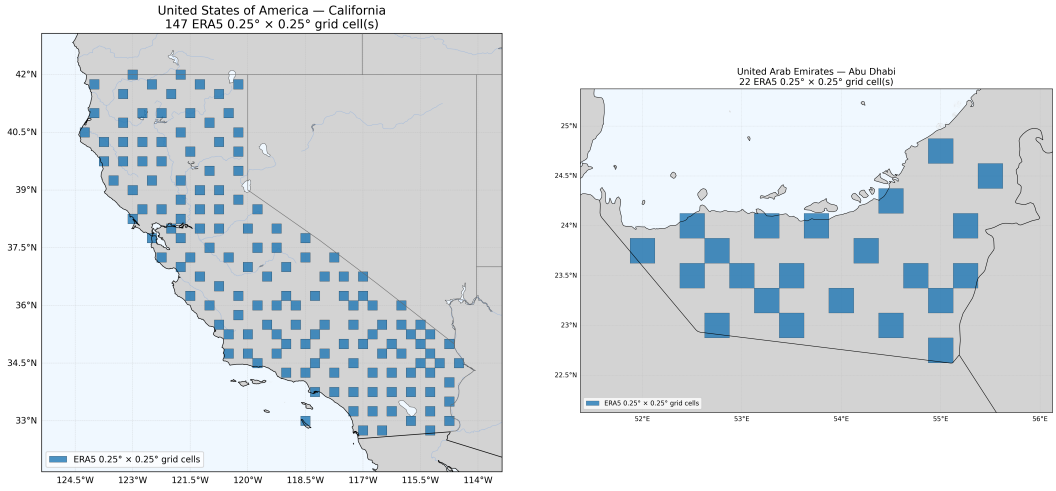


Figure 8: Examples of spatial grid-cell sampling for climatology-based normal-weather identification on California and Abu Dhabi.

derive a latitude–longitude bounding box corresponding to the affected area. We then construct a standardized metadata record containing the disaster type, locations, start and end dates, and event description.

IndDE. IndDE (Badan Nasional Penanggulangan Bencana, 2026) provides records of events and disasters across Indonesia from a government-curated source. We retain records from 2000–2024 and exclude non-natural incidents such as residential fires, social conflicts, and technological failures. We extract the available metadata, including disaster type, event type, province, regency or city, date, and time, and translate the disaster and event types from Indonesian into English using GPT-5.5. We further group records sharing the same disaster type, event type, province, and date, and aggregate their associated locations. We also remove reported times of `08:08:08 AM` and `12:01:00 AM`, which occur anomalously frequently in the source, and convert the remaining local times to UTC. To obtain a geospatial extent for each record, we geocode each reported regency or city using OpenStreetMap, and aggregate the resulting bounding boxes to derive an enclosing latitude–longitude bounding box. We then construct a standardized metadata record containing the disaster type and subtype, date, and locations.

FraDE. FraDE (Ministère de la Transition écologique, 2026) provides records of events and disasters across France from a government-curated source. We retain records from 2000–2024 and extract the available metadata, including disaster type, start and end dates, and municipality name and code. We further group records sharing the same disaster type, start date, and end date, and aggregate their associated municipalities. We translate the disaster type names from French into English using GPT-5.5. To obtain a standardized geospatial extent for each record, we match each reported municipality to its corresponding French commune[3] and aggregate the resulting administrative boundaries to derive a latitude–longitude bounding box. We then construct a standardized metadata record containing the disaster type, start and end dates, and locations.

B.2 Climatology-Based Normal-Weather Sampling

For each first-level administrative unit (ADM1), we identify ERA5 grid cells within its geoBoundaries administrative boundary (Runfola et al., 2020) and randomly sample 20% of the cells, with at least 10 cells whenever possible. For instance, as illustrated in Figure 8, this yields 147 sampled grid cells for California and 22 for Abu Dhabi. For each sampled cell, we retrieve ERA5 hourly variables from 1991 to 2025, including 2m temperature, 2m dewpoint temperature, skin temperature, mean sea-level pressure, surface pressure, sea-surface temperature, 10m and 100m wind speed, 10m wind gust, total precipitation, surface solar radiation downwards, cloud base height, significant height of

[3]French commune boundary data

Table 10: Variable-specific climatological criteria used for normal-weather sampling. Percentiles are computed independently for each ERA5 grid cell, day of year, UTC hour, and variable over the 1991–2020 reference period using a centered ±2-day window.

Criterion	Variables
P05–P95	2m temperature; 2m dewpoint temperature; skin temperature; mean sea-level pressure; surface pressure; sea-surface temperature
≤ P90	10m wind speed; 100m wind speed; 10m wind gust; total precipitation (1h, 3h, 6h, and 24h); 3h mean sea-level pressure change; significant height of combined wind waves and swell; mean wave period
≥ P10	Dewpoint depression; surface solar radiation downwards; cloud base height

combined wind waves and swell, and mean wave period, when applicable. We additionally derive 3h, 6h, and 24h precipitation accumulations, 3h mean sea-level pressure change, and dewpoint depression. The complete set of variables and their corresponding normality criteria is summarized in Table 10. For each grid cell, variable, day of year, and UTC hour, we estimate the climatological thresholds over the 1991–2020 WMO reference period (WMO, 2017) using observations pooled from the centered ±2-day window, yielding 150 observations for each threshold estimate. We then apply the grid-cell- and ADM1-level normality rules described in Section 3.1 to identify eligible normal-weather hours. We further restrict the resulting candidate pool to 2000–2024 and perform global sampling, yielding 64,553 normal-weather samples.

C Multi-Style Grounded Caption Generation and Audit

C.1 Caption Generation

For each metadata record, we use Gemini 3.1 Pro to generate one caption in each of four styles. *Natural-event-focused* captions describe the event or phenomenon directly, with time and location information when available; *physical-framing* captions emphasize its supported spatial or temporal organization; *impact-focused* captions emphasize supported weather-, climate-, environmental-, or hazard-related effects; and *atmospheric-pattern* captions emphasize provided meteorological or environmental conditions and setup. The generation prompt requires all captions to be directly supported by the input metadata and prohibits unsupported causes, impacts, forecasts, and descriptive modifiers. When the metadata does not sufficiently support a requested style, the model is instructed to produce a minimal grounded caption rather than infer missing style-specific content. The representative metadata and generated captions are shown in Figure 9. We further report the distribution of caption lengths in Figure 10. The full generation prompt is provided below.

Grounded Caption Generation Prompt

You are an expert caption writer for climate and weather data.
Task: Generate grounded captions from the provided metadata/context.
Goal: Produce stylistically diverse captions that are strictly supported by the input, preserve informative uncertainty when present, and are suitable for text-ERA5 climate data alignment.
Hard constraints:

- Do not add any fact that is not directly supported by the input.
- Do not infer specific causes, future outcomes, identities, or impacts unless explicitly given.
- Do not describe objects, actions, regions, or conditions not present in the metadata/context.
- If a detail is missing, leave it out.
- If the input contains explicit uncertainty, disagreement, multiple reported classifications, or alternative labels, do not replace them with a single definite claim. Retain the ambiguity when it is relevant to the caption.
- Do not mention missing, unspecified, or unavailable information.
- Do not force all supported metadata into every caption. Include only the details that best fit the requested style.
- Do not refer to the metadata, narrative, source text, or report inside the caption.



Figure 9: Representative metadata records and their generated captions across record sources.

- Do not convert numeric or categorical metadata into qualitative descriptions unless that wording is directly supported by the input.

- Do not add unsupported descriptive modifiers such as compact, broad, organized, intense, weak, rapidly, slowly, strengthening, or weakening.
- Do not mention vehicles, casualties, injuries, deaths, missing people, houses, buildings, infrastructure, property damage, economic losses, or other human-impact or built-environment details. Focus only on the supported natural phenomena, weather conditions, environmental conditions, or natural hazard type.
- If a requested style is not supported, still produce that caption, but reduce it to a minimal, fully grounded restatement of the event rather than inventing missing structure, impacts, or atmospheric setup.
- Do not make every caption a full recap of the same event metadata.
- When mentioning time, use hour-only format such as 21 UTC, not 21:00 UTC.
- Do not repeat the same system name, event name, storm name, phenomenon name, or cyclone name in every caption.
- Do not mention exact time in every caption. Use time only when it is important to the grounded description or helps distinguish the event more clearly.

Writing requirements:

- Write exactly one caption for each style.
- Keep each caption natural, clear, and concise.
- Ensure each caption reads smoothly as a standalone caption.
- Each caption must stand on its own and should not assume prior context.
- Avoid generic referents such as ”the system” when a more natural grounded alternative can be used.
- Prefer caption-like phrasing over report-like phrasing.
- Make the captions stylistically different by varying emphasis, structure, and phrasing, not by introducing unsupported facts or unsupported interpretations.
- Avoid repeating the same wording, sentence structure, or lead phrase across captions.
- Prefer a single sentence, but use 2 to 3 sentences when needed to capture supported details.
- Correct obvious surface errors such as typos, stray punctuation, spacing issues, duplicated words, and minor formatting problems when the intended meaning is clear.
- Write all location names in plain ASCII. Use common English names or ASCII romanized forms for locations that contain accents, non-Latin scripts, or special punctuation.
- Do not discard informative input details solely because they are uncertain or reported differently; maintain their uncertainty status when relevant to the caption.
- If the input includes too many location names, you may summarize them with a shorter shared parent location only when that summary is clearly supported by the input and does not introduce a new regional interpretation.
- If you mention a date, you must include the full supported date including the year.
- Before writing each caption, select only the input details most relevant to that style, and do not reuse the same full set of details across all four captions unless the input does not provide enough distinct supported details to differentiate them further.
- When the input lacks enough distinct supported details for strong stylistic separation, keep weakly supported captions shorter and more selective rather than repeating the same full metadata summary.
- Express numeric metadata naturally, and avoid raw technical notation when it does not read smoothly as a caption.

Allowed caption styles:

1. natural_event_focused
   - Focus on what is happening in a straightforward and natural way.
   - Do not include unrelated human-interest details, casualty counts, or incidental impacts.
   - Date and time may be included when it helps anchor the event naturally.
2. physical_framing
   - Focus on how the event or conditions were distributed, organized, extended, or expressed across space or time.
   - Prefer wording about extent, spread, concentration, coverage, or visible physical conditions when supported.
   - Do not simply restate the natural-event caption with reordered wording.
   - Do not include time unless it is needed to describe the supported spatial or temporal arrangement.
3. impact_focused
   - Focus on supported climate-, weather-, environmental, or hazard-related effects of the event.
   - Avoid named individuals, casualty counts, and highly itemized damage unless explicitly needed.

- Do not center unrelated human-interest or incidental impacts.
- If only human-loss impacts are supported, describe the event itself without adding consequences.
- Do not include time unless it is needed to anchor the reported impacts.

4. atmospheric_pattern
   - Focus on atmospheric, meteorological, or environmental setup only when it is explicitly described in the input.
   - Do not introduce meteorological structure unless it is supported.
   - If no such pattern is explicitly supported, keep the caption grounded by describing the reported conditions.
   - Do not include time unless it is needed to ground the reported setup or evolution.

Final check before answering:

- Is every statement directly supported by the input?
- Are unsupported causes, impacts, and forecasts excluded?
- Were obvious surface errors cleaned up without changing the meaning or introducing unsupported interpretations?
- If multiple locations were summarized, is the shared location label clearly supported by the input?
- Is each caption aligned with its style?
- Is each caption concise and natural?
- Is the output valid JSON with exactly the required keys?

Output format:
Return valid JSON only, with exactly the required keys and string values.
Do not output markdown, comments, code fences, or any text outside the JSON object.
Input metadata/context:
`{input_placeholder}`

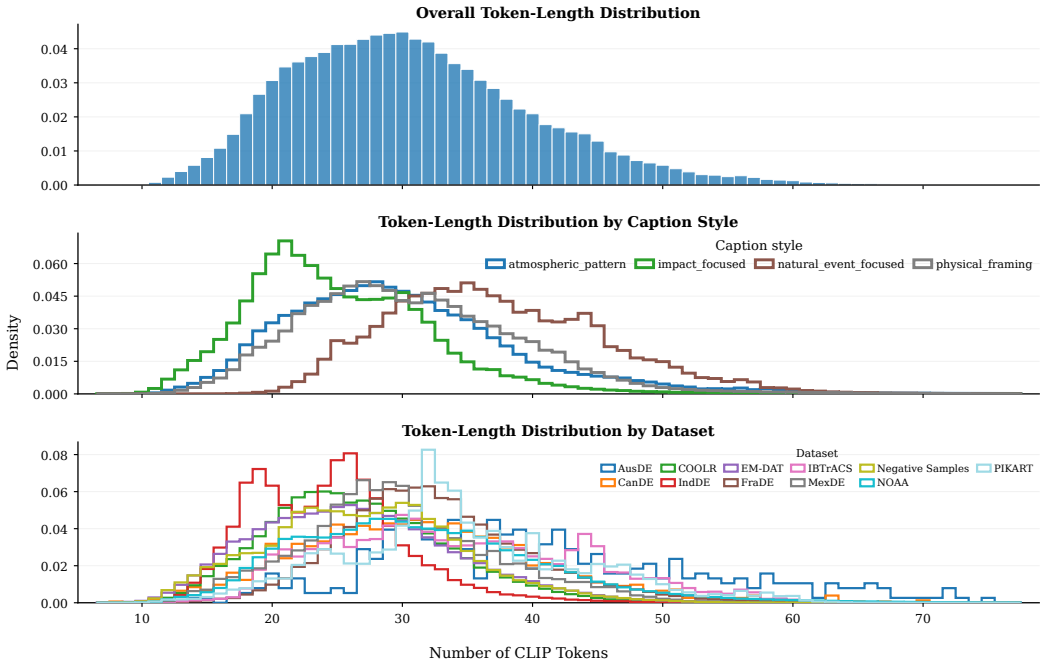


Figure 10: Distribution of generated caption lengths measured by using the CLIP text tokenizer. We report the overall distribution, distributions across the four caption styles, and distributions across record data sources.

C.2 Caption Audit and Error Analysis

Caption audit. A trained graduate annotator manually audits 440 captions, comprising 10 records from each of 11 record sources (i.e. 10 expert-curated natural event and disaster sources and the

normal-weather samples) across all four caption styles. Each caption is evaluated for *date/time accuracy*, *location accuracy*, *event/phenomenon accuracy*, *trigger/cause accuracy*, *unsupported added detail*, *unsupported modifiers*, *weather/climate focus*, and *standalone clarity*. Table 11 summarizes the results, with a caption counted as passing only when all audit criteria are satisfied.

Error analysis. Among the erroneous captions, the most frequent audit failure category is weather/climate focus (7 cases), followed by location accuracy (6), unsupported modifiers (2), trigger/cause accuracy (2), unsupported added detail (1), and event/phenomenon accuracy (1). Table 12 provides representative examples. In particular, several impact-focused failures emphasize human consequences such as drowning or transport accidents rather than focusing on weather, climate, environmental, or hazard-related effects. Other errors arise from overgeneralized locations, unsupported qualitative modifiers, or descriptions that alter the identity of the reported phenomenon.

Table 11: Manual audit results for generated captions. A trained graduate annotator verified 10 samples from each record source across all caption styles, totaling 440 captions.

Caption Style	Audited	Passed	Pass Rate
Natural Event Focused	110	110	100.00%
Physical Framing	110	108	98.18%
Impact Focused	110	98	89.09%
Atmospheric Pattern	110	105	95.45%
Overall	**440**	**421**	**95.68%**

D ERA5 Data Alignment and Preprocessing

ERA5 variables. We retrieve ERA5 hourly data on single and pressure levels through the Copernicus Climate Data Store (CDS) API. Each meteorological input contains 28 single-level variables, comprising 25 atmospheric and land variables and 3 wave variables, together with 7 pressure-level variables at 925, 850, 700, 500, and 300 hPa. This results in 63 variables in total. The complete variable set is listed in Table 13.

Temporal alignment. We align each processed record to a single ERA5 hour based on its available temporal information. For IBTrACS, PIKART, and NOAA or IndDE records with an exact timestamp, we use the reported timestamp directly. For NOAA and climatology-based normal-weather records specified over a start–end time interval, we uniformly sample one hourly timestamp within the reported interval. COOLR and some IndDE records provide only a date; for these records, we uniformly sample one of the 24 hours within the reported day in local time and subsequently convert it to UTC. EM-DAT, FraDE, CanDE, AusDE, and MexDE provide date ranges, from which we uniformly sample one hourly timestamp within the reported interval. All reported or sampled timestamps are standardized to UTC before ERA5 retrieval.

Spatial alignment. The processed records contain either a spatial bounding-box extent or a single center location. NOAA, EM-DAT, PIKART, AusDE, MexDE, CanDE, IndDE, FraDE, and climatology-based normal-weather records provide bounding-box extents, whereas IBTrACS and COOLR provide only a center latitude–longitude coordinate. For IBTrACS and COOLR, we construct a $14^\circ \times 14^\circ$ region centered on the reported coordinate. For records with an existing bounding box smaller than $14^\circ \times 14^\circ$, we preserve its center and symmetrically expand each dimension to a minimum extent of 14° whenever possible. Near the northern or southern geographic boundary, where symmetric expansion would extend beyond 90° or -90° latitude, we shift the excess latitudinal extent in the opposite direction. For records whose extent exceeds 14° in either dimension, we use the larger of the latitudinal and longitudinal extents and symmetrically expand the shorter dimension to construct a square region containing the original bounding box.

Standardization. We retrieve ERA5 patches based on the selected timestamp and spatial extent. We then standardize each extracted patch to 56×56 grid cells using linear interpolation for all variables except mean wave direction, for which we use nearest-neighbor interpolation. We further normalize each variable using its mean and standard deviation computed across samples.

Table 12: Representative caption audit failures and their corresponding failure categories.

Style	Generated Caption	Audit Comment
Impact Focused	A rip current resulted in a drowning in Coastal Horry, South Carolina, on 13 June 2022.	*Weather/climate focus:* drowning is a human-impact outcome rather than a weather-, climate-, or environmental effect.
Impact Focused	A weak IVT event carried moisture at a mean magnitude of 249.224 kg/(m·s).	*Event/phenomenon accuracy:* a weak IVT event does not necessarily correspond to an atmospheric river and changes the identity of the reported phenomenon.
Impact Focused	Hazardous snow and ice accumulations from Storm Bettina led to transport accidents in German locations like Schwabisch Hall and Denzeheide.	*Weather/climate focus:* transport accidents are incidental human impacts and should not be the focus of the caption.
Impact Focused	The winter storm produced extreme snow drifts in Prince Edward County, alongside significant snow accumulation and drifting along the north shore of Lake Ontario.	*Unsupported modifier:* "extreme" is not supported by the input metadata.
Physical Framing	Significant precipitation and lake effect snow spread across Eastern Canada, with intense snow drifting extending along the north shore of Lake Ontario through to Kingston.	*Location/modifier:* the lake-effect snow is described more specifically along Lake Ontario through Kingston, thus overgeneralizing its spatial extent; *Unsupported modifier:* "intense" is unsupported.

Table 13: ERA5 variables used to construct the meteorological inputs. The 28 single-level variables are combined with 7 pressure-level variables at 5 pressure levels, yielding 63 variables in total.

Variable Group	Variables
Single-level atmospheric and land	2m temperature; 2m dewpoint temperature; 10m u-component of wind; 10m v-component of wind; instantaneous 10m wind gust; mean sea-level pressure; total precipitation; convective precipitation; snowfall; snow depth; snowmelt; runoff; volumetric soil water layer 1; volumetric soil water layer 2; volumetric soil water layer 3; volumetric soil water layer 4; convective available potential energy; convective inhibition; boundary layer height; cloud base height; total cloud cover; zero-degree level; total column water vapour; vertical integral of eastward water vapour flux; vertical integral of northward water vapour flux
Single-level wave	Significant height of combined wind waves and swell; mean wave direction; peak wave period
Pressure-level	Geopotential; temperature; u-component of wind; v-component of wind; specific humidity; vertical velocity; vorticity, each at 925, 850, 700, 500, and 300 hPa

E Additional Dataset Statistics and Distributions

Regional distribution. Figure 11 shows the geographic distribution of EarthAlign across 19 regions. As illustrated, EarthAlign spans a broad range of geographic regions with North America being the most represented, largely reflecting the NOAA Storm Events Database which contributes 286,292 records from the United States. The remaining global and country-specific sources provide additional coverage across other regions, resulting in broad geographic coverage with a notable concentration in North America.

Weather condition class distribution. Figure 12 shows the class distribution for the weather-condition classification task across the training and test sets. The two splits exhibit broadly similar distributions across most classes, with drought showing the largest relative difference. The resulting class imbalance reflects the underlying distribution of records across the constituent record sources.

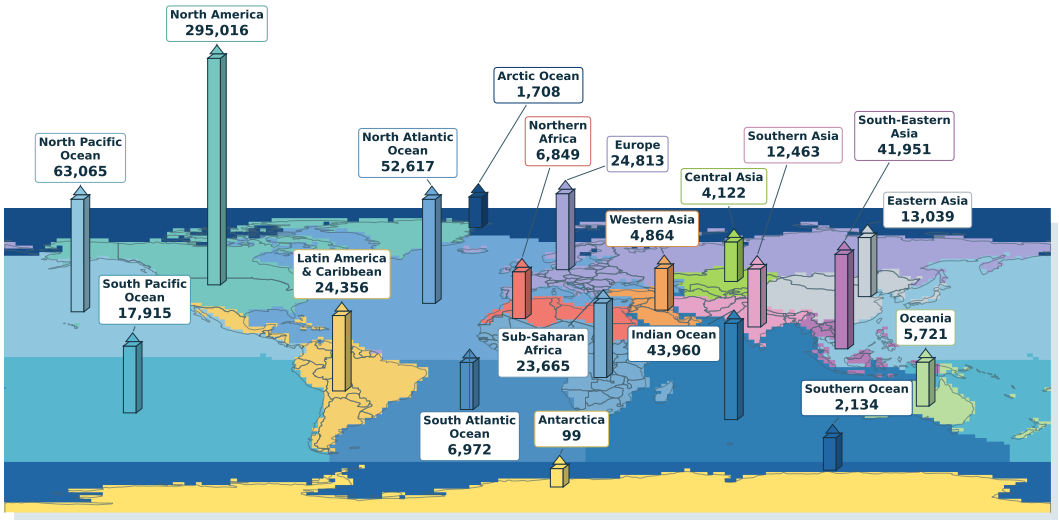


Figure 11: Regional distribution of EarthAlign records. Records spanning multiple regions contribute to more than one regional count.

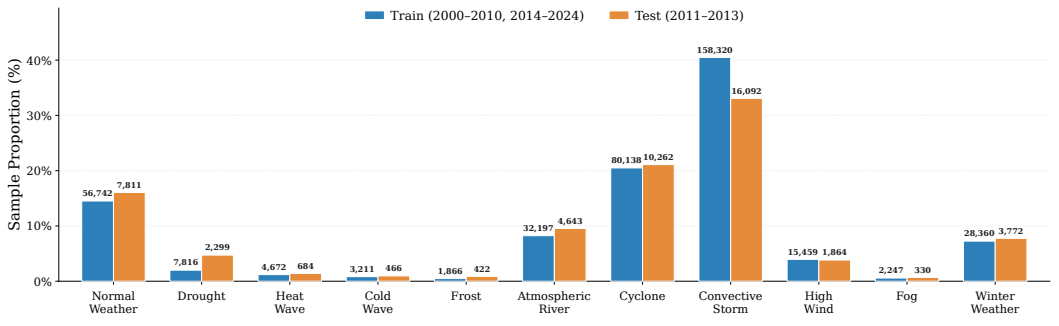


Figure 12: Class distribution for the weather condition classification task in the training and test sets.

F Additional Analysis

F.1 Weather Condition Classification Analysis

Table 14 provides category-wise classification results, with the corresponding confusion matrices shown in Figure 13. Under text-prototype inference, Meteo-CLIP and Meteo-SigLIP generally improve category-level performance over their respective CLIP and SigLIP baselines. The improvements are particularly notable for atmospheric river, convective storm, high wind, and winter weather. Under linear probing, performance increases substantially across almost all categories, with Meteo-CLIP and Meteo-SigLIP achieving macro F1 scores of 79.98 and 80.75 respectively. The consistently stronger linear-probing results indicate that the frozen meteorology representations encode substantial category-level information.

Across weather conditions, drought, heat wave, and cold wave remain the most challenging categories. Even under linear probing, their F1 scores are substantially below those of atmospheric river, cyclone, and convective storm, which exceed 94 F1. The lower performance of drought, heat wave, and cold wave cannot be solely attributed to limited test data, as frost achieves above 82 F1 with both MeteoLIP variants despite having only 422 test samples, fewer than drought, heat wave, and cold wave. The confusion matrices further show that these challenging categories are frequently confused with other meteorological conditions, including normal weather.

The lower performance on drought, heat wave, and cold wave potentially reflects their inherently temporal characteristics. For instance, the Copernicus Warm and Cold Spell indicators define such

Table 14: Per-category classification performance under *text-prototype inference* and *linear probing* across 11 weather conditions. We report class-wise F1 scores and macro F1, with the number of test samples shown in parentheses. **Bold** indicates the best result and underline indicates the second-best.

	Category	CLIP	Meteo-CLIP	SigLIP	Meteo-SigLIP
Text-Prototype Inference	Normal weather (n=7,811)	42.59 ± 4.77	50.84 ± 2.65	34.07 ± 7.09	$\mathbf{56.29 \pm 5.64}$
	Drought (n=2,299)	18.66 ± 1.33	23.42 ± 4.14	14.20 ± 2.15	$\mathbf{25.36 \pm 3.51}$
	Heat wave (n=684)	16.81 ± 3.61	28.14 ± 3.79	15.45 ± 6.20	$\mathbf{43.04 \pm 8.05}$
	Cold wave (n=466)	$\mathbf{25.49 \pm 5.61}$	17.98 ± 0.71	14.83 ± 0.81	24.77 ± 1.52
	Frost (n=422)	12.46 ± 1.73	33.32 ± 6.93	17.74 ± 7.77	$\mathbf{48.48 \pm 15.29}$
	Atmospheric river (n=4,643)	31.37 ± 10.18	$\mathbf{72.37 \pm 4.42}$	12.70 ± 1.29	69.19 ± 4.15
	Cyclone (n=10,262)	54.07 ± 7.14	$\mathbf{61.67 \pm 5.80}$	27.12 ± 4.06	42.07 ± 8.38
	Convective storm (n=16,092)	25.20 ± 4.59	$\mathbf{65.39 \pm 0.30}$	12.11 ± 8.89	57.70 ± 15.64
	High wind (n=1,864)	30.31 ± 1.76	59.29 ± 4.24	29.29 ± 6.24	$\mathbf{66.00 \pm 6.36}$
	Fog (n=330)	8.55 ± 2.22	33.77 ± 7.33	6.92 ± 2.27	$\mathbf{35.31 \pm 10.16}$
	Winter weather (n=3,772)	65.89 ± 3.33	80.70 ± 3.24	52.59 ± 7.47	$\mathbf{84.09 \pm 1.83}$
	Macro F1 ↑	30.13 ± 2.83	47.90 ± 2.09	21.55 ± 1.48	$\mathbf{50.21 \pm 2.95}$
Linear Probing	Normal weather (n=7,811)	90.37 ± 0.13	91.74 ± 0.13	90.69 ± 0.12	$\mathbf{93.00 \pm 0.04}$
	Drought (n=2,299)	48.80 ± 0.46	52.54 ± 0.89	51.41 ± 1.17	$\mathbf{56.70 \pm 0.61}$
	Heat wave (n=684)	47.30 ± 1.35	61.94 ± 1.15	52.84 ± 1.49	$\mathbf{62.41 \pm 0.86}$
	Cold wave (n=466)	48.43 ± 0.79	49.49 ± 0.95	48.98 ± 1.93	$\mathbf{54.59 \pm 1.23}$
	Frost (n=422)	75.38 ± 1.69	$\mathbf{83.14 \pm 0.54}$	75.40 ± 2.56	82.12 ± 0.96
	Atmospheric river (n=4,643)	96.46 ± 0.18	$\mathbf{97.92 \pm 0.10}$	96.86 ± 0.14	97.73 ± 0.12
	Cyclone (n=10,262)	98.11 ± 0.06	$\mathbf{98.31 \pm 0.04}$	98.24 ± 0.04	98.24 ± 0.02
	Convective storm (n=16,092)	94.07 ± 0.09	$\mathbf{95.08 \pm 0.16}$	94.25 ± 0.03	95.03 ± 0.04
	High wind (n=1,864)	74.86 ± 0.55	$\mathbf{80.84 \pm 0.74}$	78.00 ± 0.22	80.03 ± 0.22
	Fog (n=330)	62.75 ± 0.96	$\mathbf{76.81 \pm 1.63}$	62.36 ± 0.55	76.70 ± 0.55
	Winter weather (n=3,772)	88.95 ± 0.24	$\mathbf{91.98 \pm 0.08}$	90.31 ± 0.15	91.73 ± 0.22
	Macro F1 ↑	75.04 ± 0.20	79.98 ± 0.26	76.30 ± 0.54	$\mathbf{80.75 \pm 0.14}$

events from sequences of at least three hot or cold days based on daily minimum and maximum temperature anomalies[4]. Similarly, meteorological drought indicators such as the Standardized Precipitation Index (SPI) characterize precipitation deficits over accumulated time scales ranging from months to years[5]. In contrast, MeteoLIP currently takes a single meteorological snapshot as input, thereby limiting its ability to capture the temporal persistence and accumulated anomalies underlying these conditions. These findings motivate future extensions of MeteoLIP to incorporate multiple timestamps or temporal sequences, which can help analyze meteorological conditions over time.

F.2 Physical Baselines for Cyclone Intensity Estimation

Since mean sea-level pressure and 10-m wind are included in the ERA5 input, we additionally compare MeteoLIP with physical baselines derived directly from these variables. For minimum central pressure, the direct ERA5 baseline takes the minimum mean sea-level pressure within each input patch, while for maximum sustained wind speed, it takes the maximum 10-m wind speed. We also fit a linear regression from each ERA5-derived quantity to its corresponding IBTrACS target using the training set. As shown in Table 15, the direct ERA5 baselines achieve R^2 values of 51.96 for minimum central pressure and 10.78 for maximum sustained wind speed, which increase to 63.78 and 48.02 with linear regression. In comparison, linear probing of Meteo-CLIP achieves R^2 values of 76.36 and 71.34, while Meteo-SigLIP achieves 77.36 and 71.85, respectively. These results show that linear probing of the learned meteorological representations achieves better predictive performance than directly extracting or linearly calibrating the corresponding ERA5 variables.

[4]EDO INDICATOR FACTSHEET - Warm Spell and Cold Spell Duration and Intensity Indicators

[5]EDO AND GDO INDICATOR FACTSHEET - Standardized Precipitation Index (SPI)

Table 15: Physical baselines for cyclone intensity estimation. **Bold** and underlined values indicate the best and second-best results, respectively.

Variable	Method	$R^2 \uparrow$	MAE↓	RMSE↓
Wind	Direct ERA5	10.78	14.50	20.83
	ERA5 Linear Regression	48.02	12.27	15.90
	Meteo-CLIP	<u>71.34</u>	<u>8.65</u>	<u>11.80</u>
	Meteo-SigLIP	**71.85**	**8.50**	**11.70**
Pressure	Direct ERA5	51.96	7.61	13.44
	ERA5 Linear Regression	63.78	8.43	11.67
	Meteo-CLIP	<u>76.36</u>	<u>6.41</u>	<u>9.43</u>
	Meteo-SigLIP	**77.36**	**6.16**	**9.23**

Table 16: Effect of initialization on downstream fine-tuning using the same Meteo-SigLIP architecture. **Bold** and underline indicate the best and second-best results, respectively.

	Classification		Sustained Wind		Central Pressure	
Initialization	**F1** ↑	**Top-1 Acc.** ↑	R^2 ↑	**MAE** ↓	R^2 ↑	**MAE** ↓
Scratch	77.02 ± 0.76	91.56 ± 0.16	72.55 ± 1.04	8.23 ± 0.18	78.89 ± 0.74	5.65 ± 0.13
WebLI	<u>78.38 ± 0.64</u>	<u>91.88 ± 0.20</u>	<u>74.31 ± 0.28</u>	<u>7.89 ± 0.08</u>	<u>80.65 ± 0.37</u>	<u>5.36 ± 0.05</u>
MeteoLIP	**81.31 ± 0.47**	**92.97 ± 0.15**	**76.15 ± 0.35**	**7.56 ± 0.06**	**81.24 ± 0.38**	**5.24 ± 0.03**

F.3 Downstream Fine-Tuning

Using the same Meteo-SigLIP architecture, we fine-tune models initialized from scratch, the SigLIP WebLI checkpoint, and our Meteo-SigLIP pretrained checkpoint on the classification and physical variable estimation tasks. As shown in Table 16, MeteoLIP initialization achieves the best performance, achieving 81.31 F1 and 92.97 Top-1 accuracy for classification, R^2 of 76.15 with 7.56 MAE for maximum sustained wind speed, and R^2 of 81.24 with 5.24 MAE for minimum central pressure. These results indicate that meteorology-language pretraining provides a stronger initialization for task-specific fine-tuning, further demonstrating the transferability of the learned multimodal representations.

F.4 Downstream Transfer to WeatherQA

We further evaluate the transferability of Meteo-CLIP representations on WeatherQA (Ma et al., 2024), a severe-weather reasoning benchmark over the CONUS, using 7,304 training samples and 600 test samples. WeatherQA comprises a multiple-choice question answering task for the affected geographic region (*Area Affected*) and a classification task for the primary weather concern (*Concerning*).

For each WeatherQA sample, we extract spatial patch representations from the frozen Meteo-CLIP meteorology encoder, map them to the hidden space of Qwen3-VL-4B-Instruct (Bai et al., 2025) using a learnable projection layer, and adapt the language model using LoRA (Hu et al., 2022) with rank 16 and $\alpha = 32$. We train for five epochs with AdamW using a batch size of 16, learning rates of 5×10^{-5} and 2×10^{-4} for LoRA and the projection layer, respectively, and a cosine learning-rate schedule.

Meteo-CLIP achieves $49.22 \pm 4.45\%$ accuracy on *Area Affected*, which requires selecting the affected region from four candidate options, and $51.73 \pm 2.41\%$ conditional accuracy on *Concerning*, an 11-category weather-concern classification task evaluated when the affected area is correctly predicted. These results show that the frozen Meteo-CLIP representations encode spatial and meteorological information that can be effectively transferred to downstream severe-weather reasoning, further demonstrating the usefulness of the learned meteorological representations beyond the Earth-Align evaluation tasks.

WeatherQA Qualitative Example

System Prompt:
You are an AI assistant with expertise in severe weather analysis and forecasting. You are provided with a meteorological representation describing atmospheric and surface conditions across the contiguous United States. Your responsibility is to assess potential severe weather conditions, affected areas, and weather concerns.

User Prompt:
<Parameters>
The provided meteorological representation describes atmospheric and surface conditions across the contiguous United States at 11 / 12, 16UTC.
The meteorological representation is derived from a 63-variable ERA5 snapshot.
Pressure-level variables at 300, 500, 700, 850, and 925 hPa:
geopotential, temperature, zonal wind, meridional wind, specific humidity, vertical velocity, and relative vorticity.
Surface and single-level variables:
volumetric soil water content in layers 1–4; vertically integrated eastward and northward water-vapour flux; significant wave height; peak wave period; mean wave direction; snowmelt; convective precipitation; snowfall; runoff; total precipitation; convective inhibition; cloud-base height; zero-degree level; instantaneous 10-m wind gust; convective available potential energy; total column water vapour; snow depth; mean sea-level pressure; boundary-layer height; total cloud cover; 10-m zonal and meridional wind; 2-m temperature; and 2-m dewpoint temperature.
<meteorological_representation>
</Parameters>

<Thought process>
Use the provided meteorological representation to answer the following multiple-choice and classification questions:

1. Analyze the spatial and meteorological conditions represented in <Parameters> and identify regions showing signals associated with severe or hazardous weather.
2. Combine the relevant meteorological information to assess the location and severity of the weather conditions.
3. Map the conclusions to the provided options. Only give the correct option characters, without any other details or explanations. Use JSON format with the keys 'Area_Affected' for the first answer and 'Concerning' for the second answer.

</Thought process>

<Item number="1">
Question: Choose one geographical area(s) most likely to be impacted by the severe weather event from the four options provided.

A southern/eastern North Carolina and northeastern South Carolina
B Central Kansas and Oklahoma
C portions of northeast Iowa into far southwest Wisconsin
D Far southeast Nebraska...northeast Kansas...extreme northwest Missouri

Answer: <answer>Option</answer>
</Item>

<Item number="2">
Question: Choose the most likely scenario regarding severe weather concerns from the options provided. Options include whether a convective watch has benn issued, the probability of a future watch, or, in winter, potential weather phenomena related to winter storms.

A Severe potential...Watch unlikely (confidence level on the expectation of a watch is 5 or 20%)
B Severe potential...Watch possible (confidence level on the expectation of a watch is 40 or 60%)
C Severe potential...Watch likely (confidence level on the expectation of a watch is 80 or 95%)
D Heavy snow
E Severe potential...tornado watch likely (confidence level on the expectation of a watch is 80 or 95%)
F Severe potential...severe thunderstorm watch likely (confidence level on the expectation of a watch is 80 or 95%)
G Winter mixed precipitation
H Freezing rain
I Severe potential...watch needed soon (confidence level on the expectation of a watch is 95%)
J Blizzard
K Snow squall

Answer: <answer>Option</answer>
</Item>

Model Output:

```
{"Area_Affected": "A", "Concerning": "A"}
```

F.5 Training Dynamics

Figure 14a shows the training dynamics of Meteo-CLIP. As training progresses, the CLIP test loss decreases and gradually converges, while retrieval performance generally improves. We observe a similar trend for physical variable estimation, where the R^2 values increase substantially after the early training epochs and then remain relatively stable. In contrast, the classification F1 exhibits larger fluctuations and does not show a similarly consistent increasing trend.

To further examine whether these fluctuations reflect the quality of the learned meteorological representations, we perform linear probing on the frozen meteorology encoder checkpoints from epochs 6, 12, 18, 24, and 30. As shown in Figure 14b, both F1 and Top-1 accuracy generally improve with training, while Top-3 accuracy remains stable at approximately 98.8%. This indicates that although classification performance under text-prototype inference in the learned meteorology-language space can fluctuate across epochs, the frozen meteorology representations continue to become more discriminative for category-level prediction. The nearly saturated Top-3 accuracy further suggests that later training primarily improves the ranking of the correct class among the top candidate categories, which is better reflected by F1 and Top-1 accuracy.

F.6 Positional Encoding in the Meteorological Encoder

We investigate two positional encoding approaches for the meteorological encoder, i.e. learnable positional embeddings (LPE) and 2D rotary positional embeddings (RoPE) (Heo et al., 2024). For 2D RoPE, we adopt the *RoPE-Mixed* variant, which has been shown to outperform the axial formulation in vision transformers (Heo et al., 2024). Table 17 compares LPE, 2D RoPE, and their combination under both OpenAI pre-trained and scratch initialization. LPE achieves the best retrieval performance across both settings, with a larger performance gap when training from scratch. In addition, combining LPE with 2D RoPE does not provide additional gains.

Table 17: Ablation on positional embedding under pre-trained and scratch initialization.

Init.	Pos. Encoding	M2T R@1 ↑	T2M R@1 ↑
OpenAI	LPE	$\mathbf{58.79 \pm 1.43}$	$\mathbf{60.10 \pm 0.23}$
	2D RoPE	$\underline{54.93 \pm 1.81}$	56.29 ± 1.20
	LPE + 2D RoPE	53.40 ± 3.19	$\underline{56.87 \pm 0.19}$
Scratch	LPE	$\mathbf{34.46 \pm 1.78}$	$\mathbf{41.43 \pm 2.21}$
	2D RoPE	$\underline{27.74 \pm 0.69}$	$\underline{36.03 \pm 0.90}$
	LPE + 2D RoPE	25.61 ± 5.37	35.23 ± 2.18

F.7 Model Capacity

Figure 15 studies the effect of scaling the encoder from ViT-B to ViT-L. For pretrained Meteo-CLIP, ViT-L achieves a lower training loss than ViT-B (0.0490 vs. 0.0802), but a higher test loss (0.73 vs. 0.65) and lower M2T/T2M R@1 (54.24/57.21 vs. 58.79/60.10). We observe a similar trend for Meteo-SigLIP, where scaling to ViT-L reduces the training loss from 0.2395 to 0.1864, while increasing the test loss from 0.56 to 0.61 and reducing M2T/T2M R@1 from 66.07/66.79 to 63.28/61.66. In comparison, scaling the pretrained CLIP baseline from ViT-B to ViT-L slightly improves both retrieval performance and test loss. When Meteo-CLIP is trained from scratch, ViT-L achieves a lower test loss than ViT-B (0.90 vs. 0.95) and improves M2T/T2M R@1 from 34.46/41.43 to 36.68/41.86. This suggests that although increased capacity can benefit Meteo-CLIP and pretrained visual representations provide a useful initialization (Section 5.3), scaling larger pretrained vision-language encoders may require greater adaptation and additional meteorology-language supervision due to the substantial modality gap between RGB images and multivariate meteorological data. These results position EarthAlign as a challenging setting, where simply fine-tuning a larger model does not guarantee better performance.

G LIMITATIONS AND FUTURE DIRECTIONS

Although EarthAlign curates natural event and disaster records from established expert-curated databases and published datasets, their underlying reporting and coverage patterns inevitably shape the geographic, temporal, and event-distribution biases of the dataset. In particular, the United States, Indonesia, and France contain more records than many other countries due to the extensive country-level coverage provided by the NOAA Storm Events Database, IndDE, and FraDE, respectively (Table 9). In addition, MeteoLIP currently represents each sample using a single meteorological snapshot. Future work can extend this setting to multiple timestamps and temporally aggregated inputs, as well as broader Earth-system data products such as operational forecasts and other reanalysis datasets. Incorporating regional or downscaled products with higher spatial resolution may further improve the representation of localized phenomena and enable more comprehensive alignment across temporal scales, spatial resolutions, and Earth-system data sources.

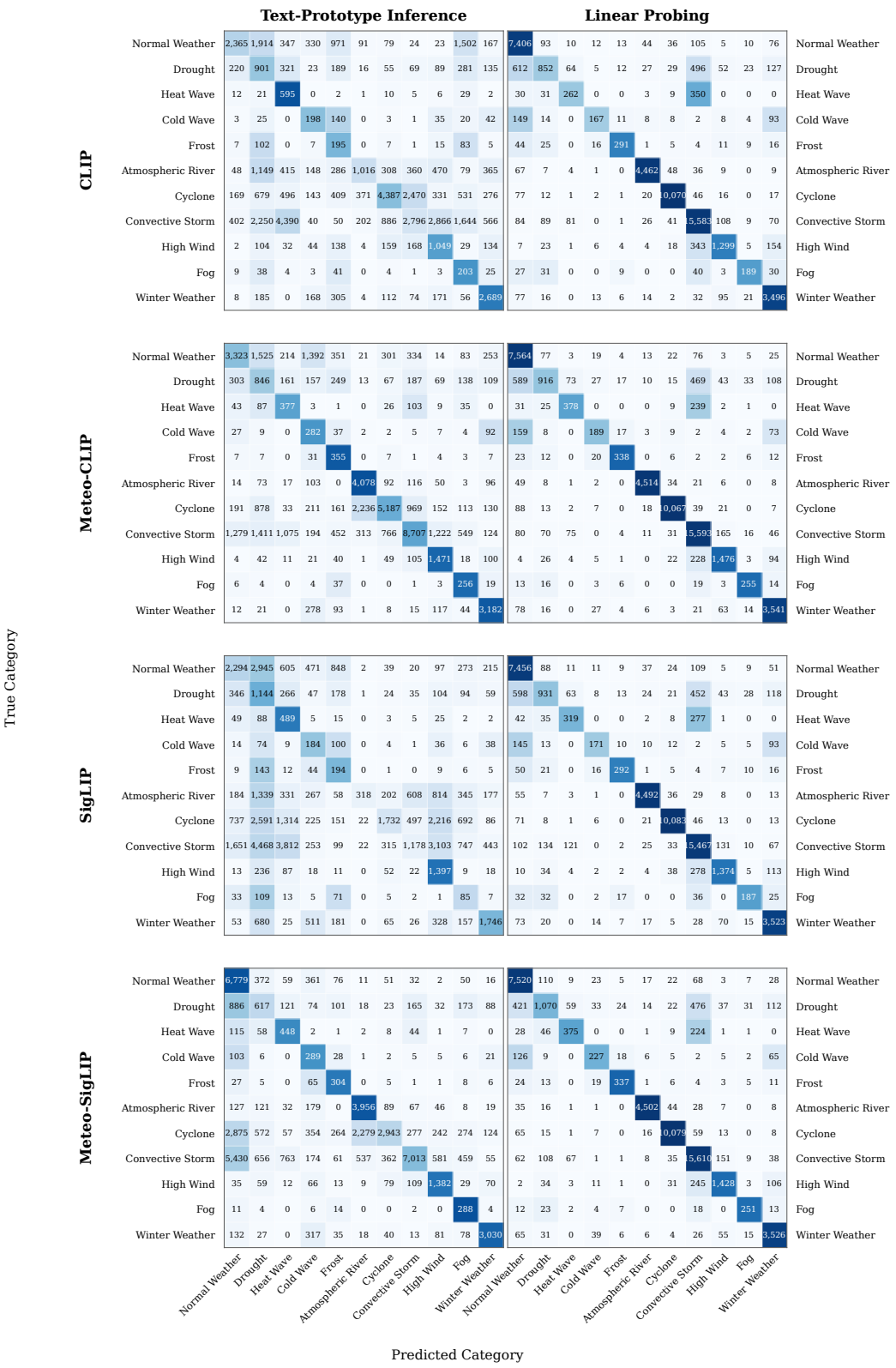
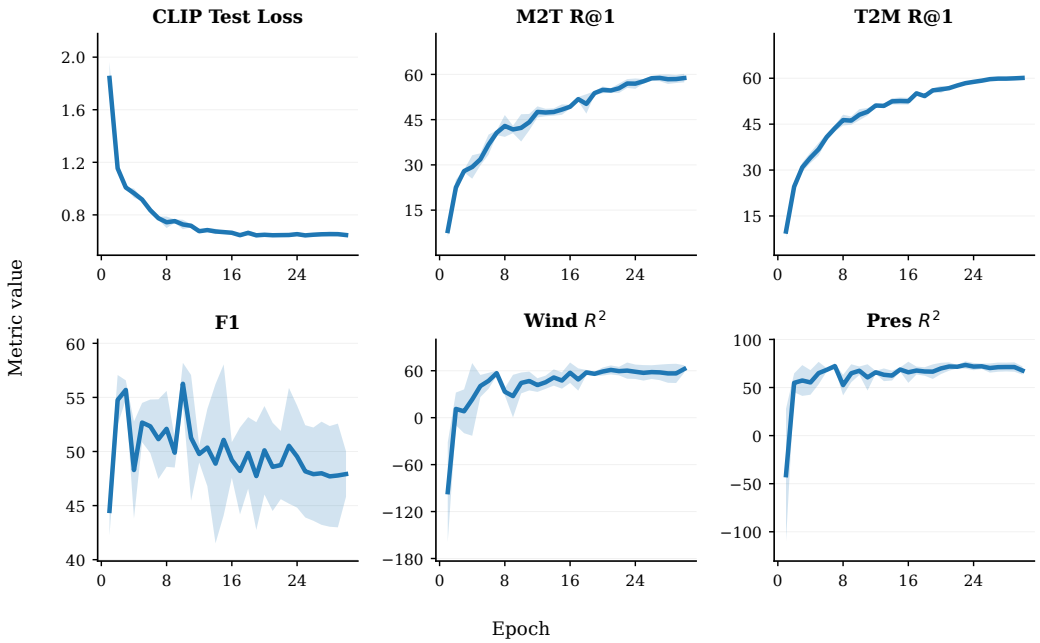
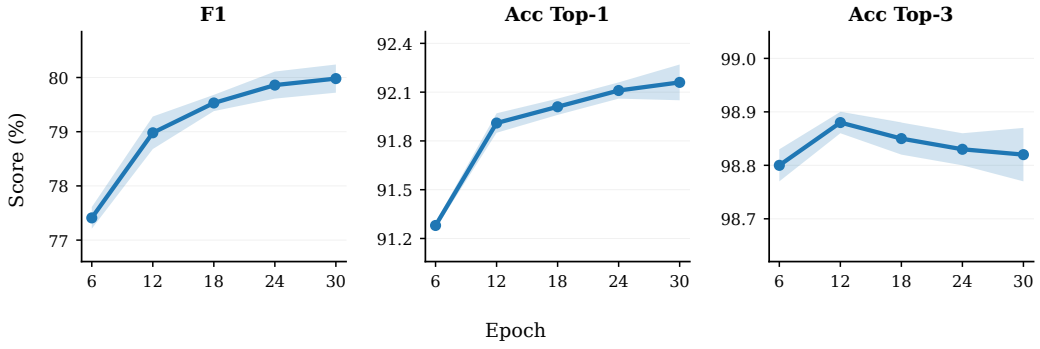


Figure 13: Confusion matrices under *text-prototype inference* and *linear probing*. Cell values report mean counts, while color intensity is normalized by true category to account for class imbalance.



(a)



(b)

Figure 14: Training dynamics of Meteo-CLIP. (a) CLIP test loss and retrieval, classification, and physical variable estimation performance across epochs. (b) Linear-probing classification performance using frozen meteorology encoder checkpoints at epochs 6, 12, 18, 24, and 30.

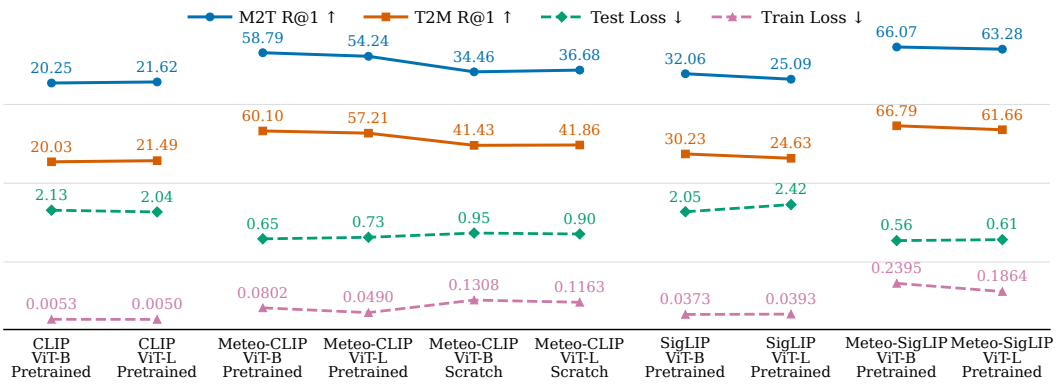


Figure 15: We compare ViT-B and ViT-L encoder variants using pretrained initialization, and additionally compare Meteo-CLIP models trained from scratch. We report M2T and T2M R@1 along with training and test loss to examine how model capacity interacts with initialization and meteorological-language inputs. The reported values are averaged over three training runs.