MeteorPred: A Meteorological Multimodal Large Model and Dataset for Severe Weather Event Prediction

Supplementary Material

1. Dataset Details

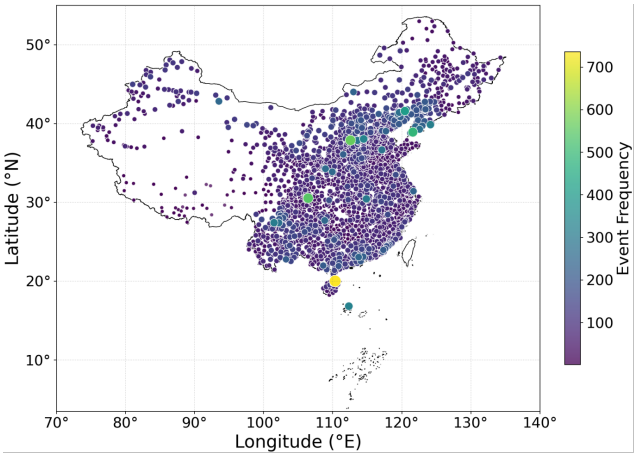
1.1. Meteorological data

In this study, to investigate the impact of atmospheric physical processes in the near-surface layer, troposphere, and stratosphere on warning issuance, we selected five key variables from the ERA5 reanalysis dataset across 37 vertical pressure levels in Tab. 2. Specifically, the 800–1000 hPa range represents the near-surface layer, 200–800 hPa corresponds to the main troposphere, and levels below 200 hPa are associated with the stratosphere.

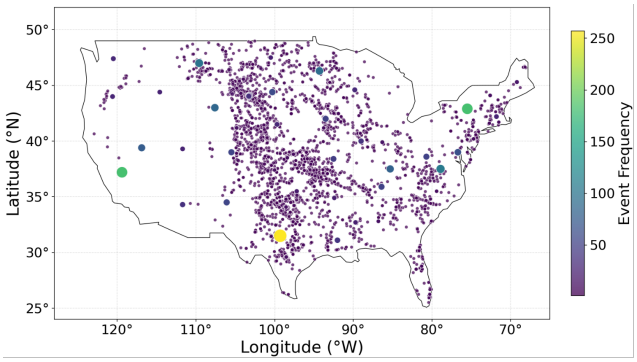
1.2. Severe Weather Event Distribution

Fig. 1 (a) displays the severe weather event data for the China region, which serves as the training and testing set for the model. The dots represent the locations of the events, with color and symbol size corresponding to the regional event frequency, clearly indicating that the densely populated eastern and southeastern regions of China are areas with high frequencies of severe weather.

Fig. 1 (b) depicts the distribution of severe weather events in the US region. Specifically, this US subset consists of 1,000 samples drawn from the NOAA Storm Events database and covers all seasons, with the proportions of each severe-weather type matched to those of the MP-Bench test set (see Tab. 1). This data is designated as the generalization validation set to test the model's ability to generalize to different geographical areas. Similar to the China data, the plot uses color and size to show event frequency, highlighting the Midwestern and Eastern parts of the US as regions prone to high severe weather occurrences.



(a) Spatial distribution of severe weather events across China. Symbol size corresponds to regional severe weather frequency.



(b) Spatial distribution of severe weather events across US. Symbol size corresponds to regional severe weather frequency.

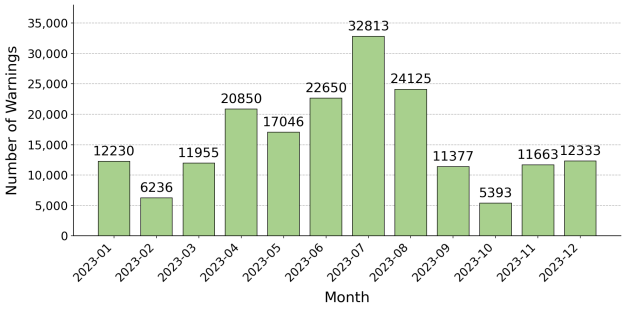
Figure 1. Spatial distribution of severe weather events of China and US.

1.3. Train-Test Split Description

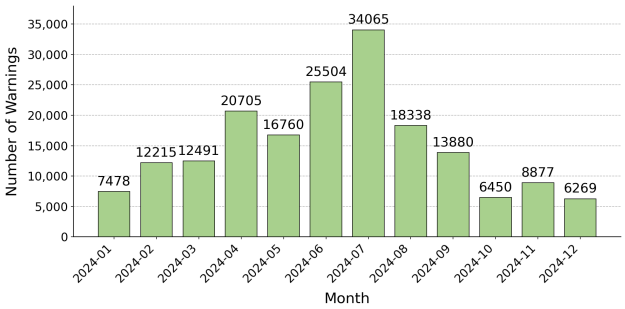
To construct a temporally independent evaluation protocol, we adopt a year-wise dataset split, using 2023 as the training set and 2024 as the test set. Tab. 1 summarizes the distribution of severe weather warnings for both years. Although the absolute event counts vary across years, the relative proportions of different severe weather types remain highly consistent, providing a stable data foundation for training–testing separation. Gale accounts for the largest proportion in both years (2023: 43.54%, 2024: 41.52%), ensuring sufficient samples for model learning and robust evaluation. Rain Storm and Heat Wave show similar ratios across years (22.24% vs. 23.50%; 9.74% vs. 11.91%), enabling reasonable generalization on mid-frequency categories. Cold Wave, Hail, Frost, and Snow Storm remain low-frequency but stable (each $< 5\%$), which is important for evaluating the model's ability to detect rare severe weather events. The number of Normal samples is almost identical between the two years (about 25k), providing a consistent baseline for distinguishing severe vs. normal weather.

1.4. QA types

MP-Bench comprises four types of QA pairs, with Fig. 6–Fig. 9 illustrating representative examples of each type.



(a) 2023 severe weather events bar chart.



(b) 2024 severe weather events bar chart.

Figure 2. Monthly bar charts of severe weather events for the years (a) 2023 and (b) 2024.

Year	Type	Num	Ratio (%)
2023	Gale	92939	43.54
	Rain Storm	47470	22.24
	Heat Wave	20787	9.74
	Cold Wave	11306	5.30
	Frost	7171	3.36
	Hail	5794	2.71
	Snow Storm	3204	1.50
	Normal	24810	11.62
	All	**213481**	**100.0**
2024	Gale	86314	41.52
	Rain Storm	48870	23.50
	Heat Wave	24767	11.91
	Cold Wave	8200	3.94
	Hail	7004	3.37
	Frost	4410	2.12
	Snow Storm	3467	1.67
	Normal	24850	11.95
	All	**207882**	**100.0**

Table 1. Statistics of severe weather warnings (2023–2024).

2. Experiment Settings

In MMLM framework, we adopt four base models, Qwen2.5-VL-7B-Instruct, LLaVA-NeXT-Video-7B, Video-LLaVA-7B, and InternVL3-8B, and attach three plug-and-play modules (DTGF, TGS, and TGCA) for fine-tuning and evaluation. All linear layers in the base models are fine-tuned with LoRA, while DTGF, TGCA, and the fusion layer are trained as additional learnable components. Unless otherwise specified, all reported results are averaged over three independent runs. The detailed training and testing settings are summarized in Tab. 4 and Tab. 5.

3. Supplementary Experiment

3.1. Analysis of Temporal Window Length

Building on the temporal-window ablation described in the main text, we further report detailed results for the T/F, RSW and NSW tasks in the appendix. Using the same 5,000 samples subset and the three temporal windows [t, t+1), [t, t+5], and [t, t+11] as in the MC experiment, we train separate models and evaluate their performance on T/F, RSW and NSW. As shown in Tab. 3, the 12-hour window [t, t+11] consistently achieves the best accuracy across both tasks, confirming that the longer temporal context is beneficial not only for MC but also for regional selection and open-ended description of severe weather events.

4. Physically Consistent in the TGCA Module

To determine the most sensitive pressure levels for five meteorological variables in the ERA5 dataset, we selected seven typical severe weather types (10 samples each) and analyzed them by calculating the average weights across all pressure levels. As shown in Fig. 3, the channel-wise attention learned by the TGCA module automatically concentrates on physically meaningful pressure levels: 500hPa meridional wind for mid-tropospheric trough–ridge patterns, 950 hPa zonal wind for near-surface gale-related flows, 300 hPa temperature for upper-level cold cores and jet structures, and geopotential height and humidity around 825–875 hPa for low-level baroclinicity and moisture supply. This alignment with classical synoptic-scale analysis suggests that the model has discovered physically consistent diagnostics.

5. Physically Consistent in the DTGF Module

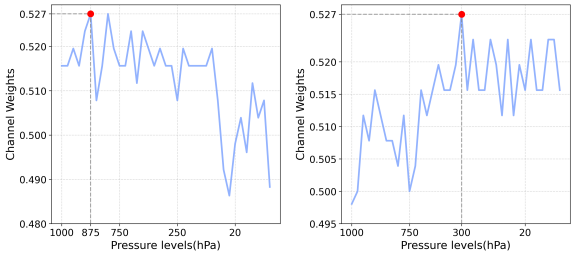
For snow storms (as shown in Fig. 4(b)), DTGF assigns clearly larger positive time-weight differences within the 1–6 h window, while the 9–11 h bins are dominated by negative values. This indicates that severe (red) snowstorm warnings rely more on the rapid intensification during the last 1–6 hours before the event, whereas milder (blue) warnings put relatively higher weights on the earlier 9–12 h evolution. Such a pattern is consistent with the CMA criteria, where red snowstorm warnings are issued when heavy snow ($\geq$ 15 mm) is expected within 6 hours, while blue warnings

Variable	Definition	Unit	Pressure Levels (hPa)
z	geopotential	gpm	1, 2, 3, 5, 7, 10, 20, 30, 50, 70, 100,
u	U-component Wind Speed	m/s	125, 150, 175, 200, 225, 250, 300, 350,
v	V-component Wind Speed	m/s	400, 450, 500, 550, 600, 650, 700, 750,
t	Temperature	K	775, 800, 825, 850, 875, 900, 925, 950,
q	Specific Humidity	kg/kg	975, 1000

Table 2. Summary of the 5 physical variables in the dataset.

Window Length	**T/F Acc ↑**	**RSW Acc ↑**	**NSW Score ↑**
[t , t+1)	67.17	52.71	1.1
[t , t+5]	72.23	54.63	1.6
[t , t+11]	**79.21**	**58.63**	**1.7**

Table 3. Analysis of Temporal Window Lengths (1h, 6h, 12h).



(a) Specific humidity. (b) Temperature.

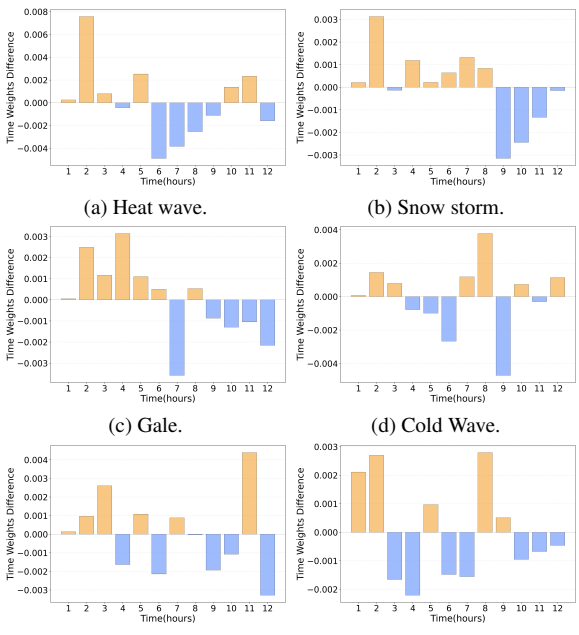
(c) U-component of wind. (d) Geopotential.

Figure 3. Examples of TGCA's Weight distribution patterns. Each subplot represents a type of severe weather event.

correspond to lighter accumulations ($\geq$ 4 mm) over a longer 12-hour window.

For gales (as shown in Fig. 4(c)), DTGF produces strong positive time-weight differences in the 1–4 h window, a pronounced negative segment around 7–9 h, and another positive peak near 11 h. This indicates that severe (red) gale warnings rely primarily on the rapid wind strengthening during the last few hours before the event, while milder (blue) warnings place relatively higher weights on the mid-range 7–9 h evolution. The additional positive peak around 11 h suggests that intense gale cases tend to exhibit stronger early precursors than blue-warning cases. Overall, this temporal pattern aligns well with the CMA criteria, where red gale warnings are issued for gales expected within 6 hours, whereas blue warnings are based on the risk of $\geq$ 6-grade winds within a much longer 24-hour window.

For hail (as shown in Fig. 4(f)), the DTGF module assigns the largest positive time-weight differences to the 1–2 h bins, while the 3–6 h bins are dominated by negative values. This means that severe (red) hail warnings rely much more on the most recent 1–2 h evolution of the storm, whereas milder warnings (e.g., orange) put relatively higher weights on the broader 3–6 h window. This pattern is highly consistent with the CMA operational criteria, where red hail warnings are issued only when hail is highly likely within the next 2 hours, while orange warnings correspond to possible hail within 6 hours.



(a) Heat wave. (b) Snow storm.

(c) Gale. (d) Cold Wave.

(e) Frost. (f) Hail.

Figure 4. Examples of DTGF's weights difference distribution patterns.Each subplot represents a type of severe weather event. The positive bar indicates higher weights for red warnings. The negative bar indicates higher weights for blue warnings.

6. Scoring Rubric with LLM-as-a-Judge

In this section, we detail the rubric used to guide GPT-4o's scoring of NSW answers (see Fig. 10). The rubric asks GPT-4o, acting as a meteorological researcher, to compare the model's output with the reference warnings and assign a score from 0 to 5. Scores are determined primarily by whether the locations, levels (blue, yellow, orange, red), and categories (eight types: Rain Storm, Snow Storm, Gale, Frost, Cold Wave, Heat Wave, Hail, Normal) match those in the reference. A score of 5 corresponds to completely accurate predictions (all locations and all level–category pairs match), while intermediate scores (4–2) reflect varying degrees of partial coverage or minor/more substantial omissions and mismatches. cores of 1 and 0 are reserved for mostly incorrect or completely irrelevant answers, where only a small fraction—or none—of the reference warnings are correctly reproduced.

7. Case Study

To better understand how the model attends to meteorological patterns when it succeeds or fails, we compared 20 rainstorm events with correct and incorrect predictions and visualized their COE fields. Fig. 5 shows contour plots of correct samples (blue) versus incorrect samples (red) across the two attention-feature dimensions, namely magnitude and angle, for MC, T/F and RSW tasks. The peak of the incorrect-sample distribution is shifted toward higher magnitude and angle values compared to the correct-sample distribution, and its contours are more dispersed, indicating that the meteorological features the model focuses on when it makes errors are more complex and variable. In particular, many misclassified cases correspond to compound weather situations (e.g., heavy rainfall accompanied by strong winds), where the model tends to overemphasize certain signals such as wind and incorrectly predicts gale events, even when the ground-truth label is a rainstorm.

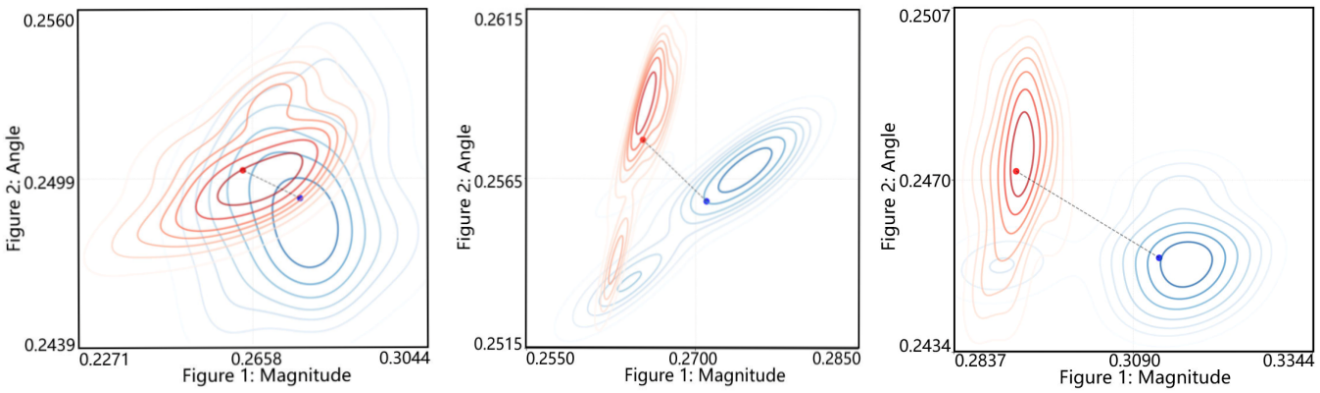


Figure 5. CoE Feature distribution of correct and incorrect sample sets in three Question types: (a) MC questions, (b) T/F questions, (c) RSW questions. Blue and Red distributions represent the correct and incorrect samples. Dataset used in this figure is testing set and model used in the figure is Qwen2.5-VL-7B-Instruct.

True/False Question

As a professional meteorologist, please analyze the provided ERA5 dataset and determine whether **Maqu County (Coordinates: [34.00°N, 102.07°E])** is currently experiencing severe weather. Respond with either **"Yes"** or "No".

Figure 6. An example of True/False question prompt for training and testing.

Regional Severe Weather Question

As a professional meteorologist, please analyze the provided ERA5 data and assess the likelihood of severe weather events occurring in **Fuzhou City (Coordinates: [26.08°N, 119.30°E])**. Please identify which types of severe weather events may occur, selecting from the following categories:

Rain Storm, Snow Storm, Gale, Cold Wave, Heat Wave, Frost, Hail.

Figure 7. An example of RSW question prompt for training and testing.

Multiple Choice Question

As a professional meteorologist, please identify the severe weather events that occurred in **Beijing (Coordinates:[39.90°N,116.40°E])** based on the input data. Please select only one applicable option from the following options:

[Rain Storm]
A1: Blue-level A2: Yellow-level A3: Orange-level A4: Red-level

[Snow Storm]
B1: Blue-level B2: Yellow-level B3: Orange-level B4: Red-level

[Gale]
C1: Blue-level C2: Yellow-level C3: Orange-level C4: Red-level

[Cold Wave]
D1: Blue-level D2: Yellow-level D3: Orange-level D4: Red-level

[Heat Wave]
E1: Yellow-level E2: Orange-level E3: Red-level

[Frost]
F1: Blue-level F2: Yellow-level F3: Orange-level

[Hail]
G1: Orange-level G2: Red-level

[Normal Conditions]
H1: No warnings issued

Figure 8. An example of multiple choice question prompt for training and testing.

National Severe Weather Question

As a professional meteorologist, you are tasked with analyzing the provided dataset to identify and characterize any severe weather events that have occurred across **China's** administrative divisions.
Please focus on the following regions with their respective coordinates:

Hebei(Coordinates:[38.04°N,114.51°E]),
Shanxi(Coordinates:[37.87°N,112.55°E]),
Liaoning(Coordinates:[41.80°N,123.50°E]),
Jilin(Coordinates:[43.90°N,125.33°E]),
Heilongjiang(Coordinates:[45.76°N,126.64°E]),. . . . . .

Determine what kind of severe weather occurred in each region. Output only the area where severe weather occurs For each detected event, use the following structured format: [Region Name] issues a [Event Type] [Severity Level].

Definitions:
1. Region Name (Administrative Divisions)
2. Event Type (Rain Storm/Snow Storm/Gale/Cold Wave/Heat Wave/Frost/Hail)
3. Severity Level (Blue/Yellow/Orange/Red)

Figure 9. An example of NSW question prompt for training and testing.

GPT-4o's scoring criterion

You are a meteorological researcher. Your task is to compare the reference (ground-truth) answers with the outputs from a multi-modal meteorological model, assign a score from 0 to 5, and explain your reasoning. Each answer consists of multiple warning entries separated by commas, each formatted as: <Location><Level><Category>. Categories (8 types): Rain Storm, Snow Storm, Gale, Frost, Cold Wave, Heat Wave, Fog No Severe Weather. Levels (4 colors): Blue, Yellow, Orange, Red. Use the following open-ended scoring rubric (0-5) and explain why you chose that score:

5 – Completely accurate:

- All locations match exactly.
- For each warning, Level and Category both match perfectly.

4 – Accurate but slightly incomplete:

- Locations match exactly.
- Categories match perfectly; Levels partially match.

3 – Partially accurate:

- Most locations match, with minor omissions or discrepancies.
- Categories partially match; Levels partially match.

2 – Contains some errors:

- Some reference locations are missing in the model output.
- Categories partially match; Levels partially match.

1 – Mostly incorrect:

- Most reference locations are missing or incorrect.
- Only a small fraction of entries match in Location, Level, or Category.

0 – Completely incorrect or irrelevant:

- No locations match.
- No entries match in Level or Category.

Please score the following:
Standard answer: [Insert standard answer here]
Multimodal meteorological foundation model: [Insert student's answer here]
Please score the model's output according to the criteria and explain the reasoning for the score given.

Figure 10. GPT-4o's scoring criterion.

Parameter	Value
Model Parameters	
model_name_or_path	/model_pre-trained_weights
adapter_name_or_path	-
trust_remote_code	true
Method Parameters	
stage	sft
do_train	True
finetuning_type	lora
quantization_method	bitsandbytes
template	specific_model_templates
flash_attn	auto
dataset	train_data
cutoff_len	9500
max_samples	220000
preprocessing_num_workers	16
LoRA Parameters	
lora_rank	8
lora_alpha	16
lora_dropout	0
lora_target	all
additional_target	gating_mlp, linear, text_proj, mlp_channel
Training Parameters	
per_device_train_batch_size	2
gradient_accumulation_steps	8
learning_rate	0.00005
num_train_epochs	3.0
lr_scheduler_type	cosine
max_grad_norm	1.0
warmup_steps	0
packing	False
report_to	none
bf16	True
optim	adamw_torch
ddp_timeout	180000000
include_num_input_tokens_seen	True
ddp_find_unused_parameters	False
seed	42
Output Parameters	
output_dir	/path/to/output
logging_steps	8
save_steps	200
plot_loss	True
overwrite_output_dir	True
save_only_model	False

Table 4. Model Training Parameters

Parameter	Value
Model Parameters	
model_name_or_path	/model_pre-trained_weights
adapter_name_or_path	/lora-trained_weights
trust_remote_code	true
Method Parameters	
stage	sft
do_train	False
finetuning_type	lora
quantization_method	bitsandbytes
template	specific_model_templates
flash_attn	auto
dataset_dir	data
eval_dataset	test_data
cutoff_len	9500
max_samples	100000
preprocessing_num_workers	16
Evaluation Parameters	
per_device_eval_batch_size	4
predict_with_generate	True
max_new_tokens	512
top_p	0.7
temperature	0.95
do_predict	True
seed	42
Output Parameters	
output_dir	/path/to/output

Table 5. Model Evaluation Parameters