Environmental Data Science (2026), 5: e19, 1–11
doi:10.1017/eds.2026.10050

CAMBRIDGE UNIVERSITY PRESS

DATA PAPER

SynopticBench: evaluating vision-language models on generating weather forecast discussions of the future

Timothy Higgins[1], Antonios Mamalakis[2,3] and Chirag Agarwal[3]

[1]Environmental Institute, University of Virginia, USA
[2]Department of Environmental Sciences, University of Virginia, USA
[3]School of Data Science, University of Virginia, USA
Corresponding author: Timothy Higgins; Email: hay3fm@virginia.edu

Received: 11 May 2026; **Revised:** 11 May 2026; **Accepted:** 04 June 2026

Keywords: multimodality; numerical weather prediction; vision language model; weather forecasting

Abstract

Recent advances in visual-language models (VLMs) have led to significant improvements in a plethora of complex multimodal tasks like image captioning, report generation, and visual perception. However, generating text from meteorological data is highly challenging because the atmosphere is a chaotic system that is rapidly changing at various spatial and temporal scales. Given the complexity of atmospheric phenomena, it is critical to verifiably quantify the effectiveness of existing VLMs on weather forecasting data. In this work, we present SynopticBench, a high-quality dataset consisting of 1,367,041 text samples of Area Forecast Discussions created by the National Weather Service over the continental United States paired to images of 500 mb geopotential height, 2 m temperature, and 850 mb wind velocity in weather forecasts. We also present Synoptic Phenomena Alignment and Coverage Evaluation (Space), a novel evaluation framework that can be used to effectively estimate the quality of text descriptions of synoptic weather phenomena. Extensive experiments on generating forecast discussions using state-of-the-art VLMs show the sensitivity of existing evaluation metrics in this domain and enable further exploration into synoptic weather and climate text generation.

Impact Statement

To the best of our knowledge, this is the first study to generate text discussions from images of numerical weather forecasts. We introduce (i) the largest benchmark dataset, comprising 1,367,041 image-text samples, to run experiments that explore the potential to generate text discussions from numerical weather prediction; and (ii) a methodology for evaluating generated text discussions of synoptic weather and climate.

1. Introduction

Recent advances in vision-language models (VLMs) have enabled substantial progress on complex multimodal tasks such as image captioning, document understanding, visual question answering (QA), and multimodal summarization. As these models improve in capability and are increasingly deployed or

This research article was awarded Open Data and Open Materials badges for transparent practices. See the Data Availability Statement for details.

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

proposed for deployment in high-stakes settings, there is a growing need for evaluation frameworks that go beyond surface-level text similarity and instead quantify whether generated language is faithful, spatially grounded, and domain-relevant.

To this end, weather forecasting is a particularly consequential and technically challenging high-stakes domain. Numerical weather prediction (NWP) is deeply integrated into business operations, agriculture, government decision-making, public safety, and daily life for billions of people (Jones et al., 2000; World Meteorological Organization, 2015; Jain and Jain, 2017; Uccellini and Ten Hoeve, 2019; Ukhurebor et al., 2022). Vast observation networks and reanalysis products have enabled continual improvements in model development, data assimilation, and assessments of predictive reliability (Randles et al., 2017; Hersbach et al., 2020). However, atmospheric dynamics are chaotic and span a wide range of spatial and temporal scales, which is why forecast verification relies on many complementary skill metrics (e.g., Anomaly Correlation Coefficient, RMSE, MAE, and skill scores) to characterize different aspects of performance (Brady and Spring, 2021). In contrast, despite the operational importance of narrative forecasting products, methods to verifiably quantify the reliability of textual representations of atmospheric states and evolution **remain limited**.

Recent years have seen a growing body of work that explores multimodal learning for atmospheric data, including fine-tuning frontier VLMs and dataset creation for meteorological QA and event understanding. Examples include CLLMate (Li et al., 2024), which pairs extreme-event news with reanalysis for QA; RadarQA (He et al., 2025), which introduces radar annotations and heuristics for VLM fine-tuning; and hazard-focused datasets such as SEVIR (Veillette et al., 2020) and GridRad-Severe (Murphy et al., 2023). Additional efforts pair structured meteorological inputs with textual outputs for QA-oriented benchmarks, such as MeteorPred (Tang et al., 2025), Climate IQA (Chen et al., 2025), Zephyrus (Varambally et al., 2025), and WeatherQA (Ma et al., 2024). While these studies demonstrate the promise of multimodal models for atmospheric applications, direct comparisons across benchmarks are often confounded by rigid constraints on target outputs, spatial scales, forecast horizons, and training data choices. Moreover, many evaluations emphasize QA accuracy or rely heavily on traditional text metrics (BLEU, ROUGE, METEOR, BERTSCORE) and/or LLM-judge scoring, which can be insufficient for assessing whether generated discussions correctly identify synoptic phenomena, capture their spatial extent, and reflect scale-dependent relevance to a forecast domain.

1.1. Present work

To address all the aforementioned gaps, we introduce a new benchmark and evaluation methodology designed specifically for the generation and verification of synoptic-scale forecast discussions. We present SYNOPTICBENCH, a large-scale dataset comprising 1,367,041 National Weather Service (NWS) Area Forecast Discussions paired with forecast images of 500 mb geopotential height, 2 m temperature, and 850 mb wind velocity over the continental United States. In contrast to existing benchmarks, SYNOPTICBENCH is the first benchmark to evaluate NWP meteorological samples across both mesoscale and synoptic spatial scales for all weather (see Table 3). Critically, we also propose Synoptic Phenomena Alignment and Coverage Evaluation (SPACE), an evaluation framework that explicitly accounts for the scale-dependent nature of atmospheric phenomena and the differing spatial precision required across variables (e.g., precipitation versus pressure patterns). SPACE is designed to evaluate distinct phenomena separately while incorporating spatial coverage and relevance, enabling more faithful estimation of a VLM's skill in generating meteorological text. Building on SYNOPTICBENCH, we train models for synoptic discussion generation and fine-tune widely used VLMs. Our results show that the ability of the models to conduct physical reasoning is improved from fine-tuning. The key goal of SYNOPTICBENCH is to advance the study and understanding of VLMs in weather forecasting. Therefore, we share the dataset, metric, and fine-tuned weights of a range of VLMs to support reproducible research and more rigorous, domain-appropriate evaluation of VLMs in weather forecasting (Table 1).

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

Table 1. *Comparison between* SynopticBench *and existing work using multimodal datasets for atmospheric data*

Dataset	Meteorological data	Text size	Spatial scale	Target
GridRad-Severe (Murphy et al., 2023)	RS	x	SC, MC, MS	Extreme events
SEVIR (Veillette et al., 2020)	RS	x	MS	Extreme events
WeatherQA (Ma et al., 2024)	Reanalysis	8000	MS	Extreme events
ClimateIQA (Chen et al., 2025)	Reanalysis	762,120	MS, Synoptic	Extreme events
CLLMate (Li et al., 2024)	Reanalysis	41,000	MS	Extreme events
RadarQA (He et al., 2025)	RS	69,000	Mesoscale	All weather
MP-Bench (Tang et al., 2025)	Reanalysis	421,363	MS	Extreme events
Zephyrus Bench (Varambally et al., 2025)	Reanalysis	2158	MS, Synoptic, Global	All weather
SynopticBench	**NWP**	**1,367,041**	**MS and Synoptic**	**All weather**

Note. Here, we present the type of meteorological data (RS→Remote sensing; NWP→Numerical weather prediction), the number of text samples, the relevant spatial scales (SC→Single-cell; MC→Multi-cell; MS→Mesoscale), and the target of each study. SynopticBench is the first large-scale benchmark that generates text from NWP and contains 1,367,041 text samples.

2. SynopticBench

Here, we detail the dataset building pipeline (Section 2.1), data collection (Section 2.2), data preprocessing (Section 2.3), and our proposed Space metric (Section 2.5).

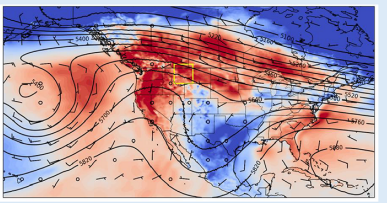
2.1. Dataset building pipeline

One of the largest bottlenecks that stunts progress for multimodal models for atmospheric data is the scarcity of high-quality text that can be paired with numerical data. We propose using Area Forecast Discussions (AFDs) created by the NWS as a potential solution to the multimodal data problem. The AFD text data are publicly available and were extracted from the Iowa Environmental Mesonet (https://mesonet.agron.iastate.edu/wx/afos/list.phtml). AFDs exist at NWS stations across the United States and are typically issued multiple times per day. They reliably describe weather forecasts to an exceptionally high degree of detail, often informing the public about the weather a week in advance (National Research Council, 2006). In addition, they describe numerous physical variables, vertical levels, spatial scales, and processes. AFDs also typically include high-quality descriptions of the temporal evolution of weather phenomena, creating potential for training models to generate text that describes how atmospheric data changes over time.

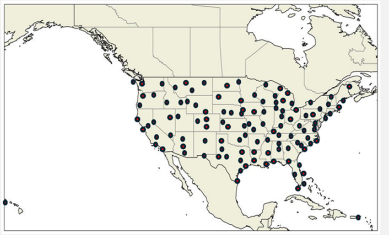
2.2. Dataset collection

We create SynopticBench by pairing the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) model (National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce, 2015) to AFDs. This dataset contains GFS forecasts issued four times per day from January 1, 2016, to December 31, 2025 (14,159 total forecasts) paired with **1,367,041** AFD text samples. Each GFS forecast consists of 2 m temperature, precipitation rate, 500 mb geopotential height, 850 mb zonal wind, and 850 mb meridional wind that is matched to the nearest AFD in time (always within several hours) for each location. We pair each forecast with 117 text samples, as there are 117 NWS station locations used in the study (*Figure 1*). This creates consistency among the text samples for descriptions of the primary synoptic features occurring over the US for a given forecast.

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press



Prompt: Analyze these weather charts showing mean 2 meter temperature over the 2-day period (shaded contours), 500 mb geopotential height (unshaded contours), and 850 mb wind velocity (wind barbs) for the Great Falls, Montana forecast region and generate a detailed forecast discussion of the large-scale features relevant to the area. The forecast region is shown in the yellow box.
Answer: "High pressure at the surface and aloft will remain centered inland of the West Coast, maintaining a dry and stable northwest flow aloft over the region."



Sample 1: Great Falls, MT discussion, January 1st, 2016
Sample 2: Denver, CO discussion, January 1st, 2016
...
Sample 117: Boston, MA discussion, January 1st, 2016
Sample 118: Great Falls, MT discussion, January 2nd, 2016
...
Sample 1,367,041: Boston, MA discussion, December 31st, 2025

Figure 1. *An example case of a single sample from the training set (top panel). Each training sample image has a yellow box indicating the location of the discussion. The example answer is a filtered AFD. All of the locations used in the discussions are shown in the bottom panel. The format of the training samples is also shown, with 117 AFDs matched to each forecast.*

2.3. Preprocessing

Before we begin training and evaluating VLMs, we apply several data preprocessing steps to both AFDs and GFS forecasts. During preprocessing, we first discard all lead times beyond 48 hours. Although AFDs primarily align with GFS forecasts, they are sometimes based on other models as well, which diverge as the lead time increases due to the chaotic nature of the atmosphere. The spatial domain is cropped from 15°N to 65°N and 60°W to 160°W. We plot mean 2 m temperature anomalies in shaded contours along with mean 500 mb geopotential height in line contours and 850 mb wind velocity in the form of wind barbs. A yellow 5° × 5° box is centered on the location of each discussion that is paired to an image.

Keeping the goal of matching text that can be represented in the images, we preprocessed the information in the AFDs for the experiments. AFDs are often informed by GFS forecasts but can use information from a variety of different models. We remove sentences including any reference to other forecast models or synoptic pressure systems that do not exist at the scale or vertical level that the images show. We also only use sentences including keywords that do describe synoptic pressure systems and temperature anomalies including *trough, ridge, low pressure, high pressure, cold front, and warm front.* The full list of inclusion and exclusion keywords is included in the Supplementary Appendix. When evaluating synoptic phenomena with lead times up to 48 hours, we can also expect robust consistency of large-scale patterns. The information described in AFDs occurs chronologically, often including the day of the week on which different phenomena occur. Once a day of the week that is more than two days after the issue day is mentioned, all text that follows is removed to stay consistent with the forecast lead times.

2.4. Baselines

We used four different baselines to add perspective to the results. Gemini-3.1 Pro is used as a baseline for current state-of-the-art models. The nearest neighbor baseline is created by first computing the SSIM scores between each test set image and all training set images. The text paired with the training set image with the highest SSIM score is used as the text for each sample in the baseline. The climatology is created by choosing text from a random sample in the training set from the same month and location as each sample in the test set. The blind LLM baseline strips the image tokens from test before running LLaMA-3.2-11B.

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

2.5. Synoptic *phenomena* alignment and coverage evaluation (SPACE) metric

Traditional metrics for text evaluation are not sufficient for text describing atmospheric processes because they do not include any physical reasoning. The purpose of SPACE is to evaluate the ability of the VLM to generate text that predicts the correct phenomenon in the target location. Evaluating location accuracy from text can be a complex task because many terms may differ entirely but represent similar locations (e.g., The Rockies versus Colorado) and the precision of location terms may be more or less acceptable depending on the scale of the phenomenon (e.g., "trough moving through the Denver area" vs. "trough moving through Colorado" is preferred more than "rain in the Denver area" vs. "rain in Colorado"). To accommodate these relations, we introduce a spatial hierarchy tree specifically created for evaluating text describing synoptic phenomena. The hierarchy comprises three levels that capture terms commonly used to detail the locations of synoptic weather phenomena in North America (see *Figure 2* for an example of a small portion of the hierarchy tree).

SPACE can be applied toward various types of synoptic weather phenomena that have positive and negative phase categorizations (e.g., high/low pressure systems, warm/cold fronts, wet/dry events). Next, to demonstrate the methodology of SPACE, we use pressure systems as an example. We begin SPACE by matching pressure terms to locations. If "low pressure," "high pressure," "trough" (excluding instances of "shortwave trough"), or "ridge" exists in the text, the term is first matched to any location that follows in the sentence. For a sample of text, this creates a list of pressure systems matched to different locations. Pressure terms are then matched to other instances for locations that are equal or related (see connections in *Figure 2*). Locations with a common relative can also match with the exception of those that cover a large portion of the continent (see Supplementary Appendix). This logic creates groups of pressure objects for both the generated and the AFD text sample. Objects can be grouped for single discussions (SPACE-local) or aggregated across all locations (SPACE-aggregate) for a single forecast. If any object from a group has the same pressure sign and a location related to any object in the corresponding text sample, the object groups form a match. We consider "trough" to be synonymous with "low pressure," and "ridge" to be synonymous with "high pressure." The match score, s_m, is a measure of the quality of the matches and is defined below:

$$s_m = 1 - \left| \frac{m_L^{\text{pred}}}{m_L^{\text{pred}} + m_H^{\text{pred}}} - \frac{m_L^{\text{ref}}}{m_L^{\text{ref}} + m_H^{\text{ref}}} \right|, \tag{2.1}$$

where m_L^{pred} is the number of matching *negative phase* objects in the predicted text, m_L^{ref} is the number of matching *negative phase* objects in the reference text, m_H^{pred} is the number of matching *positive phase*

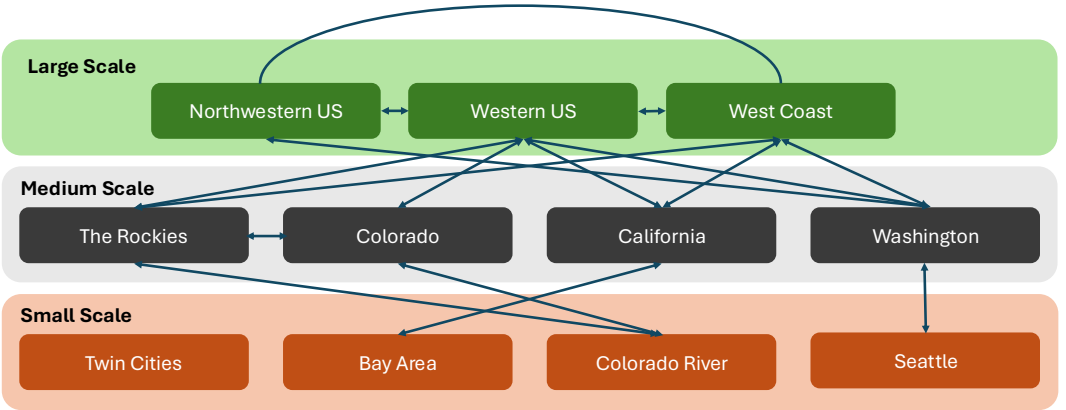


Figure 2. *Several examples of matching large- (green), medium- (purple), and small-scale (orange) location keywords. Blue lines indicate the potential matches that these locations would make if found in the predicted or reference text.*

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

objects in the predicted text, and m_H^{ref} is the number of matching *positive phase* objects in the reference text. The coverage ratio, r_c, is the percentage of pressure objects that are included in matches and is defined as:

$$r_c = \frac{\sum \left(\left(m_L^{\text{pred}} + m_H^{\text{pred}} \right) + \left(m_L^{\text{ref}} + m_H^{\text{ref}} \right) \right)}{n_L + n_H}, \tag{2.2}$$

where n_L is the total number of *negative phase* objects and n_H is the total number of *positive phase* objects in the combined predicted and observed text. The coverage ratio is designed to penalize the generated text for creating hallucinations and for failing to mention key terms. The final SPACE score is calculated as the product of the matching score and coverage ratio, that is, $s = s_m \cdot r_c$.

3. Results

3.1. Traditional metrics are not sufficient

Base model versions for LLaVA-v1.5-7B, LLaVA-v1.5-13B, Qwen2-VL-7B, and LLaMA-3.2-11B-Vision were run on the test set. We first evaluate each of the four base models, fine-tuned models, and baselines on traditional skill metrics (Table 2). LoRA fine-tuning improved performance on all traditional metrics except METEOR for LLaMA-3.2-11B-Vision. The climatology baseline had the highest METEOR score, while fine-tuned Qwen2.5-VL-7B had the highest Bertscore, ROUGE-L, and F1 scores. Gemini-3.1-Pro had the lowest Bertscore, ROUGE-L score, METEOR score, and F1 score while having the highest LLM-as-a-judge score. The climatology baseline used text from real NWS AFDs issued from the same exact station, which may have allowed it to have relatively high scores for traditional metrics by using similar discussion styles. Despite slight variations in the performance of all models, all scores were low, indicating that the language in the generated discussions had little overlap with the NWS discussions. The traditional scores cannot determine the models' ability to discuss the primary relevant features in the images, which creates a need for further evaluation.

Table 2. *Traditional metrics for base models, fine-tuned models, and baselines are shown*

Model	Bertscore	ROUGE-L	METEOR	F1	Gemini-2.5-Flash
LLaVA-v1.5-7B-base	.7143 ± .0001	.1091 ± .0001	.1327 ± .0001	.1471 ± .0002	.1565 ± .0031
LLaVA-v1.5-13B-base	.7192 ± .0001	.1086 ± .0001	.1376 ± .0001	.1498 ± .0002	.1896 ± .0032
LLaMA-3.2-11B-Vision-base	.7051 ± .0001	.1043 ± .0002	.1212 ± .0002	.1313 ± .0002	.2623 ± .0041
Qwen2.5-VL-7B-base	.7136 ± .0001	.1083 ± .0001	.1447 ± .0001	.1591 ± .0003	.2566 ± .0035
LLaVA-v1.5-7B-LoRA	.7814 ± .0002	.1629 ± .0003	.1353 ± .0003	.1943 ± .0004	.2426 ± .0048
LLaVA-v1.5-13B-LoRA	.7858 ± .0002	.1663 ± .0003	.1407 ± .0003	.2010 ± .0004	.2472 ± .0052
LLaMA-3.2-11B-Vision-LoRA	.7773 ± .0001	.1530 ± .0003	.1174 ± .0003	.1867 ± .0003	.2757 ± .0047
Qwen2.5-VL-7B-LoRA	**.7927** ± .0002	**.1681** ± .0003	.1601 ± .0004	**.2184** ± .0004	.2824 ± .0053
Gemini-3.1-Pro	.6913 ± .0004	.0703 ± .0005	.1456 ± .0008	.1030 ± .0009	**.4185** ± .0071
Nearest Neighbor	.7662 ± .0001	.1194 ± .0002	.1395 ± .0002	.1748 ± .0003	.1785 ± .0047
Climatology	.7776 ± .0001	.1434 ± .0002	**.1632** ± .0003	.2035 ± .0003	.2358 ± .0049
LLama-3.2-11B (blind)	.7257 ± .0001	.1101 ± .0002	.1503 ± .0002	.1544 ± .0003	.2574 ± .0054

Note. We calculate the mean and standard error of the mean Bertscore, ROUGE-L, METEOR, F1, and LLM-judge scores across samples within the test set. The range of possible scores for all metrics is 0–1, with 1 *being a perfect score.*
The bold values are the highest value for each column.

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

3.2. SPACE scores can reduce evaluation uncertainty

Although SPACE could be used for various types of weather phenomena, we demonstrate the utility of SPACE on pressure systems in this study. Here, we show how SPACE scores can help the user understand how each model performs in identifying the correct relevant pressure systems for the area and referring to them in the correct locations. SPACE-aggregate uses all locations at a given time to create a large sample size of pressure objects matched to specific locations for evaluation. It is useful for understanding the model's ability to determine the general large-scale pressure features that occur in the forecast. SPACE-local uses a single location at a given time to create a much smaller sample size of pressure objects but can be useful for understanding the model's ability to discuss the features impacting a specific location. All SPACE scores for pressure systems are shown in Table 3. SPACE-aggregate scores are considerably higher than SPACE-local scores because the increased sample size of text increases the chances that relevant locations for synoptic phenomena are mentioned in both the generated text and the AFDs. Theoretically, a model that always randomly guesses between high pressure and low pressure would have a match score of roughly 0.5 without any knowledge of climatology, which indicates that the fine-tuned models likely all have some skill in distinguishing between high and low pressure. The local coverage ratios (r_c) are lower than the match scores for every model, demonstrating that predicting the phenomena in the correct location is a more challenging task for these VLMs than accurately describing the correct polarity of synoptic phenomena.

The Qwen2.5-VL-7B fine-tuned model had the highest SPACE-local scores, while Climatology had the highest SPACE-aggregate scores. The fine-tuned models had higher SPACE-local scores than all of the base models and baselines. The SPACE-aggregate scores for the fine-tuned models were generally higher than those of the base models with the exception of LLaMA-3.2-11B-Vision. The improvement of SPACE-aggregate scores over base models is impacted by both improved coverage ratios and match scores. The SPACE scores often agree with the traditional metrics, with the fine-tuned Qwen-VL-7B configuration generally having the highest scores. One clear exception is Gemini-3.1-Pro, which has some of the lowest scores when evaluated on traditional metrics and fairly high SPACE scores. The SPACE scores and traditional metrics also suggest that increasing the parameters from 7 billion to 13 billion in LLaVA does not drastically improve generated discussion quality and the fine-tuning approach is far more effective.

3.3. Case studies illustrate the advantages of SPACE

An example of the evaluation of both SPACE and traditional metrics in two different cases is shown in *Figure 3*. In Case 1, there is a high-pressure system over the southwestern United States, which is the most influential pressure system relevant to the station location (Tuscon, Arizona). The NWS discussion mentions high pressure in the region, and the generated VLM text also discusses high-pressure systems in the same region. This leads to a perfect SPACE-local score for pressure systems. The traditional metrics for this case were generally close to average for this model. In Case 2, there is a high-pressure system over much of the southeastern US, which is the most influential pressure system to the station location (Columbia, South Carolina). The NWS discussion mentions a ridge in the forecast area while the VLM's generated text discusses a trough in the southeastern United States. This leads to a SPACE score of 0. Despite having a SPACE-local score of 0, the traditional metrics were unable to show a clear difference in the quality of both cases. Case 2 also shows some instances of hallucinations in which the model discusses phenomena that it could not possibly see based on the image (e.g., 1 inch of snow), which does not penalize the SPACE score because SPACE only evaluates one type of phenomenon at a time. This case study demonstrates a clear example of the reliability of SPACE scores relative to traditional metrics for the evaluation of individual types of phenomena in weather discussions. Several additional cases are shown in the Supplementary Appendix.

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

Table 3. SPACE was used for synoptic pressure system evaluation for base models, fine-tuned models, and baselines

Model	s_s^{loc}	s_m^{loc}	r_c^{loc}	s_s^{agg}	s_m^{agg}	r_c^{agg}
LLaVA-v1.5-7B-base	.0976 ± .0008	.4592 ± .0056	.2158 ± .0013	.4177 ± .0008	.5039 ± .0056	.9467 ± .0013
LLaVA-v1.5-13B-base	.1553 ± .0007	.4404 ± .0061	.3615 ± .0012	.4698 ± .0007	.5044 ± .0061	.9317 ± .0012
LLaMA-3.2-11B-Vision-base	.1194 ± .0008	.6296 ± .00013	.1936 ± .0012	.6492 ± .0008	.6690 ± .0013	.9648 ± .0012
Qwen2.5-VL-7B-base	.0985 ± .0018	.5072 ± .0032	.1950 ± .0031	.5831 ± .0018	.6221 ± .0032	.9279 ± .0031
LLaVA-v1.5-7B-LoRA	.2235 ± .0015	.6575 ± .0023	.3364 ± .0018	.6486 ± .0015	.6714 ± .0023	.9635 ± .0018
LLaVA-v1.5-13B-LoRA	.2604 ± .0015	.6601 ± .0021	**.3900** ± .0018	.6410 ± .0015	.6629 ± .0021	.9648 ± .0018
LLaMA-3.2-11B-Vision-LoRA	.2424 ± .0016	.6554 ± .0023	.3654 ± .0019	.6091 ± .0016	.6397 ± .0023	.9488 ± .0019
Qwen2.5-VL-7B-LoRA	**.2626** ± .0015	**.6744** ± .0020	.3845 ± .0018	.6294 ± .0015	.6498 ± .0020	.9663 ± .0018
Gemini-3.1-Pro	.2005 ± .0037	.5825 ± .0034	.3445 ± .0058	.6685 ± .0037	.6775 ± .0034	**.9864** ± .0058
Nearest Neighbor	.1569 ± .0010	.5485 ± .0015	.2866 ± .0015	.6187 ± .0010	.6393 ± .0015	.9661 ± .0015
Climatology	.2009 ± .0016	.5655 ± .0012	.3554 ± .0018	**.6808** ± .0016	**.6900** ± .0012	.9856 ± .0018
Llama-3.2 (blind)	.1745 ± .0010	.4967 ± .0014	.3542 ± .0017	.5690 ± .0010	.5794 ± .0014	.9813 ± .0017

Note. We calculate the mean and standard error of the SPACE-local and SPACE-aggregate scores (s_s), match scores (s_m), and coverage ratios (r_c) within the test set. The range of possible scores for all metrics is 0 to 1, with 1 being a perfect score.
The bold values are the highest value for each column.

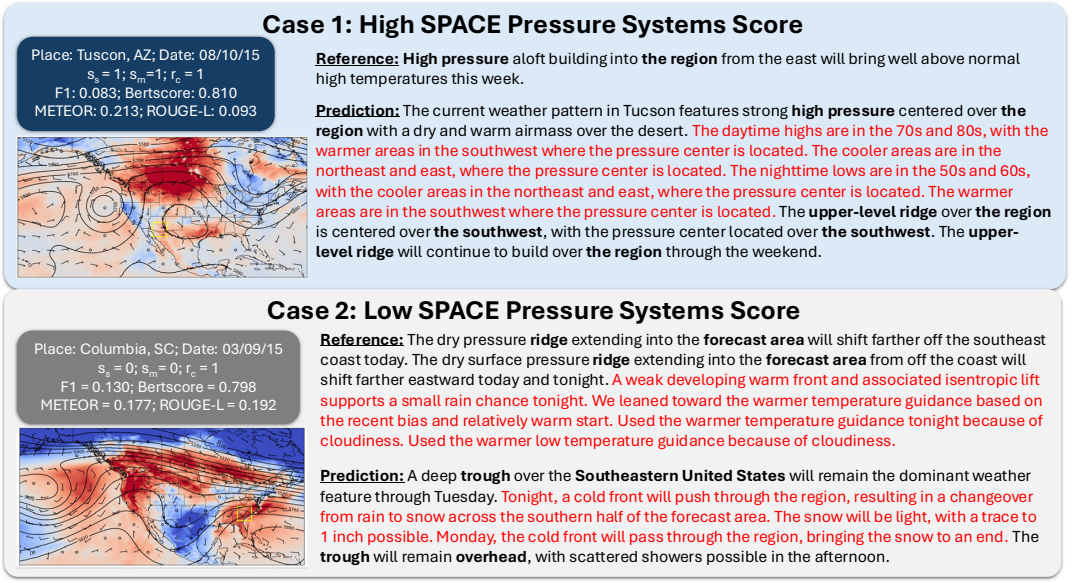


Figure 3. *Two cases demonstrating differences between* SPACE *scores and traditional skill metrics. The reference text samples are filtered NWS AFDs, and the prediction text samples are generated from the fine-tuned version of LLaVA-v1.5-7B. The terms in bold are used to compute* SPACE *scores for pressure systems. Sentences that are irrelevant to the* SPACE *scores are shown in red.*

4. Conclusion

In this work, we introduce a dataset with more than one million samples that we use to train a VLM for discussion generation from images of weather forecasts across North America. We also introduce a skill metric (SPACE) to evaluate the ability of language models to identify important synoptic features in the correct locations. Fine-tuning VLM projector weights with LoRA adapters leads to greater overall performance than the base models on both SPACE and traditional metrics. Climatology and fine-tuned Qwen2.5-VL-7B had the highest scores in most of the metrics in this study, with Qwen2.5-VL-7B excelling more in describing features relevant to a precise location. The failure of any model to score higher than climatology in the SPACE-aggregate scores demonstrates a key limiting factor in the ability of the models in this study to accurately describe synoptic features throughout the entire system.

The dataset used in this study pairs highly complex forecast discussions from the NWS with images of weather forecasts. This study performed experiments on synoptic pressure systems as an example use case and focused entirely on the United States, which creates an opportunity for future work. Human expert evaluation for text describing various types of atmospheric phenomena would also be valuable for strengthening confidence in any current or future metrics. The size and complexity of the dataset allow the user to perform experiments on other types of phenomena, specific locations or regions, spatial scales, and temporal ranges. Although the fine-tuned models often outperformed the base models on SPACE scores, improved model architectures and loss functions, agentic frameworks, and video-language models could add additional benefits.

Weather forecasts are highly complex, and complete descriptions of them therefore often require long and detailed discussions. None of the models used in this study comes close to rivaling the reliability nor complexity of NWS AFDs, so the evaluation of text generation for weather and climate applications must begin with basic isolated tasks. SPACE scores excel in capturing similarities in some of the most important details of text samples. They can help evaluate the ability of the models to demonstrate physical understanding of atmospheric processes rather than only learning common textual patterns. By

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press

determining the ability of the model to do so, the user can better understand when to stop training, which can help prevent overfitting.

Although SPACE scores are a useful method for evaluating text for multimodal language models for weather and climate applications, they do not fully capture the level of quality in generated forecast discussions, and additional metrics that evaluate text would help create a more comprehensive evaluation framework. Future improvements could involve temporal accuracy (more relevant for longer lead times), system evolution (more relevant for models that can handle videos), and vertical levels (more relevant for models that can handle more channels).

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/eds.2026.10050.

Acknowledgments. The authors thank the reviewers of this work for their dedication toward supporting the research community and providing comments that were critical toward improving the manuscript.

Author contribution. Conceptualization, Investigation, Methodology, Project administration, Validation and Writing - review & editing: T.H., A.M., and C.A.; Data curation, Formal analysis, Software, Visualization and Writing - original draft: T.H.; Funding acquisition, Resources, Supervision: A.M., and C.A. T.H., A.M., and C.A. contributed new analytic metrics and wrote the manuscript. T.H. wrote all the codes, retrieved, processed, and harmonized datasets, and performed the analyses for technical validation of the new resource. A.M. and C.A. conceived the study.

Competing interests. The author declare none.

Data availability statement. The code used to conduct this study can be found at Higgins (2026). The dataset can be found here: https://huggingface.co/datasets/Aikyam-Lab/Synoptic-Bench.

Author contribution. Conceptualization, Investigation, Methodology, Project administration, Validation and Writing - review & editing: T.H., A.M., and C.A.; Data curation, Formal analysis, Software, Visualization and Writing - original draft: T.H.; Funding acquisition, Resources, Supervision: A.M., and C.A.

Ethics statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding statement. This work was funded by the University of Virginia Environmental Institute Climate Fellows Program Grant DN002057.

Provenance statement. This article is part of the Climate Informatics 2026 proceedings and was accepted in Environmental Data Science on the basis of the Climate Informatics peer-review process.

References

Brady R and Spring A (2021) Verification of weather and climate forecasts. *Journal of Open Source Software 6*(59), 2781.

Chen J, **Zhou P**, **Hua Y**, **Chong D**, **Cao M**, **Li Y**, **Chen W**, **Zhu B**, **Liang J and Yuan Z** 2025 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*. Toronto, ON: ACM, pp. 5322–5333.

He X, **You Z**, **Gong J**, **Liu C**, **Yue X**, **Zhuang P**, **Zhang W and Bai L** (2025) *RadarQA: Multi-modal Quality Analysis of Weather Radar Forecasts*, Version Number: 1.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **De Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **Rosnay D P**, **Rozum I**, **Vamborg F**, **Villaume S, and Thépaut JN** (2020) *The ERA5 global reanalysis*, *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049 (en).

Higgins T (2026) Synoptic-Bench, Zenodo, Version 1.0. https://doi.org/10.5281/zenodo.20548726.

Jain H and Jain R (2017) Big data in weather forecasting: Applications and challenges. In *International Conference on Big Data Analytics and Computational Intelligence (ICBDAC) (Chirala, Andhra Pradesh, India), IEEE, March 2017*, Piscataway, NJ, USA: IEEE pp. 138–142.

Jones JW, **Hansen JW**, **Royce FS and Messina CD** (2000) *Potential benefits of climate forecasting to agriculture*, agriculture, *Ecosystems & Environment 82*(1–3), 169–184.

Li H, **Wang Z**, **Wang J**, **Wang Y**, **Lau AKH and Qu H** (2024) *CLLMate: A Multimodal Benchmark for Weather and Climate Events Forecasting*, Version Number: 2.

Ma C, **Hua Z**, **Anderson-Frey A**, **Iyer V**, **Liu X and Qin L** (2024) *WeatherQA: Can Multimodal Language Models Reason about Severe Weather?*, Version Number: 2.

Murphy AM, **Homeyer CR and Allen KQ** (2023) Development and investigation of GridRad-severe, a multiyear severe event radar dataset. *Monthly Weather Review 151*(9), 2257–2277.

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce (2015) *NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, Artwork Size: 574.507 Tbytes Pages: 574.507 Tbytes*.

National Research Council (2006) *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. Washington, D.C: National Academies Press.

Randles CA, **Da Silva AM**, **Buchard V**, **Colarco PR**, **Darmenov A**, **Govindaraju R**, **Smirnov A**, **Holben B**, **Ferrare R**, **Hair J**, **Shinozuka Y and Flynn CJ** (2017) The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *Journal of Climate 30*(17), 6823–6850. https://doi.org/10.1175/JCLI-D-16-0609.1.

Tang S, **Xu J**, **Zhang J**, **Chen Y**, **Jin Q**, **Shen L**, **Liu C and Xiang S** (2025) *MeteorPred: A Meteorological Multimodal Large Model and Dataset for Severe Weather Event Prediction*, Version Number: 2.

Uccellini LW and Ten Hoeve JE (2019) Evolving the National Weather Service to build a weather-ready nation: Connecting observations, forecasts, and warnings to decision-makers through impact-based decision support services. *Bulletin of the American Meteorological Society 100*(10), 1923–1942.

Ukhurebor KE, **Adetunji CO**, **Olugbemi OT**, **Nwankwo W**, **Akinola Samson Olayinka CU and Hefft DI**, *Precision agriculture: Weather forecasting for future farming*, *AI, Edge and IoT-Based Smart Agriculture*, Amsterdam, Netherlands: Elsevier, 2022, pp. 101–121 (en).

Varambally S, **Fisher M**, **Thakker J**, **Chen Y**, **Xia Z**, **Jafari Y**, **Niu R**, **Jain M**, **Manivannan VV**, **Novack Z**, **Han L**, **Eranky S**, **Cachay SR**, **Berg-Kirkpatrick T**, **Watson-Parris D**, **Ma YA and Yu R** (2025) *Zephyrus: An Agentic Framework for Weather Science*, Version Number: 1.

Veillette M, **Samsi S and Mattioli C** (2020) SEVIR: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. In Larochelle H, Ranzato M, Hadsell R, Balcan MF and Lin H (eds.), *Advances in Neural Information Processing Systems*, Vol. *33*. Curran Associates, Inc, pp. 22009–22019.

World Meteorological Organization (2015) *Valuing Weather and Climate: Economic Assessment of Meteorological and Hydrological Services*. Geneva, Switzerland: World Meteorological Organization.

Cite this article: Higgins T, Mamalakis A and Agarwal C (2026). SynopticBench: evaluating vision-language models on generating weather forecast discussions of the future. *Environmental Data Science*, 5: e19. doi:10.1017/eds.2026.10050

https://doi.org/10.1017/eds.2026.10050 Published online by Cambridge University Press