BeoGrid-Bench: Can Foundation Models Understand Multimodal Gridded Geo-Spatial Data?

Anonymous Authors¹

Abstract

We present GeoGrid-Bench, a benchmark designed to evaluate the ability of foundation models to understand geo-spatial data in the grid structure. Geo-spatial datasets pose distinct challenges due to their dense numerical values, strong spatial and temporal dependencies, and unique multimodal representations including tabular data, heatmaps, and geographic visualizations. To assess how foundation models can support scientific research in this domain, GeoGrid-Bench features large-scale, real-world data covering 16 climate variables across 150 locations and extended time frames. The benchmark includes approximately 3,200 question-answer pairs, systematically generated from 8 domain expert-curated templates to reflect practical tasks encountered by human scientists. These range from basic queries at a single location and time to complex spatiotemporal comparisons across regions and periods. Our evaluation reveals that vision-language models perform best overall, and we provide a fine-grained analysis of the strengths and limitations of different foundation models in different geo-spatial tasks. This benchmark offers clearer insights into how foundation models can be effectively applied to geo-spatial data analysis and used to support scientific research.¹

1. Introduction

Geo-spatial data pose distinct challenges for foundation models due to their inherent spatio-temporal dependencies and exceptionally high data density. Unlike typical tabular records for knowledge retrieval (Zhang et al., 2023a;



Figure 1. Overview of GeoGrid-Bench. The benchmark features questions generated from templates that vary by location, time period, and climate variable, then rewritten with natural language context. Each question is paired with multimodal input—either heatmaps as images or tabular grids of numerical values. We evaluate models on their ability to solve the queries through different modalities—natural language, code, or vision. Ground-truth answers capture find-grained aspects like overall trends, spatial references (from top-left to lower-right), coordinate references (row and column indices), and label references (textual marks on the maps), whenever available.

Pasupat & Liang, 2015; Zhang et al., 2025) or natural images, climate data exists in structured, gridded formats with complex, interconnected numerical values often represented through modalities such as tables, heatmaps, or geographic images spanning across space and time. These data are typically organized in highly structured, gridded formats that encode interconnected numerical values across spatial and temporal dimensions. Each data point is not an isolated unit but part of a dense, multi-dimensional array that reflects physical processes, environmental interactions, or geographical phenomena evolving over time. Meanwhile, models

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

¹All code and data will be made publicly available upon acceptance.

can also easily get lost in the context (Liu et al., 2023) withoverwhelming volumes of values per sample.

057 Informed decision-making in fields such as disaster re-058 sponse, climate science, and urban development depends 059 on the ability to detect and interpret patterns across regions 060 and over time. However, there remains a lack of bench-061 marks that directly address the unique challenges posed by 062 geo-spatial gridded data. Most existing efforts docus on 063 object detection, semantic segmentation, object counting, 064 captioning, or scene understanding of Earth observation im-065 ages (Lacoste et al., 2023; Danish et al., 2024; Zhang & 066 Wang, 2024; Zheng et al., 2023; Wang et al., 2024; Muhtar 067 et al., 2024; Bazi et al., 2024; Kuckreja et al., 2024), func-068 tion calls to the Geographic Information System (GIS) or 069 SQL queries for data retrieval (Krechetova & Kochedykov, 070 2025; Jiang & Yang, 2024; Ning et al., 2025; Mooney et al., 2023; Zhang et al., 2023b), or simplified query setups that overlook the spatial-temporal complexities in practical geospatial analysis (Bhandari et al., 2023). 074

075 To understand how foundation models can assist geo-spatial 076 data analysis, we introduce GeoGrid-Bench, a benchmark 077 explicitly designed to evaluate model performance on mul-078 timodal, real-world geo-spatial data. We adopt domain 079 expert-curated query templates to reflect realistic questions that practitioners would encounter in geo-spatial anal-081 ysis-providing data in both tabular and image formats. 082 These tasks range from simple queries about a fixed loca-083 tion and time to more complex analyses involving multiple locations and temporal comparisons. For each template, we 085 develop oracle code that is applied uniformly to all query 086 instances, enabling scalable and consistent generation of 087 question-answer pairs. Our contributions can be summa-088 rized as follows: 089

• Large-scale, real-world data: A domain-centric benchmark built on large-scale, real-world climate projection data, presented in multimodal formats commonly used by actual practitioners, including structured numerical tables and geographic visualizations.

090

091

092

093

094

095

096

097

098

099

100

104

105

106

108

109

- Scalable query generation: A systematic user query generation pipeline based on domain expert-designed templates, reflecting diverse and realistic scientific challenges.
- **Comprehensive evaluation:** Evaluation of foundation models with language, coding, multimodal, and reasoning capabilities across find-grained answer aspects and data modalities to diagnose their strengths and weaknesses in geo-spatial analysis tasks.

Through comprehensive evaluations, we find that visualizing dense, gridded geo-spatial data as heatmaps is the most accessible format for existing foundation models to interpret. In contrast, models struggle to generate flawless code for completing these tasks. Across all model types, identifying broad trends proves easier than making fine-grained regional distinctions, and models exhibit varying strengths and weaknesses depending on the task. With GeoGrid-Bench, we aim to shed light on the strengths and limitations of current foundation models when applied to multimodal geo-spatial data, a core yet underexplored format in climate science. Our goal is to support and advance the development of practical AI-assisted tools that can aid scientific research and decision-making.

2. BeoGrid-Bench: Overview of Data Features and Tasks

GeoGrid-Bench aims to reflect the real-world challenges that scientists face when analyzing geo-spatial data at scale. To achieve this, it features *large-scale, real-world* geospatial data sourced and sampled from ClimRR (Argonne National Laboratory, 2023), capturing the complexity of environmental conditions across North America. An overview of user-model interaction is shown in Figure 1.

GeoGrid-Bench is built to capture the unique grid structure. Climate projection data are typically organized across spatial grids and time sequences, resulting in dense, highdimensional arrays. The data is inherently interconnected, with each point influenced by its geographic neighbors and historical context. This structure poses unique challenges: models must capture spatio-temporal dependencies and handle variability across scales to derive meaningful insights.

Geo-spatial data is also inherently multimodal, presented as tabular data, heatmaps, or geographic visualizations, with each format sharing alignment across a spatial grid structure. Each grid cell encodes a rich array of numerical data that captures localized atmospheric behavior and climate dynamics over time. This multimodal grid structure makes our GeoGrid-Bench an ideal testbed for foundation models designed to reason across space, time, and modality. To perform well, foundation models must integrate spatial context from neighboring cells, understand temporal trends across multi-year projections, and interpret information presented in diverse formats and patterns.

To capture the wide range of questions concerning practitioners at the forefront of geo-spatial analysis, we surveyed 13 domain experts in natural hazard risk domains, resulting in 8 template questions based on their input and around 3200 query instances in GeoGrid-Bench. Each template includes placeholders based at one or two geographic locations, time frames, and climate variables, requiring one to eight data frames. This design allows us to generate a scalable set of scientifically concrete queries that reflect analytical goals. Specifically, GeoGrid-Bench evaluates the following capabilities of foundation models: 1. Identifying regions with
the most significant patterns. 2. Comparing data across
different locations and times. 3. Analyzing temporal trends
and seasonal variations. 4. Interpreting data in multimodal
formats.

1163. ConstructingGeoGrid-Bench At Scale

118 GeoGrid-Bench features diverse real-world geo-spatial 119 data. Now we discuss our sample curation process, and a 120 visual illustration is included in Figure 4 in the appendix. 121 Each data sample is formed by extracting a specific climate-122 location-time slice from the ClimRR (Argonne National 123 Laboratory, 2023) dataset. We sample from the 16 climate 124 variables listed in the appendix. For each climate variable, 125 we select around 50 locations where this climate variable 126 is the most prominent, resulting in a total of 150 distinct 127 locations across all climate variables, a subset of ClimRR. 128 For example, the benchmark includes more regions in South-129 ern California for wildfire risk, while precipitation-related 130 examples are more concentrated in the Pacific Northwest to 131 reflect region-specific climate concerns.

132 We render each data sample in either a **tabular** or **image** 133 format, both structured over a spatial grid. For a given lo-134 cation and its longitude and latitude, we retrieve all grid 135 cells within a square region with edge size 84 to 144 km 136 around it, resulting in approximately 50 to 150 entries in 137 the 12-by-12 km grid. In the tabular modality, we prepare 138 each table with numerical values, a caption, and row and 139 column indices as textual strings. In the image modality, we 140 prepare three types of visualization with increasing infor-141 mation densities: (1) A standalone heatmap, (2) A heatmap 142 with overlaid numerical annotations at each grid cell, and 143 (3) A heatmap overlaid on an actual geographic base map. 144 Specifically, we render the tabular data as a heatmap with 145 color gradients. This heatmap is optionally added with numerical annotation of the value on each cell, or overlaid 147 on a geographic base map (OpenStreetMap contributors, 148 2024) using Folium (Folium, 2023). To maintain consis-149 tency with the tabular format, we also render row and col-150 umn indices around the heatmap. This visualization offers 151 a richer representation to mirror common practices in real-152 world analysis. To isolate the challenge of data retrieval, 153 GeoGrid-Bench provides the foundation model during eval-154 uation with all necessary data frames in either tabular or 155 image formats, focusing solely on whether the model can 156 solve the problem given the relevant information. 157

158 GeoGrid-Bench builds on expert-curated templates for scalable query generation. Based on in-depth discussions about the analytical tasks our domain experts perform, we develop eight representative question templates, which are included in Table 1 in the appendix. Each template takes as input one or two climate variables, locations, and time 164 frames and outputs a filled-in user query in our benchmark, and may require between one and eight data frames to answer. This structured approach enables the automatic generation of a wide variety of concrete, data-driven queries tailored to real-world analytical needs.

For every template, we manually craft oracle code that deterministically solves the question and prepares ground-truth answers in desired formats. *Crucially, the same oracle applies uniformly to every query generated from a given template, enabling the scalability of the benchmark. As a result, once a template and its oracle are validated, we ensure the quality of every generated instance.*

Each question is a multiple-choice with four options, all generated by the oracle code rather than a language model. Recognizing that a foundation model may excel at different aspects in answering a geo-spatial query, each query probes a different aspect in giving the answer, as shown in Figure 1. Specifically, answer options target the following aspects: (1) Overall patterns (e.g., the wildfire risk overall increases). (2) Spatial references (e.g., the highest wildfire risk occurs around the top-left region). (3) Coordinate references (e.g., the highest wildfire risk occurs around Column 204 Row 106). (4) Label references (e.g., the highest wildfire risk occurs near the textual label "Santa Clara" on the map), which is only available for the image type "heatmap overlaid on an actual geographic base map". In addition, to explore which data modalities most effectively support geo-spatial analysis, we evaluate models across three input settings: language-only, language and code, and language and vision. Detailed prompting and result parsing strategies for each setting are provided in Appendix C.

4. Experiment

Experimental Setup. We benchmark a range of state-ofthe-art closed-source and open-source models on GeoGrid-Bench. Our evaluation covers 5 models from OpenAI, including o4-mini, GPT-4.1, GPT-4.1-mini, GPT-4o, and GPT-4o-mini (OpenAI, 2024; 2025; Hurst et al., 2024), and 6 open-source models including Llama-4-Maverick, Llama-4-Scout, Llama-3.2-11B-Vision, Llama-3.2-3B, Llama-3.1-8B (Grattafiori et al., 2024; AI, 2024), and Qwen-2.5-VL-7B (Bai et al., 2025). OpenAI models are accessed via API calls, and Llama-4 models are accessed through the Lambda Inference API. Inferences for other open-source models run locally on four NVIDIA A100-SXM4 GPUs with 40GB of VRAM. For all models, we set max_new_tokens as 1024 with default temperature and sampling strategies.

4.1. Evaluation Results and Findings

Vision-language models achieve the strongest performance in geo-spatial tasks. Among the models we evalu-

<pre>model_name</pre>	data_modality	overall_accuracy
o4-mini	language and vision	0.644
GPT-4.1	language and vision	0.578
GPT-4.1-mini	language and vision	0.568
o4-mini	language-only	0.534
GPT-40	language and vision	0.518
GPT-4.1	language-only	0.512
<pre>model_name</pre>	data_modality	overall_accuracy
Llama-4-Maverick	language and vision	0.580
Llama-4-Scout	language and vision	0.508
Llama-4-Maverick	language-only	0.486
Llama-4-Scout	language-only	0.457
Qwen2.5-VL-7B	language and vision	0.413
Llama-4-Maverick	language and code	0.337
	0.2 0.4	05 06

Figure 2. Selected evaluation results for OpenAI and open-source
models across different data modalities (columns correspond to
the fine-grained answer aspects defined in Section 3). This table
shows only a subset of the full results; the complete evaluation
tables can be found in the Appendix.

190

ate, o4-mini achieves overall the highest performance, while 191 Llama-4-Maverick leads among open-source models, as shown in Figure 5 in the Appendix. Overall, models that 193 receive input in the vision modality consistently outperform those using language-only input. This suggests that 195 converting geo-spatial gridded data into heatmap visualiza-196 tions-rather than presenting models directly with large 197 volumes of raw numerical values in tabular forms-enables foundation models to more effectively interpret such data 199 200 with complex spatial-temporal patterns.

Inferior performance in code highlights the need for 202 more agentic models in geo-spatial tasks. Contrary to our expectations, foundation models leveraging programming 204 code do not outperform their language-only counterparts on 205 our task. Upon closer inspection, much of the generated 206 code is not directly executable in a single pass. For instance, models produce incomplete scripts or bugs, omit expected 208 outputs, fail to parse data, or struggle with planning over 209 geo-spatial data-ultimately requiring human intervention 210 across multiple iterations. This limitation aligns with how 211 we construct the oracle code in the benchmark. This issue is 212 more severe in open-source models like Llama, which tend 213 to produce fewer executable code. We, therefore, emphasize 214 the need for stronger agentic behaviors (Plaat et al., 2025; 215 Kapoor et al., 2024; Ng, 2024) in foundation models, where 216 we define "agentic" as the ability to autonomously generate 217 fully executable code for human end-users in a single inter-218 action, particularly when the end-users are domain scientists 219

rather than programmers.

Fine-grained geo-spatial tasks reveals different strengthweakness tradeoffs. Commercial and open-source models exhibit different strengths and weaknesses in fine-grained geo-spatial tasks, as shown in Figure 5 in the Appendix. Specifically, open-source models generally struggle more than commercial ones in identifying regions with the most significant patterns. However, both types of models perform well when comparing trends between two locations or analyzing seasonal variations at a single location. In contrast, they show weaker performance when comparing seasonal variations across multiple locations or comparing data across different locations and times.

Models perform better at identifying overall trends than fine-grained region detections. As mentioned in Figure 1, target answers captures fine-grained aspects in answering these geo-spatial queries. Evaluation results in Figure 6 (a) and (b) in the Appendix show that models perform best on the "trend" column, while accuracy drops for spatial, coordinate, or label references—highlighting a need for improvement in fine-grained regional understanding.

Heatmaps with numerical annotations enhance performance, whereas map-overlaid heatmaps pose greater challenges for vision-language models. Figure 6 (c) in the Appendix compares model performance across three input image formats. Adding numerical annotations to heatmaps improves model accuracy compared to using color gradients alone. In contrast, the most realistic format, where heatmaps are overlaid on geographic base maps, poses the challenge for all models, as the added visual complexity hinders spatial pattern recognition.

5. Conclusion

We introduced 👹 GeoGrid-Bench, a comprehensive benchmark designed to evaluate the capability of foundation models to understand multimodal gridded geo-spatial data. GeoGrid-Bench features structured, dense numerical data using real-world gridded datasets and expert-curated templates to evaluate scientifically relevant geo-spatial tasks. This integrated design enables robust and scalable assessment of foundation models across vision, language, and code modalities. Our evaluation reveals that while visionlanguage models excel at interpreting spatial patterns from heatmaps, they still struggle with fine-grained regional understanding and label-based reasoning. Meanwhile, language and code models show limited success in generating executable analysis scripts without human intervention, highlighting the need for stronger agentic behavior. These findings point to several critical areas where model capabilities must improve to meet the practical needs of geo-spatial scientific analysis.

220 Impact Statement

Overall, this work can inform the development of more
capable models to process and understand the dense numerical data, spatiotemporal dependencies, and multimodal
representations of geo-spatial data, supporting the advancement of foundation models for informed decision-making
and resilience building across a wide range of real-world
challenges.

229 We acknowledge that this dataset is limited to the United 230 States due to data availability. Additionally, our benchmark 231 focuses on geo-spatial data in gridded formats, intentionally 232 excluding other common data types such as Earth observa-233 tion and remote sensing imagery, which have already been 234 extensively studied in prior work. However, the underlying 235 framework are designed to be generalizable and can be read-236 ily applied to similar gridded geo-spatial datasets from other 237 regions. Building on this foundation, future work will focus 238 on expanding GeoGrid-Bench beyond the United States and 239 incorporating richer data modalities such as satellite im-240 agery, elevation maps, and land use data to enable broader 241 and more diverse analytical capabilities. 242

References

243

244

245

246

247

248

249

250

251

252 253

254

255

256

257

258

259

261

263

264

265

266

267

268

269

270

271

272

273

- AI, M. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2024. URL https://ai.meta.com/blog/ llama-4-multimodal-intelligence/. Accessed: 2025-04-27.
- Argonne National Laboratory. Climrr: Climate risk and resilience portal. https://climrr.anl.gov, 2023. Accessed: 2025-04-15.
- Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.
- Bazi, Y., Bashmal, L., Al Rahhal, M. M., Ricci, R., and Melgani, F. Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477, 2024.
- Bhandari, P., Anastasopoulos, A., and Pfoser, D. Are large language models geospatially knowledgeable? In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pp. 1–4, 2023.
- Danish, M. S., Munir, M. A., Shah, S. R. A., Kuckreja, K., Khan, F. S., Fraccaro, P., Lacoste, A., and Khan, S. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024.

- Folium. Folium: Python data. leaflet.js maps. https://python-visualization.github. io/folium/latest/, 2023. URL https: //python-visualization.github.io/ folium/latest/. Version 0.14.0.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv* preprint arXiv:2407.21783, 2024.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-40 system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jiang, Y. and Yang, C. Is chatgpt a good geospatial data analyst? exploring the integration of natural language into structured query language within a spatial database. *ISPRS International Journal of Geo-Information*, 13(1): 26, 2024.
- Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., and Narayanan, A. Ai agents that matter. *arXiv preprint arXiv:2407.01502*, 2024.
- Krechetova, V. and Kochedykov, D. Geobenchx: Benchmarking llms for multistep geospatial tasks. *arXiv preprint arXiv:2503.18129*, 2025.
- Kuckreja, K., Danish, M. S., Naseer, M., Das, A., Khan, S., and Khan, F. S. Geochat: Grounded large visionlanguage model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27831–27840, 2024.
- Lacoste, A., Lehmann, N., Rodriguez, P., Sherwin, E., Kerner, H., Lütjens, B., Irvin, J., Dao, D., Alemohammad, H., Drouin, A., et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36:51080–51093, 2023.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023.
- Mooney, P., Cui, W., Guan, B., and Juhász, L. Towards understanding the geospatial skills of chatgpt: Taking a geographic information systems (gis) exam. In *Proceedings of the 6th ACM SIGSPAtIAL international workshop on AI for geographic knowledge discovery*, pp. 85–94, 2023.
- Muhtar, D., Li, Z., Gu, F., Zhang, X., and Xiao, P. Lhrsbot: Empowering remote sensing with vgi-enhanced large multimodal language model. In *European Conference on Computer Vision*, pp. 440–457. Springer, 2024.

275 Ng, A. Welcoming diverse approaches keeps ma-Zhang, Y., Wei, C., Wu, S., He, Z., and Yu, W. Geogpt: 276 chine learning strong. June 2024. URL https: Understanding and processing geospatial tasks through 277 //www.deeplearning.ai/the-batch/ an autonomous gpt. arXiv preprint arXiv:2307.07930, 278 welcoming-diverse-approaches-keeps-machine2023brning-strong/. 279 Zheng, H., Zhang, C., Guan, K., Deng, Y., Wang, S., Rhoads, Ning, H., Li, Z., Akinboyewa, T., and Lessani, M. N. An 280 B. L., Margenot, A. J., Zhou, S., and Wang, S. Segment autonomous gis agent framework for geospatial data re-281 any stream: Scalable water extent detection with the segtrieval. International Journal of Digital Earth, 18(1): 282 ment anything model. In NeurIPS 2023 Computational 2458688, 2025. 283 Sustainability: Promises and Pitfalls from Theory to De-284 OpenAI. Openai o3 and o4-mini system card, ployment, 2023. 285 2024. URL https://openai.com/index/ 286 o3-o4-mini-system-card/. Accessed: 2025-04-287 18. 289 OpenAI. Introducing gpt-4.1 in the api, 2025. URL 290 https://openai.com/index/gpt-4-1/. Ac-291 cessed: 2025-05-12. 292 293 OpenStreetMap contributors. Openstreetmap, 2024. URL 294 https://www.openstreetmap.org/. Accessed: 295 2025-05-12. 296 297 Pasupat, P. and Liang, P. Compositional semantic parsing on 298 semi-structured tables. arXiv preprint arXiv:1508.00305, 299 2015. 300 301 Plaat, A., van Duijn, M., van Stein, N., Preuss, M., van der 302 Putten, P., and Batenburg, K. J. Agentic large language 303 models, a survey. arXiv preprint arXiv:2503.23037, 2025. 304 305 Reimers, N. and Gurevych, I. Sentence-bert: Sentence em-306 beddings using siamese bert-networks. In Proceedings 307 of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational 308 309 Linguistics, 11 2019. URL https://arxiv.org/ abs/1908.10084. 311 Wang, J., Zheng, Z., Chen, Z., Ma, A., and Zhong, Y. Earth-312 vqa: Towards queryable earth via relational reasoning-313 based remote sensing visual question answering. In Pro-314 ceedings of the AAAI Conference on Artificial Intelli-315 gence, volume 38, pp. 5481-5489, 2024. 316 317 Zhang, C. and Wang, S. Good at captioning bad at count-318 ing: Benchmarking gpt-4v on earth observation data. In 319 Proceedings of the IEEE/CVF Conference on Computer 320 Vision and Pattern Recognition, pp. 7839–7849, 2024. 321 322 Zhang, T., Yue, X., Li, Y., and Sun, H. Tablellama: Towards 323 open large generalist models for tables. arXiv preprint 324 arXiv:2311.09206, 2023a. 325 326 Zhang, X., Wang, D., Dou, L., Zhu, Q., and Che, W. A 327 survey of table reasoning with large language models. 328 Frontiers of Computer Science, 19(9):199348, 2025. 329

A. Data Curation

We illustrate our sample curation process in Figure 4. Our template questions are included in Table 1. The full list of Climate Variables in GeoGrid-Benchis included below.

Full List of Climate Variables in GeoGrid-Bench

Maximum Annual Temperature, Minimum Annual Temperature, Consecutive Days with No Precipitation, Cooling Degree Days, Fire Weather Index, Maximum Daily Heat Index, Maximum Seasonal Heat Index, Number of Days with Daily Heat Index *i*, 95°F/105°F/115°F/125°F, Heating Degree, Annual Total Precipitation, Maximum Seasonal Temperature, Minimum Seasonal Temperature, Wind Speed.

Templates that require one data frame

1. Which region in the {location1} experienced the largest increase in {variable1} during {time_frame1}?

Templates that require two data frames

2. How has {variable1} changed between {time_frame1} and {time_frame2} in the {location1}?

3. What is the correlation between {variable1} and {variable2} in the {location1} during {time_frame1}?

4. How does {variable1} compare between {location1} and {location2} during {time_frame1}?

Templates that require four data frames

5. What is the *seasonal* variation of {climate_variable1} in {location1} during {time_frame1}?

6. Which *season* in {time_frame1} saw the highest levels of {variable1} in {location1}?

7. Which of $\{location1\}$ or $\{location2\}$ experienced a greater change in $\{variable1\}$ throughout $\{time_frame1\}$ and $\{time_frame2\}$?

Templates that require eight data frames

8. How does the *seasonal* variation of {variable1} in {location1} compare to that in {location2} for {time_frame1}?

Table 1. **Template questions in GeoGrid-Bench.** We develop those questions with domain experts. Each question includes placeholders for one or two locations, time frames, and geo-spatial variables. This design enables scalable question construction while capturing varying levels of complexity based on the number of data frames involved.

B. Evaluation Results

The evalution result tables are shown in Figure 5 and 6.

C. Inference Prompts and Result Parsing

To evaluate models across different modalities, we design prompts for three settings: language-only, language and code, and language and vision. Each prompt is designed to be simple yet encourage model response with desired style and consistent answer formatting.

- Language-only: models receive data in tabular format with instructions "Think step by step before making a decision. Then, explicitly state your final choice after the special phrase "####Final Answer" followed by (a), (b), (c), or (d). Please don't use programming code.".
- Language and programming code: models receive data in tabular format with instructions "Please write Python code to answer the question and show the complete script. You must include a print statement at the end of the code that outputs the final answer using the special phrase "###Final Answer' followed by (a), (b), (c), or (d)."
- Language and vision: models receive climate data in one of the three image formats with instructions "Analyze this image and answer the question. Think step by step before making a decision. Then, explicitly state your final choice after the special phrase "####Final Answer" followed by (a), (b), (c), or (d).".

In each mode, we provide the model with the user query, the relevant data (in either tabular or image format), all four multiple-choice options, and system instructions as inputs. We extract the model's final answer following the special tokens



Figure 3. We prepare every data sample in one of the four formats: (a) 2D table as a textual string. (b) standalone heatmap; (c) heatmap with overlaid numerical annotations at each grid cell; (d) heatmap overlaid on an actual geographic base map. These formats reflect real-world climate data practices and differ markedly from typical natural images seen by foundation models. More in Appendix D.



Figure 4. **Overview of the example curation process.** Each example in GeoGrid-Bench is constructed by combining a query template with sampled climate variables, locations, and time frames from real-world climate data. Each template is paired with a corresponding oracle code that deterministically generates target answers for all filled-in question instances under that template.

"###Final Answer" to facilitate answer parsing. If the model fails to provide an explicit option (a), (b), (c), or (d), we use a sentence embedding model (Reimers & Gurevych, 2019) to identify the most similar option based on the model's response. When the model outputs Python code, we execute the code in a shell environment to extract the final answers.

D. Examples of Data Visualizations for All Query Templates

			Which of							
			{location1} or					What is the		How does the
			experienced a greater change	What is the seasonal	Which region in the {location1}	Which season in	How does {climate_variabl	between {climate_variabl	How has {climate_variabl	variation of {climate_variabl
			in {climate_variabl}	variation of {climate_variabl	experienced the largest increase	{time_frame1} saw the highest	el} compare between	el} and {climate_variabl	el} changed between	el} in {location1}
			<pre>el} throughout {time_framel}</pre>	el} in {location1}	in {climate_variabl	levels of {climate_variabl	{location1} and {location2}	<pre>e2} in the {location1}</pre>	{time_framel} and	<pre>compare to that in {location2}</pre>
model_name	data_modality	overall_accuracy	and {time_frame2}?	during {time_frame1}?	el} during {time_framel}?	el} in {location1}?	during {time_frame1}?	during {time_frame1}?	<pre>{time_frame2} in the {location1}?</pre>	for {time_framel}?
o4-mini	language and vision	0.644	0.667	0.673	0.743	0.813	0.623	0.453	0.453	0.724
GPT-4.1	language and vision	0.578	0.640	0.593	0.660		0.523	0.313	0.400	0.673
GPT-4.1-mini	language and vision	0.568	0.633	0.517	0.600		0.487	0.373	0.453	0.680
o4-mini	language-only	0.534	0.790	0.800	0.470	0.210	0.590	0.570	0.510	0.333
GPT-40	language and vision	0.518	0.613	0.407	0.630	0.773	0.447	0.370	0.380	0.525
GPT-4.1	language-only	0.512	0.690	0.670	0.450	0.530	0.540	0.470	0.450	0.293
GPT-4.1-mini	language-only	0.511	0.640	0.670	0.450	0.500	0.580	0.440	0.470	0.333
GPT-4o-mini	language and vision	0.462	0.573	0.657	0.363	0.700	0.400	0.437	0.373	0.192
o4-mini	language and code	0.453	0.650	0.660	0.420	0.150	0.500	0.470	0.530	0.242
GPT-4o-mini	language-only	0.437	0.630	0.550	0.410	0.400	0.270	0.570	0.380	0.283
GPT-4.1-mini	language and code	0.427	0.470	0.570	0.560	0.200	0.410	0.440	0.420	0.343
GPT-40	language-only	0.423	0.630	0.420	0.430	0.400	0.370	0.420	0.430	0.283
GPT 40 mini	language and code	0.412	0.500	0.580	0.350	0.290	0.430	0.420	0.440	0.283
GPT-40-min	language and code	0.369	0.440	0.530	0.270	0.260	0.330	0.400	0.390	0.333
Overall	language and vision	0.554	0.470	0.520	0.590	0.230	0.496	0.330	0.412	0.554
Overall	language and vision	0.483	0.676	0.622	0.442	0.408	0.430	0.494	0.448	0.305
Overall	language and code	0.406	0.506	0.572	0.388	0.230	0.400	0.416	0.418	0.313
Overall	all	0.481	0.602	0.588	0.476	0.474	0.455	0.433	0.426	0.392
				{location1} or {location2}		What is the		How does the		
model_name	data_modality	overall_accuracy	What is the seasonal variation of {climate_variabl el} in {locationl} during {time_framel}?	experienced a greater change in {climate_variabl el} throughout {time_frame1} and {time_frame2}?	Which season in {time_framel} saw the highest levels of {climate_variabl el} {locationl}?	<pre>tetrition between {climate_variabl el} and {climate_variabl e2} in the {location1} during {time_frame1}?</pre>	How does {climate_variabl el} compare between {location1} and {location2} during {time_frame1}?	variation of {climate_variabl el} in {location1} compare to that in {location2} for {time_frame1}?	How has {climate_variabl el} changed between {time_framel} and {time_frame2} in the {locationl}?	Which region in the {location1} experienced the largest increase in {climate_variabl e1} during {time_frame1}?
model_name Llama-4-Maverick	data_modality language and vision	overall_accuracy 0.580	What is the seasonal variation of {climate_variabl el} in {location1} during {time_frame1}? 0.627	experienced a greater change in {climate_variabl el} throughout {time_frame1} and {time_frame2}? 0.667	Which season in {time_frame1} saw the highest levels of {climate_variabl el} {location1}?	<pre>climate_variabl els and {climate_variabl els and {climate_variabl el2} in the {location1} during {time_frame1}? 0.373</pre>	How does {climate_variabl el} compare between {location1} and {location2} during {time_frame1}? 0.467	variation of {climate_variabl el} in {location1} compare to that in {location2} for {time_frame1}? 0.721	How has {climate_variabl el} changed between {time_frame1} and {time_frame2} in the {location1}? 0.443	Which region in the {location] experienced the largest increase in {climate_variabl e1} during {time_frame1}? 0.537
model_name Llama-4-Maverick Llama-4-Scout-	data_modality language and vision language and vision	overall_accuracy 0.580 0.508	What is the seasonal variation of {climate_variabl el} in {location1} during {time_frame1}? 0.627 0.463	experienced a greater change in {climate_variabl el} throughout {time_framel} and {time_frame2}? 0.667 0.607	Which season in {time_frame1} saw the highest levels of {climate_variabl e1} in {location1}? 0.807 0.727	<pre>{climate_variabl e1} and {climate_variabl e2} in the {location1} during {time_frame1}? 0.373 0.517</pre>	How does {climate_variabl el} compare between {location1} during {time_frame1}? 0.467 0.393	variation of {climate_variabl el} in {location1} compare to that in {location2} for {time_frame1}? 0.721 0.556	How has {climate_variabl el} changed between {time_frame1} and {time_frame2} in the {location1} 0.443 0.378	Which region in the {location} experienced the largest increase ictimate_variabl e1} during {time_frame1}? 0.537 0.423
model_name Llama-4-Maverick Llama-4-Scout- Llama-4-Maverick	data_modality language and vision language and vision language-only	overall_accuracy 0.580 0.508 0.486	What is the seasonal variation of {climate_variabl el} in {locationl} during {time_frame1}? 0.627 0.463 0.570	experienced a greater change in {climate_variabl el} throughout {time_framel} 0.667 0.607 0.620	Which season in {time_frame]} saw the highest levels of {climate_variabl el} in {location1}? 0.807 0.727 0.460	<pre>climate_variabl {climate_variabl e2} and {climate_variabl e2} in the {location1} during {time_frame1}? 0.373 0.517 0.370</pre>	How does {climate_variabl el} compare between {location1} and {location2} during {time_frame1}? 0.467 0.393 0.540	<pre>variation of {climate_variabl</pre>	How has {climate_variabl el} changed between {time_frame1} and the {location1}? 0.443 0.378 0.470	<pre>Which region in the {location1} experienced the largest increase in {climate_variabl e1} during {time_frame1}? 0.537 0.423 0.430</pre>
model_name Llama-4-Maverick Llama-4-Scout- Llama-4-Maverick Llama-4-Scout	data_modality language and vision language and vision language-only language-only	overall_accuracy 0.580 0.508 0.486 0.457	What is the seasonal variation of {climate_variabl el} in {location1} during {time_frame1}? 0.627 0.463 0.570 0.490	experienced a greater change in {climate_variabl el} throughout {time_framel} 0.667 0.607 0.620 0.480	Which season in {time_frame]} saw the highest levels of {climate_variabl el} in {location1}? 0.807 0.727 0.460 0.530	<pre>climate_variabl {climate_variabl {climate_variabl e2} in the {location1} during {time_frame1}? 0.373 0.517 0.370 0.510</pre>	How does {climate_variabl el} compare between {location1} and {location2} during {time_frame1}? 0.467 0.393 0.540 0.410	<pre>variation of {climate_variabl</pre>	How has {climate_variabl el} changed between {time_frame1} and the {location1}? 0.443 0.378 0.470 0.402	<pre>Which region in the {location1} experienced the largest increase in {climate_variabl e1} during {time_frame1}? 0.537 0.423 0.430 0.430</pre>
model_name Llama-4-Maverick Llama-4-Scout- Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B	data_modality language and vision language and vision language-only language and vision	overall_accuracy 0.580 0.508 0.486 0.457 0.413	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.463 0.570 0.469 0.490 0.490	experienced a greater change in {climate_variabl el} throughout {time_framel} 0.667 0.607 0.620 0.480 0.420	Which season in {time_frame1} saw the highest levels of {climate_variab1 el} in {location1}? 0.727 0.727 0.460 0.530 0.507	<pre>{climate veriabl el} add {climate veriabl e2} in the {location1} during {time_frame1}? 0.373 0.517 0.370 0.510 0.523</pre>	How does {climate_varial el} compre termen {location1} during {time_framel}? 0.467 0.393 0.540 0.410 0.330	<pre>variation of {climate_variabl</pre>	How has {climate_vriable el} changed el} changed and time_frame1} 0.431 0.378 0.432 0.432 0.432 0.432 0.337	Which region in the {location1} experienced the largest increase e1} during {time_framel}? 0.537 0.423 0.430 0.430 0.430 0.430
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Owen2.5-VL-7B Llama-4-Maverick	data_modality language and vision language and vision language and vision language and vision language and code	overall_accuracy 0.580 0.508 0.486 0.457 0.413 0.337	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.463 0.570 0.460 0.460 0.440 0.630	experienced a greater change if {climate_variabl el} throughout {time_framel} 0.667 0.607 0.620 0.480 0.420 0.350	Which season in {time_frame1} saw the highest levels of {climate_variabl el} in {location1}? 0.272 0.460 0.530 0.507 0.100	<pre>{climate veriabl el} add {climate veriabl e2} in the {locationl} during {time_framel}? 0.373 0.517 0.370 0.510 0.523 0.320</pre>	How does { {climate_varial el} compare beb beb doestion2} during {time_framel}? 0.467 0.393 0.540 0.410 0.330 0.540	<pre>variation of {climate_variabl</pre>	How has {climate_vriabl el}_changed el}_changed and time_frame1} int the {location}? 0.443 0.378 0.440 0.400	<pre>Which region in the {location1} experienced the largest increase in{climate_variabL e1} during {time_framel}? 0.537 0.423 0.430 0.430 0.430 0.430 0.430</pre>
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-4-Maverick Llama-3.2-3B	data_modality language and vision language and vision language and vision language and vision language and code language any	overall_accuracy 0.680 0.508 0.486 0.457 0.413 0.337 0.312	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.463 0.570 0.463 0.570 0.440 0.630	experienced a greater change [climate variabl el} throughout {time_framel} 0.667 0.667 0.660 0.480 0.480 0.420 0.350 0.350	Which season in {time_frame1} saw the highest Levels of {climate_variabl location1}? 0.807 0.727 0.460 0.530 0.507 0.100 0.410	<pre>{climate veriabl el} add {climate veriabl e2} in the {location1} during {time_frame1}? 0.373 0.517 0.370 0.510 0.523 0.320 0.280</pre>	How does (climate_variable) e1} coppare between location1] during (time_frame1)? 0.467 0.393 0.540 0.410 0.330 0.410 0.320	<pre>variation of {climate_variabl</pre>	How has {climatery veriable el} changed between {time_trame1} and the {location}}? 0.443 0.378 0.440 0.400 0.337 0.400 0.280 0.2	Which region in the {location1} experienced the largest increase ell during ell during (time_framel)? 0.423 0.430 0.430 0.430 0.430 0.430 0.240 0.367 0.230 0.240
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-4-Maverick Llama-3.2-3B Llama-4-Scout	data_modality language and vision language and vision language and vision language and vision language and code language and code language and code	overall_accuracy 0.680 0.508 0.486 0.457 0.413 0.337 0.312 0.312 0.312	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.463 0.570 0.463 0.0570 0.440 0.630 0.0290 0.470	experienced a greater change [climate variabl el} throughout {time_framel} 0.667 0.667 0.660 0.480 0.480 0.420 0.350 0.350 0.350 0.350	<pre>Which season in {time_frame1} saw the injhest levels of {climate_variab1 {location1}? 0.807 0.727 0.460 0.530 0.507 0.100 0.410 0.200</pre>	<pre>{climate veriabl el] and {climate veriabl e2} in the {locationl} during {time_framel}? 0.373 0.517 0.370 0.510 0.523 0.320 0.320 0.320 0.320</pre>	How does (cimate_variable) ei} compare between between during du	<pre>variation of {climate_variabl</pre>	How has {climatery variable between between itme_frame2} inn dtme_frame2) inn 0.433 0.378 0.470 0.400 0.337 0.400 0.280 0.280 0.290 0.	Which region in the {location1} experienced the largest increase inclination1 {climate_variabL e1} during dif_dime_frame1}? 0.423 0.430
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-4-Scout Llama-3.2-3B Llama-4-Scout Qwen2.5-VL-7B	data_modality language and vision language and vision language and vision language and vision language and vision language and code language and code language and code	overall_accuracy 0.580 0.508 0.486 0.457 0.413 0.337 0.312 0.311 0.296	What is the seasonal variation of {climate_variabt el} in {location1} during {time_frame1}? 0.627 0.643 0.6570 0.4400 0.6400 0.6300 0.0400 0.04700 0.04700 0.04700 0.04700 0.04700 0.04700 0.0470000000000	experienced a greater change if (climate variabl el} throughout {time_framel} 0.667 0.667 0.620 0.480 0.480 0.420 0.350 0.350 0.350	Which season in {time_frame1} saw the highest lovals - {climate_variabl el} in {location1}? 0.807 0.727 0.460 0.530 0.530 0.507 0.100 0.410 0.200 0.190	<pre>{climate veriabl el] and {climate veriabl e2} in the {locationl} during {time_framel}? 0.373 0.517 0.370 0.510 0.523 0.320 0.280 0.320 0.320 0.320 0.420</pre>	How does el} compare between { {location}} and {location} during	<pre>variation of {climate_variabl</pre>	How has {clianct variable between between {time_frame2} in {time_frame2} in 0.438 0.378 0.430	Which region in the {location1} experienced the largest increase int climate_variabL during {time_frame1}? 0.537 0.433 0.430
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-4-Scout Qwen2.5-VL-7B Qwen2.5-VL-7B Qwen2.5-VL-7B Llama-2.29B	data_modality language and vision language and vision language and vision language and vision language and code language and code language and code language and code	overall_accuracy 0.580 0.588 0.486 0.457 0.413 0.337 0.312 0.311 0.298 0.286 0.285	What is the seasonal variation of {climate_variabl el} in {location1} during {time_frame1}? 0.627 0.643 0.570 0.440 0.640 0.0400 0.0400 0.0400 0.0400000000	experienced a greater change if (climate variabl el} throughout {time_framel} 0.667 0.620 0.620 0.480 0.480 0.480 0.480 0.350 0.350 0.350 0.350	Which season in {time_frame1} saw the highest tevels of {climate_varient e1} in {location1}? 0.807 0.727 0.460 0.530 0.507 0.0100 0.410 0.200 0.190 0.320 0.020	<pre>{climate yeriabl el] and {climate yeriabl e2} in the e2 in the e10 and e10 and e2 in the e10 and e10 and</pre>	How does el} compare between location]} and (location]} and (location]} during	variation of {climate_variabl el} in {location1} compare to that in {location2} for {time_framel}? 0.721 0.556 0.424 0.364 0.380 0.424 0.364 0.380 0.192 0.293 0.182 0.300 0.140	How has {clianet, cransba between time_frame2) in the {location}} 0.443 0.443 0.440 0.440 0.440 0.450 0.500 0	Which region in the {location1} experienced the largest increase int {climate variabl el} during {time_frame1}? 0.537 0.423 0.430000000000
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-4-Scout Llama-3.2-3B Llama-4-Scout Qwen2.5-VL-7B Qwen2.5-VL-7B Qwen2.5-VL-7B Qwen2.5-VL-7B Llama-3.2-3B Llama-3.2-3B	data_modality language and vision language and vision language and vision language and vision language and code language and code language and code language and code	overall_accuracy 0.580 0.608 0.486 0.457 0.413 0.337 0.312 0.311 0.298 0.286 0.286 0.284	What is the seasonal variation of {climate_variabl el} in {location1} during {time_frame1}? 0.627 0.643 0.6570 0.4400 0.6400 0.6400 0.6400 0.6400 0.6400 0.6400 0.6400 0.6400 0.6400 0.6400 0.65000 0.65000 0.65000 0.65000 0.65000 0.65000 0.650000000000	experienced a greater change if (climate variabl el} throughout {time_framel} 0.667 0.620 0.620 0.620 0.620 0.620 0.0350 0.0350 0.0350 0.0350 0.0550 0.0550 0.0550 0.0550	Which season in {time_frame1} saw the highest Levels of {climate_variable {} {limate_variable } 0.807 0.727 0.460 0.530 0.530 0.530 0.530 0.530 0.000 0.100 0.320 0.320 0.030	<pre>{climate yeriabl elimate yeriabl el] and {climate yeriabl elimate</pre>	How does el} compare between location]} and (location]} and (location]} and (location]} during during during constant co	Variation of {climate_variabl el} in {location1} compare to that in {location2} for {time_frame}}? 0.721 0.720 0.721 0.7	How has {clianet, changed betwee time_frame2) in the {location}} 0.443 0.443 0.440 0.440 0.440 0.400 0.	Which region in the {location1} experienced the largest increase interval time_frame1}? 0.537 0.423 0.430 0.200 0.200 0.220
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-4-Scout Qwen2.5-VL-7B Qwen2.5-VL-7B Qwen2.5-VL-7B Llama-3.2-3B Llama-3.2-3B Llama-3.2-1B-VLama-	data_modality language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and code language and code	overall_accuracy 0.580 0.508 0.486 0.457 0.413 0.337 0.312 0.311 0.298 0.286 0.264 0.261	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.643 0.6570 0.4400 0.6400 0.6400 0.6400 0.6300 0.6300 0.6500 0.6500	experienced a greater change if (climate variable el} throughout (time_framel) 0.667 0.620 0.620 0.620 0.620 0.620 0.030 0.030 0.030 0.030 0.030 0.030 0.0350 0.0350 0.0550 0.000 0.0119	Which season in {time_frame]; saw the highest levels of ellimate_variet ellimate_variet o.807 0.727 0.460 0.530 0.530 0.530 0.530 0.530 0.000 0.410 0.320 0.320 0.020 0.030	<pre>{climate variabl</pre>	How does el} compare betwee location]} and {location]} and {location} during during divention during	Variation of {climate_variabl el} in {location1} compare to that in {location2} for {time_frame1}? 0.721 0.720 0.721 0.7	Hov has {clianet, changed between time_frame2} ithe {location}} 0.443 0.443 0.440 0.4	Which region in the {location1} experienced the largest increase increase idunt time_frame1? 0.537 0.423 0.430 0.200 0.200 0.2200 0.2200
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Uama-4-Scout Qwen2.5-VL-7B Llama-3.2-3B Llama-4-Scout Qwen2.5-VL-7B Qwen2.5-VL-7B Llama-3.2-18 Llama-3.2-18-Vision Llama-3.2-118-Vision	data_modality language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and code language and code language and code	overall_accuracy 0.580 0.588 0.486 0.457 0.413 0.337 0.312 0.311 0.298 0.286 0.266 0.264 0.264 0.261 0.233	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.643 0.6570 0.4400 0.6400 0.6400 0.6300 0.6300 0.6300 0.6500 0.6500 0.6500 0.6500 0.6500 0.6500	experienced a greater change if (climate variable el} throughout (time_framel) 0.667 0.620 0.620 0.620 0.620 0.0480 0.0480 0.0480 0.0480 0.0350 0.0350 0.0350 0.0550 0.0070 0.060 0.0419 0.0419 0.0419	Which season in {time_frame]; saw the highest tevels of {climate_variet {limate_variet {limate_variet 0.807 0.727 0.460 0.530 0.530 0.507 0.0100 0.400 0.320 0.320 0.020 0.030 0.034 0.043	<pre>{climate variabl</pre>	How does e1 compare between	variation of {climate_variabl e] in {location1} compare to that in {location2} for (time_frame1)? 0.721 0.721 0.721 0.721 0.721 0.721 0.721 0.721 0.721 0.723 0.723 0.723 0.723 0.723	How has {clianet, changed between time_frame2, in clianet, frame2, i	Which region in the {location1} experienced the largest increase increase it during {time_frame1}? 0.537 0.423 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.240 0.240 0.250 0.250
model_name Llama-4-Maverick Llama-4-Scout Llama-4-Scout Qwen2.5-VL-7B Llama-4-Maverick Llama-3.2-3B Llama-4-Scout Qwen2.5-VL-7B Qwen2.5-VL-7B Llama-3.2-3B Llama-3.2-1B Llama-3.2-1B-Vision Llama-3.2-11B-Vision	data_modality language and vision language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and code language and code language and vision	overall_accuracy 0.580 0.508 0.486 0.457 0.413 0.337 0.311 0.238 0.286 0.266 0.264 0.261 0.261	What is the seasonal variation of {climate_variabl el} in {location1} during {time_framel}? 0.627 0.643 0.6570 0.4400 0.4400 0.6400 0.6570 0.63000 0.63000 0.63000 0.63000 0.630000000000	experienced a greater change if (climater variable el) throughout (time_framel) 0.667 0.620 0.620 0.620 0.620 0.620 0.030 0.0300 0.0300 0.0300 0.0550 0.0070 0.060 0.0309 0.0319 0.0319	Which season in {time_frame]; saw the highest tevels of {elimate_varient {location}}7 0.807 0.727 0.460 0.530 0.507 0.0100 0.400 0.020 0.020 0.030 0.020 0.030 0.030 0.034 0.031	<pre>{climate_</pre>	How does (cimate_variabs fcoartion1) and flocation2) and flocation3) and flocation3) and flocation3) and flocation3) and flocation3) flocat	Variation of {climate_variabl el] in {location1} compare to that in {location2} for 0.7210	How has {clime_freatbook time_freatbook the {location}} 0.443 0.443 0.443 0.440	Which region in the {location1} experienced the largest increase identities and el} during el} during el] duri
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Qwen2.5-VL-7B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-1B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision	data_modality language and vision language and vision language-only language-only language and vision language and vision language-only language-only language-only	overall_accuracy 0.580 0.588 0.486 0.457 0.413 0.337 0.311 0.238 0.286 0.266 0.264 0.261 0.264 0.264 0.264	What is the seasonal variation of {climate_variabl el} in {location1} during time_framel}? 0.630 0.630 0.630 0.0400 0.0400 0.0400 0.0400 0.0400 0.0500 0.650 0.650 0.650 0.650 0.650	experienced a greater change if (climater variable el) throughout (time_framel) 0.667 0.620 0.620 0.620 0.620 0.620 0.0370 0.620 0.0370 0.0350 0.0350 0.0350 0.0550 0.0070 0.0109 0.019 0.019 0.020	Which season in {time_frame]; saw the highest tevels of {climate_variabi (location]}7 0.807 0.727 0.460 0.530 0.507 0.0100 0.400 0.020 0.030 0.030 0.030 0.030 0.330	<pre>{climate_eriabl el) and {climate_variabl el) and {climate_variabl el) and {locationl} during during {time_frame.}? 0.370 0.510 0.523 0.0300 0.520 0.020 0.0300 0.020 0.0310 0.0420 0.040 0.020 0.</pre>	How does (linete_variabs) (control) and (location) and (location) and (location) (lo	Variation of {climate_variabl el] in {location1} compare to that in {location2} for document	How has {clime_freatly the {location}} } Clime_freatly Clime_freatly Clime_freatly Clime_freatly Clime_freatly Clime C	Which region in the {location1} experienced the largest increase iduring time_framel}7 0.537 0.423 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.240 0.250 0.250 0.250 0.250 0.250 0.250 0.250 0.250
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Owen2.5-VL-7B Llama-3.2:3B Llama-3.2:3B Llama-3.2:3B Llama-3.2:3B Llama-3.2:1B-Vision Llama-3.2:1B-Vision Llama-3.2:1B-Vision Llama-3.2:18-Vision	data_modality language and vision language and vision language-only language-only language-only language-only	overall_accuracy 0.580 0.588 0.486 0.457 0.413 0.337 0.331 0.331 0.238 0.266 0.265 0.264 0.263 0.264 0.233	<pre>What is the seasonal variation of {climate_variabl el} in {locationl} during time_framel}? 0.627 0.627 0.637 0.637 0.0470 0.0470 0.0470 0.05700 0.05700 0.05700 0.05700 0.05700 0.0570000000000</pre>	experienced a greater change if (climater variable el) throughout (time_framel) and (time_framel)? 0.660 0.660 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620 0.620	Which season in {time_frame]; saw the highest levels of {clune1; no.807 0.727 0	<pre>{climate_ariabl el) and {climate_variabl el) and {climate_variable el) and {climateevariable el) an</pre>	How does {climate_variabs betwears betwears betwears betwears betwears betwears betwears betwears betwears close clo	variation of {climate_variabl el] in compare to that in {location1} compare to that in {location2} for {time_framel} 0.721 0.624 0.0364 0.0364 0.0360 0.0192 0.0300 0.0192 0.0300 0.0192 0.0300 0.0192 0.0300 0.0192 0.0300 0.0192 0.0300 0.0192 0.0300 0.0172	How has cliane (charge of the second the (location)) 100000000000000000000000000000000000	Which region in the {location] argest increase in{climate variable el] during time_framel]? 0.537 0.423 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.430 0.240 0.240 0.240 0.240 0.250 0.270 0.250 0.270 0.250 0.270 0.270 0.270 0.270 0.270 0.270 0.270 0.270 0.270 0.270 0.270 0.270 0.270
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Uwen2.5-VL-7B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-1B- Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-18B Ulama-3.2-18B Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.2-18 Llama-3.1-88 Llama-3.2-88 Llama-3.2-	data_modality language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and code language and vision language and vision language and vision language and vision language and vision language and vision language-ony language-ony	overall_accuracy 0.580 0.486 0.457 0.413 0.311 0.311 0.298 0.286 0.286 0.286 0.284 0.281 0.281 0.233 0.204	<pre>What is the seasonal variation of {climate_variabl el} in {location1} during (time_frame)? 0.0470 0.0470 0.0470 0.0470 0.0470 0.0470 0.0470 0.0470 0.0470 0.0477 0.0471</pre>	experienced a greater change if climate variabl ell throughout {time_framel} 0.667 0.667 0.667 0.660 0.040 0.040 0.040 0.040 0.0350 0.0350 0.05500 0.05500 0.05500000000	Which season in {time_frame]} saw the highest tevels of {climate_variable {climate_variable 0.807 0.727 0.	<pre>{climate_arisbl el] and {climate_varisbl el] and {climate_varisbl el] in the during during during ftime_framel}? 0.370 0.570 0.370 0.523 0.0320 0.0320 0.0320 0.0320 0.240 0.240 0.240 0.240 0.240</pre>	How does (Linexe_variable between between between location2) (Lo	variation of {climate_variabl el] in compare to that in {location1} compare to that in {location2} for {time_framel} 0.556 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0564 0.0565 0.0565 0.0565 0.0575 0.05550 0.05550 0.055500000000	Hov has sbetwase sbet	Which region in the {location} experienced the largest increase idition idition {time_frame} 0.423 0.423 0.423 0.423 0.423 0.423 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.240 0.250
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Owen2.5-VL-7B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.1-8B Llama-3.2-1B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision Llama-3.2-11B-Vision	data_modality language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and code language and code language and vision language and vision	overall_accuracy 0.580 0.486 0.457 0.413 0.317 0.312 0.312 0.318 0.286 0.286 0.286 0.286 0.286 0.281 0.281 0.233 0.204 0.173	What is the Seasonal variation of {climate_variabl {location1} during {location2} 0.627 0.638 0.627 0.648 0.627 0.648 0.627 0.630	experienced a greater change if climate variable entitioned framel {time_framel} 0.667 0.667 0.667 0.667 0.667 0.660 0.480 0.4	Which season in {fime_frame]; saw the highest levels of {clunate_variabl el}; in {location}?7 0.027 0.0307 0.0307 0.0307 0.0307 0.0301 0.0301 0.0302 0.0303 0.0304 0.0304 0.0304 0.0305 0.0306 0.0307 0.0307 0.0308 0.0309 0.0301 0.310 0.310 0.312 0.312 0.312 0.312 0.312 0.312 0.312 0.312 0.312 0.312 0.312	<pre>{climate veriabl el] and {climate veriabl e2} in the approximate approxim</pre>	How does (linet_variable between between between location2) (location2) and location2) and and and and and and and and	variation of {climate_variabl e}i in {location1} compare to that in {location2} for ftime_frame}? 0.721 0.556 0.0424 0.0360 0.0424 0.0360 0.0420 0.0300 0.0400 0.0300 0.0400 0.0300000000		<pre>Which region in the {location1} experienced the largest increase id_gest increase id_during {time_frame]? 0.537 0.423 0.420 0.4</pre>
model_name Llama-4-Maverick Llama-4-Maverick Llama-4-Scout Uama-4-Scout Owen2.5-VL-7B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.2-3B Llama-3.1-8B Llama-3.2-11B-Vision	data_modal.ity language and vision language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and vision language and vision	overall_accuracy 0.580 0.480 0.457 0.413 0.437 0.312 0.312 0.312 0.318 0.286 0.286 0.286 0.286 0.286 0.281 0.233 0.204 0.433 0.232 0.435	What is the Seasonal variation of [climate_variabl [location1] during [location1] during [location1] during [location1] during [location1] during during [location1] during	experienced a greater change iffer the spectra {time_framel} (time_framel)? 0.667 0.667 0.667 0.667 0.667 0.667 0.667 0.650 0.040 0.040 0.040 0.0350 0.050 0	Which season in {fime_frame]; saw the highest levels of {clunate_variabl el}; in {location}? 0.027 0.0307 0.0307 0.0307 0.0307 0.0301 0.0301 0.0301 0.0301 0.0301 0.0301 0.0301 0.0302 0.0303 0.0304 0.0310 0.0310 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0312 0.0314 0.0315	<pre>{climate veriabl el] and {climate veriabl e2} in the ac2 in the acc in t</pre>	How does el'extern (location2) during (time_frame)? (does during (does during during (does during during during (does during dur	variation of {climate_variabl e}; e}; n {location}; compare to that in {location}; for time_frame}; 0.721 0.556 0.0424 0.364 0.364 0.364 0.364 0.364 0.364 0.364 0.364 0.364 0.364 0.300 0.414 0.300 0.301 0.302 0.303 0.304 0.305 0.306 0.307 0.308 0.309 0.301 0.302 0.303 0.303 0.304 0.305 0.306 0.314		Which region in the {location] experienced the largest increase idition] ctimate_variabl el} during during time_frame]? 0.537 0.423 0.423 0.423 0.424 0.420 0.421 0.422 0.423 0.423 0.420 0.420 0.420 0.420 0.200
model_name Liama-4-Maverick Liama-4-Maverick Liama-4-Scout Owen2.5-VL-7B Liama-4-Scout Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.1-8B Liama-3.2-11B-Vision Liama-3.2-11B-Vision Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.2-11B-Vision Liama-3.1-8B Liama-3.1-8B Liama-3.2-11B-Vision Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.2-11B-Vision Liama-3.2-11B-Vision Liama-3.2-11B-Vision Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.1-8B Liama-3.2-11B-Vision Liama-3.1-8B Liama-3.1-	data_modal.ity language and vision language and vision language and vision language and vision language and vision language and vision language and code language and code language and code language and vision language and visi	overall_accuracy 0.580 0.480 0.457 0.413 0.437 0.312 0.311 0.311 0.314 0.314 0.314 0.314 0.264 0	What is the seasonal variation of {climate_variabl el} in {location1} during {time_frame}? 0.627 0.0463 0.0570 0.463 0.0400 0.0400 0.0470 0.0370 0.0370 0.0585 0.0464 0.0464 0.0464	experienced a greater change if climate variabl estime_framel} (time_framel)? 0.667 0.667 0.667 0.667 0.667 0.667 0.667 0.650 0.420 0.350	Which season in {time_frame]; saw the highest levels of {climate_variabl el} in {location]}7 0.0727 0.0727 0.0807 0.	<pre>{climate veriabl el] and {climate veriabl e2} in the {bccitonl} e2; in the {bccitonl} during {time_framel}? 0.373 0.517 0.370 0.510 0.523 0.0520 0.0280 0.0280 0.0280 0.0280 0.0280 0.0280 0.0280 0.0280 0.0281 0.0281 0.0240 0.240 0</pre>	How does el'event (location2) during (time_framel)? during (time_framel)? during (location2) during (locat	<pre>variation of {climate_variabl</pre>		Which region in the {location] experienced the largest increase idition] ctimate_variabl el} during during time_framel;7 0.537 0.423 0.423 0.423 0.424 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.420 0.200
model_name Liama-4-Maverick Liama-4-Maverick Liama-4-Maverick Liama-4-Scout Owen2.5-VL-7B Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.2-3B Liama-3.2-1B-Vision Liama-3.2-11B-Vision Liama-3.2-	data_modality language and vision language and vision language and vision language and vision language and voice language and vision language and vision	overal_accuracy 0.680 0.484 0.457 0.413 0.437 0.312 0.311 0.28400000000000000000000000000000000000	What is the sessonal variation of {climate_variabt during {time_frame}? 0.627 0.0463 0.0570 0.440 0.450 0.040 0.040 0.040 0.040 0.0470 0.0470 0.0470 0.0500 0.050 0.050 0.050 0.0500000000	experienced a greater change if climate variable ell throughout {time_framel} 0.667 0.667 0.667 0.667 0.667 0.667 0.667 0.667 0.667 0.670 0.670 0.670 0.03700 0.03700 0.03700 0.0370000000000	Which season in {time_frame]; saw the highest levels of {climate_variabl el} in {location]}? 0.2727 0.460 0.530 0.507 0.020 0.400	<pre>{climate veriabl el] and {climate veriabl e2} in the c2 in th</pre>	Hev des (ciast_vrish (cortion2) during (toertion2) during (cortion	variation of {climate_variabl wariabl variabl in {location}; compare to that in {location2; for 0.721 0.556 0.424 0.556 0.424 0.556 0.424 0.556 0.424 0.556 0.424 0.360 0.122 0.361 0.302 0.133 0.306 0.140 0.303 0.304 0.305 0.4100 0.301 0.302 0.140 0.303 0.304 0.305 0.457 0.265 0.265 0.265 0.265 0.266 0.314 0.7	↓ vinks {time_trans {time_trans {time_trans 0.433 0.434 0.437 0.438 0.439 <tr< td=""><td>Which region in the {location] experienced the largest increase idition] ctimate_variabl el} during during time_framel? 0.537 0.423 0.423 0.423 0.420 0.420 0.420 0.360 0.201 0.2020 0.2030 0.201 0.2020 0.2030 0.2030</td></tr<>	Which region in the {location] experienced the largest increase idition] ctimate_variabl el} during during time_framel? 0.537 0.423 0.423 0.423 0.420 0.420 0.420 0.360 0.201 0.2020 0.2030 0.201 0.2020 0.2030 0.2030

Figure 5. Evaluation results. The top table shows OpenAI models and the bottom table shows open-source models. Each row corresponds to one model with one data modality—language-only, language and code, or language and vision, while each column represents a query template in Table 1.

GeoGrid-Bench: Can Foundation Models Understand Multimodal Gridded Geo-Spatial Data?

495														
496														
497	model_name	data_modality	overall_accuracy	trend	spatial_ref	coordinate_ref	label_ref	model_name	data_modality	overall_accuracy	trend	spatial_ref	coordinate_ref	label_ref
498	o4-mini	language and vision	0.644	0.686	0.565	0.515	0.667	Llama-4-Maverick	language and vision	0.580	0.696	0.426	0.375	0.483
400	GPT-4.1	language and vision	0.578	0.668	0.451	0.451	0.655	Llama-4-Scout	language and vision	0.508	0.607	0.361	0.308	0.414
499	GPT-4.1-mini	language and vision	0.568	0.655	0.475	0.426	0.529	Llama-4-Maverick	language-only	0.486	0.517	0.287	0.472	nai
500	o4-mini	language-only	0.534	0.474	0.426	0.555	nan	Llama-4-Scout	language-only	0.457	0.478	0.269	0.463	nai
501	GPT-4o	language and vision	0.518	0.599	0.457	0.412	0.517	Qwen2.5-VL-7B	language and vision	0.413	0.445	0.380	0.342	0.345
502	GPT-4.1	language-only	0.512	0.505	0.361	0.597	nan	Liama-4-Mavenck	language and code	0.337	0.262	0.231	0.407	nar
502	GPT-4.1-mini	language-only	0.511	0.482	0.333	0.639	nan	Liama-4-Scout	language and code	0.311	0.276	0.333	0.343	nar
503	GPT-4o-mini	language and vision	0.462	0.508	0.306	0.300	0.425	Qwen2.5-VL-7B	language-only	0.298	0.309	0.148	0.250	nar
504	o4-mini	language and code	0.453	0.369	0.398	0.546	nan	Qwen2.5-VL-7B	language and code	0.286	0.317	0.222	0.250	nar
505	GPT-4o-mini	language-only	0.437	0.448	0.278	0.437	nan	Llama-3.2-3B	language and code	0.265	0.202	0.361	0.343	nar
505	GPT-4.1-mini	language and code	0.426	0.340	0.417	0.588	nan	Llama-3.1-8B	language and code	0.264	0.183	0.361	0.352	nar
506	GPT-40	language-only	0.423	0.436	0.185	0.555	nan	Llama-3.2-11B-Vision	language and code	0.261	0.191	0.330	0.330	nar
507	GPT-4.1	language and code	0.412	0.371	0.324	0.513	nan	Llama-3.2-11B-Vision	language and vision	0.233	0.297	0.274	0.267	0.176
500	GPT-4o-mini	language and code	0.369	0.366	0.343	0.328	nan	Llama-3.2-11B-Vision	language-only	0.204	0.232	0.167	0.130	nar
500	GPT-40	language and code	0.367	0.356	0.343	0.462	nan	Llama-3.1-8B	language-only	0.173	0.256	0.250	0.176	nar
509			(a)			, , , , , , , , , , , , , , , , , , ,				(E)			
510						0.2 0.3	Accuracy	0.5 0.6						
511				mo	del data_mo	dality overall_acc	uracy heatmap	heatmap_with_ann	otations heatma	p_overlayed_on_map				
511				04-1	mini language and	1 vision	0.644 0.653			0.578				
512			Li	ama-4-Mave	erick language and	1 vision	0.580 0.578		0.641	0.521				
513				GPT	-4.1 language and	1 vision	0.578 0.573		0.625	0.537				
514				GPT-4.1-	mini language and	d vision	0.568 0.568		0.620	0.517				
514				GP1	-40 language and	1 vision	0.518 0.524		0.573	0.457				
515				Llama-4-S	cout language and	i vision	0.507 0.506		0.552	0.464				
516				GP1-40-1	mini language and	1 vision	0.462 0.464		0.501	0.422				
517			Liama	Jwen2.5-VL	-78 language and	1 vision	0.077 0.000		0.458	0.365				
517			Liama	-3.2-11B-VR	sion language and	1 VISION	0.277 0.269		0.273	0.208				
518			(c)			0.30 0.35	0.40 0.45 0.50	0.55 0.60 0.65	J					
519							Accuracy							
520														

Figure 6. More evaluation results. (a) OpenAI models and (b) open-source models evaluated under different data modalities. Columns represent fine-grained answer aspects defined in Section 3, including trend, spatial references, coordinate references, and label references. There exist NaN values since the label reference is only available for the vision modality. (c) vision-language models, which are evaluated on three visualization types, as mentioned in Section 3 and Figure 3.

Which region in Philadelphia, PA experienced the largest increase in maximum annual temperature during historical period?

,			
maximum annual temperature	maximum annual temperature	maximum annual temperature	maximum annual temperature
C500 C501 C502 C503 C504 C505 C506 C507 C508 C509	56.1 57.4 58.6 59.8 61.1	96,1 57,4 58,6 58,8	an own and a second and a second at the seco
R197 NaN 56.42 56.63 57.16 57.59 57.96 58.46 58.63 58.56 NaN	C499 (C401 C402 (C404 (C40) (C		
R196 56.13 56.54 56.91 57.35 58.02 58.53 59.92 59.00 58.83 NaN		8 (555) 551 552 553 553 553 553 553 553 553 553 553	All and a second as a seco
R195 56.75 57.35 57.61 57.81 59.32 60.12 59.85 59.38 59.21 59.19	nua de la companya de	84.9427.30 59.75 57.84 84.9427.30 57.24 57.24 57.24 57.24 57.24 57.24 57.25 57	Tan H H H H H H H H
R194 57.54 57.89 58.18 59.33 59.90 60.61 59.87 59.61 59.57 59.39		84831333 5331 538 543 532 532 533 533 533 533 533 533 533 53	
R193 57.33 57.90 58.20 59.59 60.45 61.00 60.90 59.92 59.89 59.77		8(22)737 57.9 8(2)57.	Sist and the second sec
R192 57.11 57.78 58.17 59.95 60.86 61.09 59.92 59.92 60.12 60.15	șio	52.99 55.56 53.57 62.58 63.99 53.52 53.59 91.5967.53 91.5967.53	
R191 57.37 57.99 58.56 59.81 60.05 60.02 59.93 60.02 60.25 60.34		81.6958 57 59.55 68.00 69.37 59.30 69.32 6	
R190 NaN 58.38 59.16 60.89 60.16 60.06 60.05 60.15 60.29 60.47		9 (1993) 17 (0.10) (0.1	and the state of t
R189 NaN 58.91 59.63 60.00 60.01 60.03 60.22 60.36 60.51 NaN		ALEXAND 6234 6235 6235 6235 6235	AND A Contract of the Contract
R188 NaN 59.71 60.02 60.20 59.21 60.19 NaN NaN NaN NaN			Converter Annual A
(a)	(b)	(c)	(d)

Figure 7. We prepare every data sample in one of the four formats: (a) 2D table as a textual string. (b) standalone heatmap; (c) heatmap with overlaid numerical annotations at each grid cell; (d) heatmap overlaid on an actual geographic base map. These formats reflect real-world climate data practices and differ markedly from typical natural images seen by foundation models.

User Query



Figure 8. **Template 1:** Which region in {location1} experienced the largest increase in {climate_variable1} during {time_frame1}? This example takes location1 = New York city, NY, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.



Figure 9. **Template 2:** How has {climate_variable1} changed between {time_frame1} and {time_frame2} in the {location1}? This example takes location1 = New York city, NY, climate_variable1 = maximum annual temperate, time_frame1 = historical period, and time_frame2 = mid-century period (RCP-4.5).



Figure 10. Template 3: What is the correlation between {climate_variable1} and {climate_variable2} in the {location1} during {time_frame1}? This example takes location1 = New York city, NY, climate_variable1 = maximum annual temperate, climate_variable2 = minimum annual temperate, and time_frame1 = historical period.



Figure 11. Template 4: How does {climate_variable1} compare between {location1} and {location2} during {time_frame1}? This
 example takes location1 = New York city, NY, location2 = Los Angeles, CA, climate_variable1 = maximum annual temperate, and
 time_frame1 = historical period.



Figure 12. **Template 5**: What is the *seasonal* variation of {climate_variable1} in {location1} during {time_frame1}? Same data is used in **Template 6**: Which *season* in {time_frame1} saw the highest levels of {climate_variable1} in {location1}? This example takes location1 = New York city, NY, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.



Figure 13. **Template 7.** Which of {location1} or {location2} experienced a greater change in {climate_variable1} throughout {time frame1} and {time frame2}? This example takes location1 = New York city, NY location2 = Los Angeles CA, climate variable1



Figure 14. **Template 8.** How does the *seasonal* variation of {climate_variable1} in {location1} compare to that in {location2} for {time_frame1}? This example takes location1 = New York city, NY, climate_variable1 = maximum annual temperate, and time_frame1 = historical period.