GCA Framework: A GCC Countries–Grounded Dataset and Agentic Pipeline for Climate Decision Support

Muhammad Umer Sheikh[1], Khawar Shehzad[2], Salman Khan[1,3], Fahad Shahbaz Khan[1,4], Muhammad Haris Khan[1]

[1]Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2]University of Missouri, USA
[3]Australian National University, Australia
[4]Linköping University, Sweden
{muhammad.sheikh}@mbzuai.ac.ae

Abstract

Climate decision-making in the GCC states increasingly demands systems that can translate heterogeneous scientific and policy evidence into actionable guidance, yet general-purpose large language models (LLMs) remain weak both in region-specific climate knowledge and grounded interaction with geospatial and forecasting tools. We present the **GCA framework**, which unifies (i) **GCA-DS**, a curated multimodal dataset grounded in the GCC states, and (ii) **Gulf Climate Agent (GCA)**, a tool-augmented agent for climate analysis. GCA-DS comprises ∼200k question–answer pairs spanning governmental policies and adaptation plans, NGO and international frameworks, academic literature, and event-driven reporting on heatwaves, dust storms, and floods, complemented with remote-sensing inputs that couple imagery with textual evidence. Building on this foundation, the GCA agent orchestrates a modular tool pipeline grounded in real-time and historical signals and geospatial processing that produces derived indices and interpretable visualizations. Finally, we benchmark open and proprietary LLMs on climate tasks in the GCC states and show that domain fine-tuning and tool integration substantially improve reliability over general-purpose baselines.

1 Introduction

Climate change is among the most consequential challenges for societies worldwide, yet its impacts and policy responses are profoundly shaped by regional context (Hewitson et al., 2014). The Gulf region, encompassing the **United Arab Emirates, Saudi Arabia, Qatar, Kuwait, Oman, and Bahrain**, faces a unique constellation of climate hazards: extreme heat, dust storms, flash floods and rapid coastal erosion. These challenges demand decision support systems that can distill complex geophysical data and policy documents into actionable insights. While recent advances in large language models (LLMs) and vision–language models (VLMs) have made it possible to *access and summarize* climate information at scale, general-purpose models often fall short on fine-grained tasks such as numerical reasoning over climate visualizations (trend/anomaly plots), tool-mediated retrieval of spatiotemporal variables (AQI, rainfall, discharge), and jurisdiction-specific policy interpretation. High-stakes decisions in infrastructure and policy require numerically precise, source-grounded answers that adhere to domain conventions (Bulian et al., 2024; Mastrandrea et al., 2011), yet many models struggle with figure understanding and visual reasoning (Masry et al., 2022; Methani et al., 2020). Existing climate QA benchmarks either rely on small expert-curated corpora or automatically generated datasets that suffer from noise and weak validation, and they seldom incorporate multimodal evidence or region-specific context (Manivannan et al., 2025).

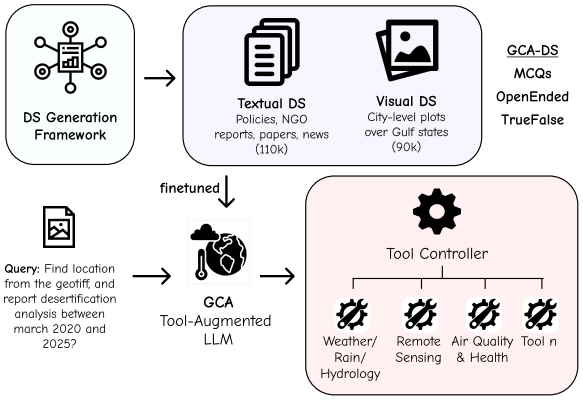

Figure 1: Overview of Gulf Climate Agent (GCA) Framework. We curate a Gulf-focused multimodal QA dataset, GCA-DS and fine-tune a tool-augmented LLM that routes user queries to specialized climate tools to produce grounded, interpretable outputs

Within climate NLP, recent benchmarks and datasets have begun to systematize evaluation and introduce multimodal supervision. For exam-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42494–42506

ple, Bulian et al. (2024) studies the adequacy of LLM responses to climate questions under a fine-grained evaluation framework; Manivannan et al. (2025) propose an automated benchmark creation pipeline (ClimaGen) and expert-in-the-loop evaluation; CPIQA introduces a figure-grounded climate QA benchmark from scientific articles, targeting retrieval-augmented setups (Mutalik et al., 2025). Despite these advances, to our knowledge there is *no* publicly available resource that (i) is explicitly Gulf-grounded across policy, hazards, and geospatial evidence at large scale, and (ii) is paired with an agentic tool-augmented pipeline designed for Gulf climate objectives.

We introduce **Gulf Climate Agent (GCA)**, a framework that bridges general-purpose LLM reasoning with specialized climate tools and datasets tailored to the Gulf (see Figure 1). **Semi-automated data-generation framework.** We construct a semi-automated Gulf-specific dataset comprising approximately **200k** question–answer pairs. The dataset is sourced from government policies and adaptation strategies, NGO reports and international frameworks, academic papers on climate and sustainability, news articles describing recent heatwaves, dust storms and floods, and geospatial/remote-sensing resources (e.g., Sentinel-2 imagery and Google Earth Engine). Our data generation pipeline combines automated extraction and synthesis with human-in-the-loop verification to enable scale while maintaining reliability, grounding responses in authoritative Gulf sources rather than generic internet text.

Agentic pipeline with climate-specific tools. We develop an agentic pipeline that orchestrates LLM reasoning with climate-specific tool suites. Our modular architecture links the LLM to specialized modules, including heat forecasting, flood-risk prediction, carbon-footprint estimation, and coastal-erosion analysis, augmented with geospatial processing to produce interpretable maps and trends (heat maps, flood indices, air-quality trajectories, shoreline-change profiles). General-purpose retrieval and retrieval-augmented generation (RAG) components complement these modules, enabling multi-step reasoning that integrates evidence from documents and structured data.

Benchmarking and fine-tuning. We fine-tune and benchmark both open and proprietary LLMs (e.g., GPT-family models, Claude, Qwen) on Gulf-centric tasks, including climate question answering, policy summarisation, tool-invocation accuracy, and geospatial reasoning. Our benchmark suite includes manually annotated 91 questions and evaluation metrics for factuality, numerical precision, and tool-use reliability. Fine-tuning on the curated dataset yields domain-specialized models that improve over general-purpose baselines, highlighting the benefits of regional adaptation and tool integration.

2 Related Work

Climate QA and Multimodal Climate Benchmarks. Climate QA has evolved from machine reading over curated documents to evaluating foundation models and retrieval-augmented generation (RAG) on technical sources. Early systems such as Climate Bot (Rony et al., 2022) demonstrated document-grounded climate QA via CCMRC. More recent benchmarks broaden scope and difficulty: ClimaQA (Manivannan et al., 2025) builds expert-in-the-loop evaluation sets from graduate-level textbooks, CPIQA (Mutalik et al., 2025) introduces figure-grounded climate QA from scientific articles for RAG settings, and MMClima (Sheikh et al., 2025) expands multimodal coverage with expert-validated QA over figures and text. Complementarily, ClimateIQA (Chen et al., 2024) targets meteorology-style visual reasoning over heatmaps. Despite these advances, existing resources are largely *global* and rarely couple regional governance documents with city-level spatiotemporal evidence, which is central to Gulf decision contexts.

Evaluating Climate Knowledge and Scientific Reasoning. Several studies assess LLM climate competence and scientific reliability, highlighting gaps between fluent generation and faithful, source-grounded responses (Bulian et al., 2024; Zhu and Tiwari, 2023; Kurfali et al., 2025). In parallel, general scientific QA benchmarks (e.g., ScienceQA, SciQAG, SciQA) provide methodology for multimodal supervision and evaluation design, but remain domain-general and do not test climate-specific tool use or geospatial reasoning (Lu et al., 2022; Wan et al., 2024; Auer et al., 2023). Related work on chart and plot QA further shows persistent failure modes in numerical extraction and reasoning over scientific visualizations (Masry et al., 2022; Methani et al., 2020), motivating tool-mediated computation for climate settings.

Agentic Tool Use for Climate Workflows. Agentic systems increasingly operationalize tool-

Dataset	Size	Automated	Validated	Multimodal	Region	Remote Sensing	Topic Covered
Climate Crisis QA (2024)	19,241	✓	✗	✗	General	✗	5
Pirá 2.0 (2024)	2,250	✗	✓	✗	General	✗	4
Climate-FEVER (2020)	1,535	✗	✓	✗	General	✗	3
CPIQA (2025)	54,612	✓	Partial	✓	General	✗	5
ELLE (2025)	1,130	✗	✓	✗	General	✗	6
ClimaQA-Gold (2025)	566	✓	✓	✗	General	✗	5
ClimaQA-Silver (2025)	3,000	✓	✗	✗	General	✗	4
MMClima (2025)	104,902	✓	✓	✓	General	✗	5
GCA-DS (ours)	**201,410**	✓	Partial	✓	Gulf-specific	✓	**12**

Table 1: Comparison of climate- and environment-focused QA datasets. "Partial" denotes limited human validation (e.g., expert-guided prompting or subset review). Topic Covered reports the number of represented topic types.

augmented reasoning for complex tasks. ClimateAgent (Kim et al., 2025) proposes a climate workflow agent and benchmark, while general tool-learning benchmarks (e.g., AgentBench, ToolLLM/ToolBench, ToolPlanner, StableToolBench) study planning, tool invocation, and evaluation stability across domains (Liu et al., 2024; Qin et al., 2023; Wu et al., 2024; Guo et al., 2024). However, these works do not provide a Gulf-grounded multimodal dataset paired with a compact climate tool suite and regression-style evaluation tailored to Gulf hazards and governance.

Summary and Positioning. Existing climate QA benchmarks and evaluations are primarily global and rarely integrate regional policy grounding with city-level multimodal spatiotemporal evidence (Table 1). To our knowledge, no prior work releases a unified Gulf-focused stack that couples (i) a large multimodal dataset with remote-sensing evidence and (ii) an agentic tool pipeline together with a dedicated tool-use regression benchmark. GCA addresses this gap by aligning Gulf-specific supervision (GCA-DS) with tool-augmented inference for decision-oriented climate queries.

3 Curation of GCA-DS

We curate a Gulf-focused multimodal climate QA dataset (GCA-DS) to support regional grounding and tool-augmented reasoning for decision-relevant queries with samples given in Figure 2. The corpus contains **200k** question–answer pairs spanning **MCQ**, **open-ended**, and **true/false** formats, with **110k** text-only items and **90k** visual-temporal items as show in Figure 3.

Textual data (110k). The textual split aggregates Gulf-relevant sources including government climate policies and adaptation strategies, NGO and international organization reports (by Gulf state), academic literature on regional climate and resilience, and news/emergency documentation covering recent extremes (heatwaves, dust storms, floods). Items are written to be evidence-grounded in the underlying documents.

Visual-temporal data (90k). The visual split is city-level across **Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the UAE**. For each city, we construct time-indexed visual artifacts (e.g., trends and anomaly plots) from standardized meteorological and land-surface variables (ERA5/ERA5-Land) (Copernicus Climate Change Service (C3S), 2026a; European Centre for Medium-Range Weather Forecasts (ECMWF), 2026; Copernicus Climate Change Service (C3S), 2026b), atmospheric composition and air-quality variables (CAMS) (Copernicus Atmosphere Monitoring Service (CAMS), 2026b,c,a,d,e), and hydrology signals for flooding (GloFAS) (Copernicus Emergency Management Service (CEMS), 2026c,b). We additionally include high-resolution CMIP6 HighResMIP simulations to support scenario-style questions beyond reanalysis (Haarsma et al., 2016).

3.1 Textual Data Generation

To construct the textual component of the gca-ds dataset, we employ a semi-automated pipeline that couples retrieval-grounded web acquisition with structured parsing and controlled QA synthesis. The design goal is to (i) maximize topical coverage across Gulf countries/cities and climate subdomains, while (ii) maintaining traceability to authoritative sources and reducing duplication and prompt-induced drift.

Retrieval-grounded keyword expansion. We begin by generating a set of *Gulf-targeted* retrieval

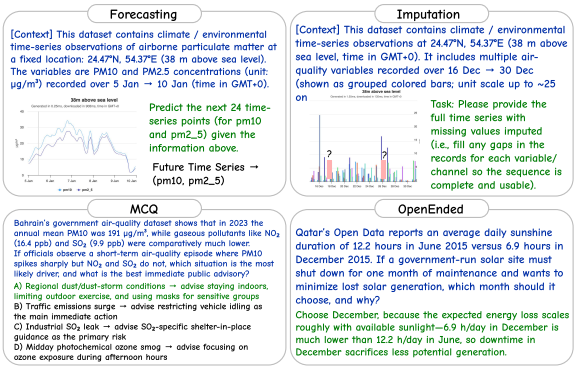

Figure 2: Example samples from the gca-ds dataset spanning text-grounded QA and visual-temporal QA over Gulf cities.

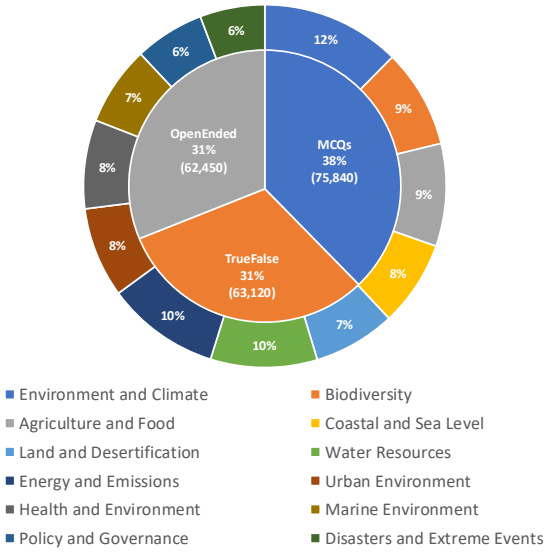

Figure 3: Distribution of GCA-DS across question types and categories.

keys that span our four textual source classes (government climate policies, NGO reports, academic papers, and event-driven news). Given an initial seed set of topic descriptors (e.g., *heatwave preparedness*, *dust storm health advisory*), we use an LLM to propose candidate keywords and query templates conditioned on *(country, city)* constraints. To prevent redundant crawling and near-duplicate keyword variants, we maintain a persistent vector index of previously used keywords. Each candidate keyword k_i is embedded into $\mathbf{e}_i$ and filtered by its maximum cosine similarity to the existing keyword set $\mathcal{K}$:

$$\text{sim}(k_i, \mathcal{K}) = \max_{k_j \in \mathcal{K}} \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\|_2 \, \|\mathbf{e}_j\|_2}, \qquad (1)$$
$$k_i \text{ kept iff } \text{sim}(k_i, \mathcal{K}) < \tau.$$

This stage yields a deduplicated pool of Gulf-conditioned queries that are subsequently used for web retrieval.

Autonomous web retrieval agent. For each keyword/query, a browsing agent issues a search and iteratively refines follow-up queries when the retrieved content is off-domain or non-authoritative. This design follows the general paradigm of *retrieval-augmented* systems that externalize knowledge access rather than relying solely on parametric memory (Lewis et al., 2020). In practice, retrieval is implemented using a combination of sparse search and dense retrieval (Karpukhin et al., 2020), and the agent maintains a lightweight interaction trace (query → click → extract) in a ReAct-style loop that interleaves reasoning with environment actions (Yao et al., 2023). The output of this stage is a set of URLs (HTML pages and PDFs) together with minimal provenance metadata (query, timestamp, source domain).

Heterogeneous document parsing. The retrieved resources include heterogeneous formats (HTML, scanned PDFs, reports with tables/figures). We therefore normalize all sources into a unified textual representation. The parser extracts main content, removes boilerplate/navigation, preserves section headers when available, and records document metadata (title, publishing organization, date, URL) to enable later citation and auditing.

Semantic segmentation and chunking. We segment each document into overlapping chunks to preserve local coherence while controlling context length. For $D = [w_1, \ldots, w_T]$, we form windows $C_m = [w_{s_m}, \ldots, w_{s_m+L-1}]$ with $L = 512$ **tokens** and **stride** $S = 384$ (128-token overlap), aligning boundaries to section/paragraph breaks when available.

Fact induction. From each chunk, we induce *atomic factual statements* that are designed to be verifiable and minimally compositional (one claim per statement when possible). This step reduces long-form narratives into a fact bank that supports diverse QA templates while improving grounding: each induced statement retains a pointer to its originating chunk and source metadata.

Instructional QA synthesis. Finally, we synthesize question–answer pairs in three formats, MCQ, open-ended, and true/false, by conditioning an

LLM on (i) one or more atomic statements, (ii) a target question type, and (iii) a rubric that enforces answerability from the provided evidence. This follows the broader idea of self-generated instructional data creation and filtering to scale instruction tuning (Wang et al., 2023). For MCQs, distractors are generated to be locally plausible but globally inconsistent with the evidence; for open-ended items, we enforce concise, evidence-anchored responses; and for true/false items, we include both entailed and contradicted variants to test factual robustness.

3.2 Visual Temporal Data Generation

The visual component is designed to capture *temporal* and *spatiotemporal* patterns that are central to Gulf climate reasoning (e.g., trend detection, anomaly identification, seasonal comparisons). We construct this split by ingesting multi-source climate and atmospheric products and rendering standardized charts that can be directly queried.

Multi-source climate data ingestion. We ingest (i) high-resolution climate simulations from CMIP6 HighResMIP models (Haarsma et al., 2016), including `CMCC-CM2-VHR4`, `FGOALS-f3-H`, `HiRAM-SIT-HR`, `MRI-AGCM3-2-S`, `EC-Earth3P-HR`, `MPI-ESM1-2-XR`, and `NICAM16-8S`; (ii) atmospheric composition and aerosol forecasts from CAMS global atmospheric composition forecasts (Copernicus Atmosphere Monitoring Service (CAMS), 2026b); (iii) air-quality reanalyses from CAMS European Air Quality Reanalysis (Copernicus Atmosphere Monitoring Service (CAMS), 2026a); and (iv) hydrological signals for flooding from GloFAS v4 reanalysis and forecasts (Copernicus Emergency Management Service (CEMS), 2026c,a). These sources support visual QA spanning meteorology (e.g., temperature, wind speed, precipitation), air quality (PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, CO, aerosols, dust), greenhouse gases (CO_2, CH_4), health-relevant indices (UV index, pollen proxies), and hydrology (river discharge).

Geocoding and city inventory. We define a Gulf city inventory for Bahrain, Kuwait, Oman, Qatar, Saudi Arabia, and the United Arab Emirates, and map each city c to its latitude–longitude pair (ϕ_c, λ_c). This enables consistent extraction from gridded products across sources.

Spatial index retrieval. Given a gridded product with coordinates $\{(\phi_i, \lambda_j)\}$, we retrieve the grid cell most relevant to a city by nearest-neighbour matching:

$$(i^*, j^*) = \arg\min_{i,j} \; d\big((\phi_c, \lambda_c), (\phi_i, \lambda_j)\big), \quad (2)$$

where $d(\cdot, \cdot)$ is a spherical distance (or an equirectangular approximation when appropriate). This step yields a consistent time series per variable and city, and it generalizes across heterogeneous grids (0.05°–0.1°, model-dependent).

Unit and schema normalization. Because sources differ in naming conventions, units, time encoding, and missingness, we map each retrieved series into a canonical schema. This includes unit conversion, timestamp normalization to a unified calendar, and harmonized variable naming. Each processed record is written to a universal CSV format that serves as both (i) the canonical data representation and (ii) metadata for plot generation and QA synthesis.

Temporal plot builder. For each city–variable pair, we segment the time series into non-overlapping 3-month windows over the last 10 years. Let $\Delta = 90$ days. For window index t, we define:

$$W_t = [t\Delta, (t+1)\Delta), \quad (3)$$

and drop windows where coverage is below a minimum completeness threshold (e.g., $< \rho$ observed timesteps). We then render standardized charts for each window; the corresponding CSV provides plot metadata (location, variable, units, time span, summary statistics), enabling controlled and reproducible QA generation.

Visual QA synthesis. Using the chart and its metadata, we synthesize QA items in MCQ, open-ended, and true/false formats across four primary categories: (i) anomaly detection (e.g., identifying abnormal spikes/drops), (ii) forecasting (trend continuation or seasonal projection from context), (iii) imputation (fill missing segments based on surrounding values), and (iv) reasoning QA (comparisons, aggregations, and multi-variable interpretation grounded in metadata). The resulting visual split contributes approximately **90k** QA pairs, complementing the textual split and enabling multimodal evaluation of climate reasoning in the Gulf.

4 GCA Agentic Architecture

Gulf Climate Agent (GCA) is a tool-augmented language agent designed to answer Gulf-focused

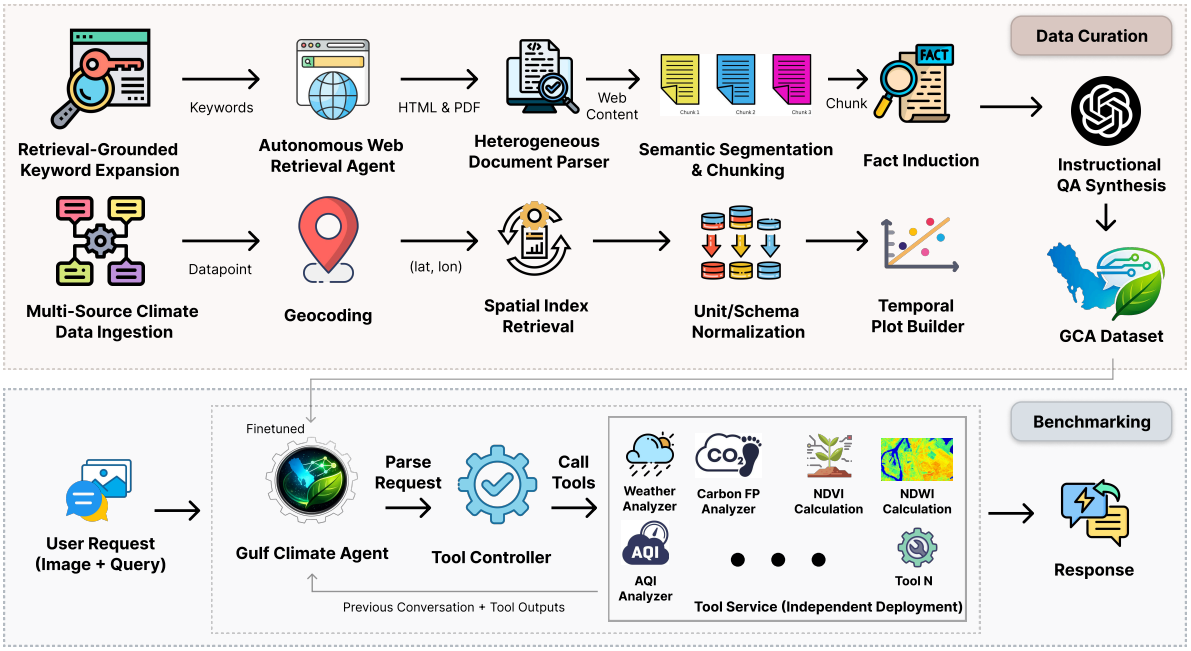

Figure 4: Gulf Climate Agent (GCA) framework. The figure summarizes multimodal dataset curation for text and visual-temporal sources, producing the gca-ds dataset, and the tool-augmented inference stack in which a fine-tuned LLM routes user requests through a tool controller to specialized climate services to generate grounded responses.

climate queries by combining (i) domain grounding from our curated multimodal dataset GCA-DS (§3) and (ii) structured access to specialised climate analytics tools as shown in Figure 4. Rather than relying on parametric knowledge alone, GCA follows an *act–observe–reason* paradigm in which the model selects tools, consumes their outputs, and iteratively refines intermediate hypotheses until it can produce an evidence-grounded response (Yao et al., 2023).

4.1 Tool Suite Overview

To keep the interface compact while covering Gulf climate objectives, we group tools into six categories (full signatures in Appendix A.1): **Remote Sensing and Land Surface** (satellite retrieval, NDVI/NDWI, desertification change analysis), **Biodiversity and Species** (bird-call and species recognition), **Web Retrieval and Summarization** (targeted search and policy/event summarization), **Carbon and Sustainability** (country/sector footprint estimation), **Air Quality and Atmospheric Composition** (AQI inquiry/forecasting/trends plus UV and pollen), and **Weather, Climate, and Hydrology** (weather/rain inquiry and forecasting, river discharge for floods, with geocoding for location resolution). This abstraction supports concise prompting and systematic evaluation of *tool invocation* as a first-class capability.

4.2 Binding Tools to the LLM

Each tool is exposed to the LLM through a structured function signature (name, arguments, and return schema). At inference time, the model receives: (i) the user query, (ii) brief tool descriptions grouped by category, and (iii) a specification requiring *typed* tool calls. The model produces either a direct answer or a tool call with arguments. Tool outputs are returned verbatim (plus metadata such as units, timestamps, and location), and are appended to the model context as *observations*.

To reduce brittle behavior, we standardize outputs across tools (e.g., normalized units, consistent timestamp formats, and explicit uncertainty when available). This allows the agent to compose multiple tools in a single query (e.g., geocode $\rightarrow$ weather_analysis $\rightarrow$ summarize) without ad hoc parsing.

4.3 Agentic Reasoning and Control

GCA adopts an iterative control loop for tool-augmented reasoning. Given an input query x, the agent maintains a trajectory of intermediate steps $\mathcal{T} = \{(a_t, o_t)\}_{t=1}^{T}$, where a_t is an action (tool call or final response) and o_t is the resulting observation. At each step, the agent first infers the dominant

Model	Format Err. (%)	Arg. Err. (%)	N/A (%)
GPT-5	19.8	11.9	6.4
Claude 4.5 Sonnet	18.6	14.7	7.1
Gemini 2.5 Pro	21.3	16.5	6.8
Qwen2.5-VL 7B	46.7	38.2	21.5
Pixtral-12B	39.4	29.8	15.2
GCA (ours)	**12.7**	**8.3**	**3.9**

Table 2: Percentages of tool-use error types on GCA. Format errors: invalid tool-call structure; argument errors: incorrect/missing schema fields; N/A: no actionable tool call when required

intent, *textual* (policy/event reporting), *numerical* (time-series inquiry), *geospatial* (satellite-derived indices), or *health/environmental* (air quality, UV, pollen), and routes to a minimal set of tool categories. It then selects a tool and arguments, executes the call, and validates schema and units; if the query requires multi-step computation, it chains additional tools and aggregates the resulting evidence. Finally, it synthesizes a response grounded in tool outputs, briefly explaining derived quantities and, when relevant, summarizing temporal trends.

5 Experiments and Results

We evaluate Gulf Climate Agent (GCA) along two axes: (i) **domain adaptation** via a single parameter-efficient fine-tuning run on our Gulf-focused multimodal dataset (§3), and (ii) **agentic competence** via a regression benchmark that probes structured tool use under a standardized agent interface. Concretely, we assess whether a model can (a) follow tool-call formatting, (b) choose the correct tool, (c) supply valid arguments, and (d) synthesize a faithful response from tool outputs.

5.1 Fine-tuning Setup

We adopt **Qwen2.5-VL 7B** as the central backbone and apply LoRA-based parameter-efficient fine-tuning (Hu et al., 2022). We use AdamW with 8-bit optimization, a cosine learning-rate schedule, weight decay 0.1, learning rate 5×10^{-5} (embedding LR 1×10^{-5}), and LoRA settings r=16, α=16 targeting `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`. We train on a unified instructional format spanning MCQ, open-ended, and true/false questions across both textual and visual splits, optimizing next-token prediction over formatted samples. This setup exposes the model to Gulf-specific variables (e.g., heat, dust-related air-quality signals, rainfall and discharge) and to tool-facing interaction patterns (location/time/units). As a result, fine-tuning targets both *domain grounding* and *tool-execution reliability* rather than general QA alone.

5.2 Agentic Benchmark for Tool Usage

To test whether models reliably *use* Gulf-relevant climate tools (rather than only describing them), we construct a regression benchmark over the final tool suite (Appendix A.1). Each instance provides a user query, tool signatures, and a gold tool-usage trace with one or more calls. The benchmark emphasizes multi-step workflows typical of Gulf climate analysis, such as `geocode_mapping` followed by temporal `weather_analysis`, or two-date satellite retrieval followed by index computation and change analysis.

Evaluation modes and metrics. We report results in two complementary evaluation modes (Table 3). **Step-by-step mode** evaluates each step against gold traces: **InstAcc** (instruction-following for the step), **ToolAcc** (correct tool choice), **ArgAcc** (correct argument *names*/schema fields), and **SummAcc** (step-conditioned summary correctness given tool outputs). **End-to-end mode** evaluates the full execution outcome: **AnsAcc** measures final answer accuracy after executing the predicted tool trace, and **AnsAcc+I** enables image generation during response composition (useful when answers require visual explanation of temporal trends).

Error taxonomy. To diagnose failures, Table 2 summarizes tool-use error rates: **Format Err.** (invalid tool-call structure), **Arg. Err.** (incorrect or missing schema fields), and **N/A** (no actionable tool call when a tool is required). These error types map directly to common agent breakdowns in practice: format errors prevent execution entirely, argument errors yield invalid or misleading tool runs, and N/A reflects premature "direct answering" without tool grounding.

Baselines. We compare against representative proprietary and open models under identical prompting, tool schemas, and execution harness: **GPT-5**, **Claude 4.5 Sonnet**, **Gemini 2.5 Pro**, **Qwen2.5-VL 7B**, and **Pixtral-12B** (`mistralai/pixtral-12b`). All models operate in a tool-augmented setting where they may either answer directly or emit structured tool calls.

5.3 Main Results

Table 3 reports the main benchmark results. Overall, **GCA is best or competitive across all metrics**

Model	Step-by-Step Mode ↑				End-to-End Mode ↑	
	InstAcc	ToolAcc	ArgAcc	SummAcc	AnsAcc	AnsAcc+I
GPT-5	88.6	92.3	**90.2**	87.4	86.3	87.0
Claude 4.5 Sonnet	86.8	90.2	88.3	85.6	84.3	84.8
Gemini 2.5 Pro	85.9	89.4	87.2	84.7	83.2	84.1
Qwen2.5-VL 7B	60.5	62.2	58.3	55.6	52.3	54.0
Pixtral-12B	66.9	68.3	64.2	60.8	56.2	58.5
GCA (ours)	**89.4**	**94.2**	89.3	**88.6**	**88.2**	**89.1**

Table 3: Tool-use benchmark results on GCA. Step-by-step mode reports instruction-following accuracy (InstAcc), tool selection accuracy (ToolAcc), argument name accuracy (ArgAcc), and step-conditioned summary accuracy (SummAcc). End-to-end mode reports final execution answer accuracy (AnsAcc) and answer accuracy with image generation enabled (AnsAcc+I).

and substantially improves tool reliability relative to its base backbone. In **step-by-step mode**, GCA attains the highest **ToolAcc** (94.2) and **SummAcc** (88.6), with strong **InstAcc** (89.4) and **ArgAcc** (89.3). In **end-to-end mode**, GCA achieves the highest **AnsAcc** (88.2) and **AnsAcc+I** (89.1), indicating that the gains in structured execution translate into improved final answers rather than only cleaner traces.

Effect of fine-tuning. Comparing GCA to the *unadapted* Qwen2.5-VL 7B baseline highlights the impact of a single LoRA run on tool competence. GCA improves by **+32.0** ToolAcc (94.2 vs. 62.2), **+31.0** ArgAcc (89.3 vs. 58.3), and **+33.0** SummAcc (88.6 vs. 55.6). These step-level gains compound into a large end-to-end improvement of **+35.9** AnsAcc (88.2 vs. 52.3) and **+35.1** AnsAcc+I (89.1 vs. 54.0). The error breakdown in Table 2 is consistent with this effect: relative to Qwen2.5-VL 7B, GCA sharply reduces **Format Err.** (12.7% vs. 46.7%), **Arg. Err.** (8.3% vs. 38.2%), and **N/A** (3.9% vs. 21.5%). This indicates that fine-tuning primarily improves (i) emitting executable tool calls and (ii) adhering to schema constraints, which are prerequisites for successful multi-step tool composition.

Comparison to strong proprietary models. GCA remains competitive with proprietary baselines. Against GPT-5, GCA improves **ToolAcc** by **+1.9** (94.2 vs. 92.3), **SummAcc** by **+1.2** (88.6 vs. 87.4), and **AnsAcc** by **+1.9** (88.2 vs. 86.3), while staying close on argument naming accuracy (89.3 vs. 90.2). Claude 4.5 Sonnet and Gemini 2.5 Pro trail GCA more clearly on end-to-end accuracy (84.3/83.2 vs. 88.2). Importantly, Table 2 shows that GCA also exhibits the lowest error rates among all evaluated models, suggesting that Gulf-specific training improves not only answer quality but also execution robustness.

Why do scores improve? We provide two quantitative signals that the gains are driven by improved *execution behavior* rather than stylistic differences. First, Table 2 shows that GCA substantially reduces tool-call failures: format errors drop to 12.7% (vs. 46.7% for Qwen2.5-VL 7B; 19.8–21.3% for GPT-5/Gemini), argument errors drop to 8.3% (vs. 38.2%), and "no-action" cases drop to 3.9% (vs. 21.5%). These reductions explain higher end-to-end accuracy because invalid or missing calls cannot be executed and therefore cannot yield grounded answers. Second, Table 3 indicates that improvements concentrate on the steps most sensitive to tool execution: relative to Qwen2.5-VL 7B, GCA improves **ToolAcc** (+32.0) and **ArgAcc** (+31.0), which in turn raises **SummAcc** (+33.0) and final **AnsAcc** (+35.9). Together, the lower error rates and the step-level gains provide direct empirical evidence that fine-tuning improves structured tool use (formatting + schema adherence), which is the dominant driver of downstream answer improvements.

6 Conclusion

We introduced **Gulf Climate Agent (GCA)**, a Gulf-focused, tool-augmented climate assistant that bridges general-purpose LLM/VLM reasoning with regional climate evidence and specialised analytics. The proposed **GCA framework** contributes (i) a semi-automated curation pipeline producing **GCA-DS**, a **multimodal** Gulf dataset of ∼**200k** QA pairs spanning policies, reports, literature, event news, and city-level visual-temporal variables; (ii) a modular tool suite for remote sensing, air quality/health indices, weather/rainfall,

hydrology, carbon, web retrieval/summarization, and geocoding; and (iii) a benchmarking and fine-tuning study showing improved tool-use reliability and end-to-end answer accuracy over strong baselines. Future work will expand continual data ingestion, strengthen validation, and add uncertainty-aware and scenario-driven analyses for Gulf climate objectives.

7 Limitations

GCA is designed for Gulf-focused climate decision support, but it has several limitations. First, **coverage and validation** of GCA-DS are semi validated: although the dataset is large and semi-automatically curated, only a subset is human-verified, and automated QA generation can introduce subtle label or grounding errors (especially for long policy documents and event reports). Second, **tool dependence** constrains reliability: the agent's outputs inherit biases, resolution limits, and missing-data patterns from upstream data products and analytical modules (e.g., spatiotemporal gaps, differing units/definitions across sources). Tool failures can also propagate through multi-step traces, even when the language model's intent is correct. Third, **benchmark scope** is necessarily selective: our regression suite emphasizes Gulf-relevant tool workflows and does not exhaustively cover all climate tasks (e.g., full uncertainty quantification, attribution, or long-horizon scenario planning). Finally, **deployment considerations** remain: real-world use requires careful monitoring, transparent provenance, and domain-expert oversight, particularly for high-stakes recommendations where uncertainty and stakeholder constraints must be explicitly communicated.

References

Sören Auer, Dante A. C. Barone, Cassiano Bartz, Eduardo G. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Dmitry Mouromtsev, and 1 others. 2023. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13:7240.

Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Huebscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauss. 2024. Assessing large language models on climate information. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 4884–4935. PMLR.

Jian Chen, Peilin Zhou, Yining Hua, Dading Chong, Meng Cao, Yaowei Li, Zixuan Yuan, Bing Zhu, and Junwei Liang. 2024. Vision-language models meet meteorology: Developing models for extreme weather events detection with heatmaps. *Preprint*, arXiv:2406.09838.

Copernicus Atmosphere Monitoring Service (CAMS). 2026a. CAMS European air quality reanalyses. Copernicus Atmosphere Data Store (ADS). Accessed 2026-01-04.

Copernicus Atmosphere Monitoring Service (CAMS). 2026b. CAMS Global atmospheric composition forecasts. ECMWF datasets. Accessed 2026-01-04.

Copernicus Atmosphere Monitoring Service (CAMS). 2026c. CAMS global greenhouse gas forecasts. Copernicus Atmosphere Data Store (ADS). Accessed 2026-01-04.

Copernicus Atmosphere Monitoring Service (CAMS). 2026d. CAMS UV index forecasts. Copernicus Atmosphere Monitoring Service (CAMS) charts. Accessed 2026-01-04.

Copernicus Atmosphere Monitoring Service (CAMS). 2026e. Ground-level pollen forecast (CAMS). European Climate and Health Observatory / Climate-ADAPT. Accessed 2026-01-04.

Copernicus Climate Change Service (C3S). 2026a. ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Data Store (C3S). Accessed 2026-01-04.

Copernicus Climate Change Service (C3S). 2026b. ERA5-Land hourly data from 1950 to present. Copernicus Climate Data Store (C3S). Accessed 2026-01-04.

Copernicus Emergency Management Service (CEMS). 2026a. GloFAS forecasts: product technical information. Copernicus Emergency Management Service (CEMS) Global Flood Awareness System. Accessed 2026-01-04.

Copernicus Emergency Management Service (CEMS). 2026b. GloFAS v4.0 basics: hydrological model and implementation set-up. Copernicus Emergency Management Service (CEMS) / ECMWF Confluence. Accessed 2026-01-04.

Copernicus Emergency Management Service (CEMS). 2026c. GloFAS v4.0 hydrological reanalysis. Joint Research Centre (JRC) Data Catalogue / Copernicus Emergency Management Service. Accessed 2026-01-04.

European Centre for Medium-Range Weather Forecasts (ECMWF). 2026. ECMWF Reanalysis v5 (ERA5). ECMWF dataset description. Accessed 2026-01-04.

Zhicheng Guo, Sijie Cheng, Hao Wang, Shihao Liang, Yujia Qin, Peng Li, Zhiyuan Liu, Maosong Sun, and Yang Liu. 2024. Stabletoolbench: Towards stable large-scale benchmarking on tool learning of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11143–11156, Bangkok, Thailand. Association for Computational Linguistics.

R. J. Haarsma, M. J. Roberts, P. L. Vidale, C. A. Senior, A. Bellucci, Q. Bao, P. Chang, and 1 others. 2016. High resolution model intercomparison project (highresmip v1.0) for cmip6. *Geoscientific Model Development*, 9:4185–4208.

Bruce Hewitson, Anthony C. Janetos, Timothy R. Carter, Filippo Giorgi, Richard G. Jones, Won-Tae Kwon, Linda O. Mearns, E. Lisa F. Schipper, and Maarten K. van Aalst. 2014. Regional context. In V. R. Barros, C. B. Field, D. J. Dokken, M. D. Mastrandrea, K. J. Mach, T. E. Bilir, M. Chatterjee, K. L. Ebi, Y. O. Estrada, R. C. Genova, B. Girma, E. S. Kissel, A. N. Levy, S. MacCracken, P. R. Mastrandrea, and L. L. White, editors, *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 1133–1197. Cambridge University Press, Cambridge, United Kingdom and New York, USA. Working Group II contribution to the IPCC Fifth Assessment Report.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Hyeonjae Kim, Chenyue Li, Wen Deng, Mengxi Jin, Wen Huang, Mengqian Lu, and Binhang Yuan. 2025. Climateagent: Multi-agent orchestration for complex climate data science workflows. *Preprint*, arXiv:2511.20109.

Murathan Kurfali, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. 2025. Climateeval: A comprehensive benchmark for nlp tasks related to climate change. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, Vienna, Austria. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations (ICLR)*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Øyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2025. Climaqa: An automated evaluation framework for climate question answering models. In *International Conference on Learning Representations (ICLR)*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Michael D. Mastrandrea, Katharine J. Mach, Gian-Kasper Plattner, Ottmar Edenhofer, Thomas F. Stocker, Christopher B. Field, Kristie L. Ebi, and Patrick R. Matschoss. 2011. The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. *Climatic Change*, 108(4):675–691.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1527–1536.

Rudra Mutalik, Abiram Panchalingam, Loitongbam Gyanendro Singh, Timothy J. Osborn, Ed Hawkins, and Stuart E. Middleton. 2025. Cpiqa: Climate paper image question answering dataset for retrieval-augmented generation with context-based query expansion. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 218–232, Vienna, Austria. Association for Computational Linguistics.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *Preprint*, arXiv:2307.16789.

Md Rashad Al Hasan Rony, Ying Zuo, Liubov Kovriguina, Roman Teucher, and Jens Lehmann. 2022. Climate bot: A machine reading comprehension system for climate change question answering. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22), AI for Good - Demos*, pages 5249–5252.

Muhammad Umer Sheikh, Hassan Abid, Khawar Shehzad, Ufaq Khan, and Muhammad Haris Khan. 2025. Mmclima: A framework for multimodal climate science data and evaluation. `https://openreview.net/forum?id=j9TdFswuZ3`. OpenReview preprint.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. 2024. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *Preprint*, arXiv:2405.09939.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Qinzhuo Wu, Wei Liu, Jian Luan, and Bin Wang. 2024. Toolplanner: A tool augmented llm for multi granularity instructions with path planning and feedback. *Preprint*, arXiv:2409.14826.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Hongyin Zhu and Prayag Tiwari. 2023. Climate change from large language models. *Preprint*, arXiv:2312.11985.

A Appendix

A.1 Tool Suite

Tool	Purpose	Inputs	Outputs
Remote sensing and land surface			
`get_satellite_image`	Retrieve a multispectral satellite image for a coordinate and date, used for downstream index and change analyses.	`lat, lon, date`	`image` + metadata.
`calculate_ndvi`	Compute NDVI from an image to quantify vegetation condition (scalar map + summary).	`image`	`ndvi_map` + stats.
`calculate_ndwi`	Compute NDWI from an image to highlight water/moisture signals (scalar map + summary).	`image`	`ndwi_map` + stats.
`desertification_analysis`	Compare two images and return land-degradation indicators (e.g., index deltas and affected area).	`image1, image2`	`change_map` + metrics.
Biodiversity and species			
`detect_bird`	Recognize bird calls from audio, returning candidate species with confidence.	`audio_clip`	(`species, conf`) list.
`detect_species`	Classify plant/animal species from an image, returning candidates with confidence.	`image`	(`species, conf`) list.
Web retrieval and summarization			
`online_search`	Targeted search for policies, reports, and event coverage; returns ranked results with snippets.	`query`	(`title,url,snippet`) list.
`summarize`	Produce a concise summary preserving key facts and implications.	`text`	`summary`.
Carbon and sustainability			
`carbon_footprint_calculation`	Estimate annual emissions for a country/industry/year given revenue (factor-based).	`country, industry, year, revenue`	tCO_2e.
Air quality and health indices			
`aqi_inquiry`	Return AQI and pollutant values for a location and date.	`lat, lon, date`	`aqi` + pollutants.
`aqi_prediction`	Forecast AQI for a location over a specified horizon.	`lat, lon, horizon`	AQI time series.
`aqi_analysis`	Summarize AQI trends and exceedances over a date range.	`lat, lon, start, end`	Stats + trend + exceedances.
`pollen_forecast`	Return forecast pollen levels for a location.	`lat, lon`	`pollen_levels`.
`uv_index_forecast`	Return UV index forecast for a location.	`lat, lon`	UV time series.
Weather, rainfall, and hydrology			
`weather_inquiry`	Return historical weather variables for a location and date.	`lat, lon, date`	Weather dict.
`weather_forecast`	Return weather forecast for the next n days.	`lat, lon, days`	Forecast series.
`weather_analysis`	Compute summary statistics and anomalies over a date range.	`lat, lon, start, end`	Stats + anomalies.
`rain_inquiry`	Return precipitation for a location and date.	`lat, lon, date`	`precip` (mm).
`rain_prediction`	Forecast precipitation for a location over a specified horizon.	`lat, lon, horizon`	Precip series.
`rain_analysis`	Summarize rainfall patterns and extremes over a date range.	`lat, lon, start, end`	Stats + events.
`river_discharge_check`	Return simulated river discharge for the nearest river grid cell at a date.	`lat, lon, date`	`discharge` (m^3/s).
Geospatial utility			
`geocode_mapping`	Resolve a region/city name to coordinates for downstream tool calls.	`region`	`lat, lon` (+ metadata).

Table 4: **Tool Suite Details.** GCA tool suite. Each tool is exposed to the LLM with typed inputs/outputs; outputs are normalized (units, timestamps) to support multi-step composition.