Designing the AI Interpreter: A Governance-First Framework for LLM-Mediated Environmental Data Communication

Michael Alvear[1], Mark Schwartz[1], Luis Cabrera[1], Christopher Smith[1], Mewcha Amha Gebremedhin[2,5], Ahmed S. Elshall[2,3,5]

[1] Department of Computing and Software Engineering, U.A. Whitaker College of Engineering, Florida Gulf Coast University, Fort Myers, Florida

[2] Department of Bioengineering, Civil Engineering, and Environmental Engineering, U.A. Whitaker College of Engineering, Florida Gulf Coast University, Fort Myers, Florida

[3] The Water School, Florida Gulf Coast University, Fort Myers, Florida

[4] DENDRITIC: A Human-Centered Artificial Intelligence and Data Science Institute, Florida Gulf Coast University, Fort Myers, Florida

[5] Faculty of Mining and Geosciences, Mekelle University, Mekelle

*Corresponding author (aelshall@fgcu.edu)

Highlights

- Framework to turn groundwater & climate data to natural language answers for the public
- Considering overlooked design considerations in the emerging field of geospatial AI agents
- Architecture of prototype AI agent with governance for trust & transparency
- Florida case shows opportunities and risks in real contexts

Abstract

Large Language Models (LLMs) offer a compelling opportunity to democratize access to complex environmental datasets by translating scientific data into plain language explanations accessible to non-technical stakeholders. A growing number of systems now pursue this goal across hydrology, flood risk management, and climate science. However, these systems have emerged as bespoke engineering solutions to specific use cases, without providing an explicit shared vocabulary for reasoning about their design choices or the risks each choice introduces. Through the development and analysis of a prototype LLM-mediated interface for South Florida groundwater and climate data, and through comparative examination of analogous systems in the literature, we identify four recurring functional concerns that such systems must address: data access, domain-adapted reasoning, natural language synthesis, and governance. To establish a standard architectural approach, we propose these concerns be concretized as an explicit four-layer design pattern (Access, Understanding, Translation, and Governance) and derive a set of layer-specific design considerations from the common failure modes observed in comparable systems. Our central finding is that the Governance Layer cannot be treated as a supplementary safeguard applied after the other layers are designed; it must serve as the organizing constraint from which Access, Understanding, and Translation decisions follow. We demonstrate this principle through CLIO (Climate and Land data Interpretation Oracle), a proof-of-concept implementation grounded in the Biscayne and Southern Everglades Coastal Transport (BISECT) hydrological model outputs, which provides of hydrological condition projections for one of the most climate-vulnerable coastal regions in the United States. The resulting framework is domain-agnostic and provides a reusable blueprint for responsible LLM-mediated scientific data communication across environmental and other data-rich domains.

Keywords

AI in hydrology, large language models, AI agents, geospatial artificial intelligence (GeoAI), democratization of climate data, groundwater, water resources

1. Introduction

The effective translation of scientific predictions into actionable decisions is a persistent challenge in environmental governance, particularly as communities confront accelerating climate change (Cabana et al., 2023; Cammarano et al., 2023; Elshall et al., 2020; Reichert et al., 2015). Over the past two decades, the hydrology and climate science communities have made substantial progress in making specialized datasets publicly available, from groundwater monitoring networks and hydrological simulations to global climate projections and coastal risk assessments as well as interactive interfaces through which those datasets can be explored (Condon et al., 2021; Eyring et al., 2016; Huggins et al., 2025; Iturbide et al., 2022; Moeck et al., 2020). This effort has been guided by the FAIR (Findable, Accessible, Interoperable, Reusable) data principles (Iturbide et al., 2022; Wilkinson et al., 2016), which have made substantial advances in data findability, interoperability, and infrastructure-level accessibility, primarily for machines and expert systems. Yet a persistent gap separates data availability from data usability. The stakeholders who most need environmental projections — municipality leaders, agricultural managers, infrastructure planners, and community organizations — are rarely equipped with the specialized tools, domain training, or computational resources that meaningful engagement with these archives demands (Dosemagen & Williams, 2022; Elliott, 2022; Hewitt et al., 2017; Koldasbayeva et al., 2024; Mabon, 2020). The result is a bottleneck between scientific infrastructure and the decision-making it is intended to support.

Large Language Models (LLMs) have emerged as a compelling mechanism for bridging this gap. Unlike programmatic APIs or interactive dashboards, which require users to conform to the organizational logic of the data, LLMs accept queries in natural language and can serve as an interpretive layer between complex model outputs and non-specialist audiences, enabling a policy stakeholder or community planner to engage with scientific results without specialized training, though how this interpretive role should be structured, and with what safeguards, is less settled than demonstrations of its feasibility suggest. A growing body of implemented systems demonstrates this potential across environmental domains: voice-enabled platforms that allow users to retrieve and visualize federal hydrological data through conversational commands (Ramirez & Demir, 2026); GPT-4-powered assistants that synthesize real-time flood warnings and geospatial data into plain language risk advice for non-technical publics (Martelo et al., 2026); LLM-orchestrated water resources management systems that decompose user requests into multi-step data ingestion and analysis workflows (He et al., 2025); and local climate services that translate regional model outputs into accessible explanations for non-specialist stakeholders (Koldunov & Jung, 2024). Beyond this communicative role, LLMs have also shown promise for accelerating scientific discovery itself (Boiko et al., 2023; X. Ji et al., 2025; Romera-Paredes et al., 2024; Wang et al., 2023),

though the role pursued here is distinct: extending expert knowledge to non-expert audiences rather than augmenting expert research workflows.

These systems represent a meaningful advance, but they have emerged largely as bespoke engineering solutions to specific use cases, each making implicit design choices about how the LLM relates to data retrieval, domain knowledge, response generation, and output accountability without an explicit shared vocabulary for reasoning about those choices or the failure modes each choice introduces. This absence has a practical consequence: it makes the risks of particular design decisions difficult to anticipate before deployment, difficult to compare across systems, and difficult to communicate to practitioners building new systems. The failure modes in question are not trivial, and in the context of AI-assisted environmental decision support they carry particular weight: the data concerns physical systems, such as water supply and flood risk, where an undetected error could conceivably lead to a consequential and irreversible decision. The risk is compounded by the fact that the non-specialist stakeholders these systems are built to serve are the audience least equipped to independently verify what the system tells them. LLMs placed in control of data retrieval exhibit unreliable spatial reasoning when translating natural language descriptions into geographic coordinates (Kizilkaya, Sajja, et al., 2025; Yan et al., 2023), degraded tool selection accuracy as toolkit complexity grows, leading to malformed calls and compounding errors across multi-step retrieval sequences (Gan & Sun, 2025), and a systematic tendency to draw on memorized training knowledge rather than retrieved evidence, even when the two are in direct conflict (Xie et al., 2023). LLMs generating natural language responses can also produce confident, grammatically correct outputs that are indistinguishable from correctly attributed claims, a well-documented hallucination risk (Z. Ji et al., 2023) that is especially dangerous when the intended audience lacks the domain expertise to detect errors (L. Zhang et al., 2025). In high-stakes scientific domains like groundwater and coastal hydrology, where a misreported salinity trend or an incorrectly bounded geographic query can mislead decisions about flood risk, water supply, or saltwater intrusion, the tolerance for undetected errors is, as it should be, very low.

This paper addresses the vocabulary gap directly. Through the development and analysis of a prototype LLM-mediated interface for South Florida groundwater and climate data, and through comparative examination of the analogous systems described above, we identify four recurring functional concerns that any such system must address: how it connects the LLM to quantitative data (data access); how it equips the LLM to reason correctly about domain-specific scientific content (domain-adapted reasoning); how it generates responses that are both accurate and accessible to non-specialists (natural language synthesis); and how it ensures that users can detect, attribute, and correct errors in LLM-generated outputs (governance). To elucidate a standard architectural approach, we propose that these

concerns be concretized as an explicit four-layer design pattern: Access, Understanding, Translation, and Governance; and derive a set of layer-specific design considerations from the failure modes observable in existing systems. This formalization creates a shared vocabulary for describing and evaluating design choices across systems, and makes design tradeoffs legible: choices in the Access Layer have direct consequences for what the Governance Layer can promise, and these dependencies are easier to reason about when the layers are explicitly named.

Our central architectural finding is that Governance cannot be treated as a supplementary safeguard applied after the other layers are designed. A system that places the LLM in the data access pipeline provides no mechanism through which a non-specialist user can independently verify its numerical outputs, because evaluating whether the LLM's retrieval steps were correct requires precisely the domain expertise such systems are designed to replace. Governance requirements must therefore serve as the organizing design constraint from which Access, Understanding, and Translation decisions follow, not the other way around. We demonstrate this principle through CLIO (Climate and Land data Interpretation Oracle), a proof-of-concept implementation grounded in the Biscayne and Southern Everglades Coastal Transport (BISECT) hydrological model, which provides high-resolution, decision-relevant projections for one of the most climate-vulnerable coastal regions in the United States (Gebremedhin et al., 2026; Swain et al., 2019). In CLIO, a user selects a geographic region and time period through an interactive map; a deterministic data processing pipeline computes a verified statistical summary of the selection; and the LLM receives only this pre-computed snapshot as its numerical ground truth, with no ability to query or modify the underlying dataset. Every numerical claim the LLM makes can be cross-referenced by the user against the displayed data panel. Every citation it provides is traceable to a specific source document and page number through a provenance widget visible alongside each response. These mechanisms operationalize a Human-on-the-Loop (HOTL) governance pattern (Baum & Laux, 2026; Parasuraman et al., 2000), in which the system operates automatically while maintaining sufficient transparency for a non-specialist user to identify and correct errors without interrupting every interaction. The resulting framework is domain-agnostic: any field in which a data-to-model-to-policy pipeline exists can instantiate the same architecture by replacing domain-specific data processing and knowledge base components while preserving the LLM's interpretive role and the governance mechanisms that bound it.

The following section develops the four-layer design pattern, describing each layer's responsibilities, design considerations, and the failure modes that motivate them. Section 3 introduces the South Florida case domain and the BISECT dataset. Section 4 describes the technical implementation of CLIO as a concrete instantiation of the framework. Section **Error! Reference source not found.** presents a qualitative demonstration of the system,

discusses its current limitations and future directions, and poses open questions for the research community. Section 6 concludes.

2. Conceptual Background

The systems mentioned in the introduction share a common set of functional requirements. In broad outline, any such system must connect the LLM to relevant scientific data, equip it to reason about that data correctly, generate accessible natural language responses, and maintain some form of output accountability. They differ substantially, however, in how each of these concerns is resolved. This section develops the vocabulary the introduction identified as missing. We organize the four recurring functional concerns as named architectural layers (Access, Understanding, Translation, and Governance) and for each describe the design approaches observable in analogous systems, the failure modes that can arise when the concerns aren't thoughtfully addressed, and the design considerations that we recommend. Section 4 then describes how each consideration is instantiated in CLIO.

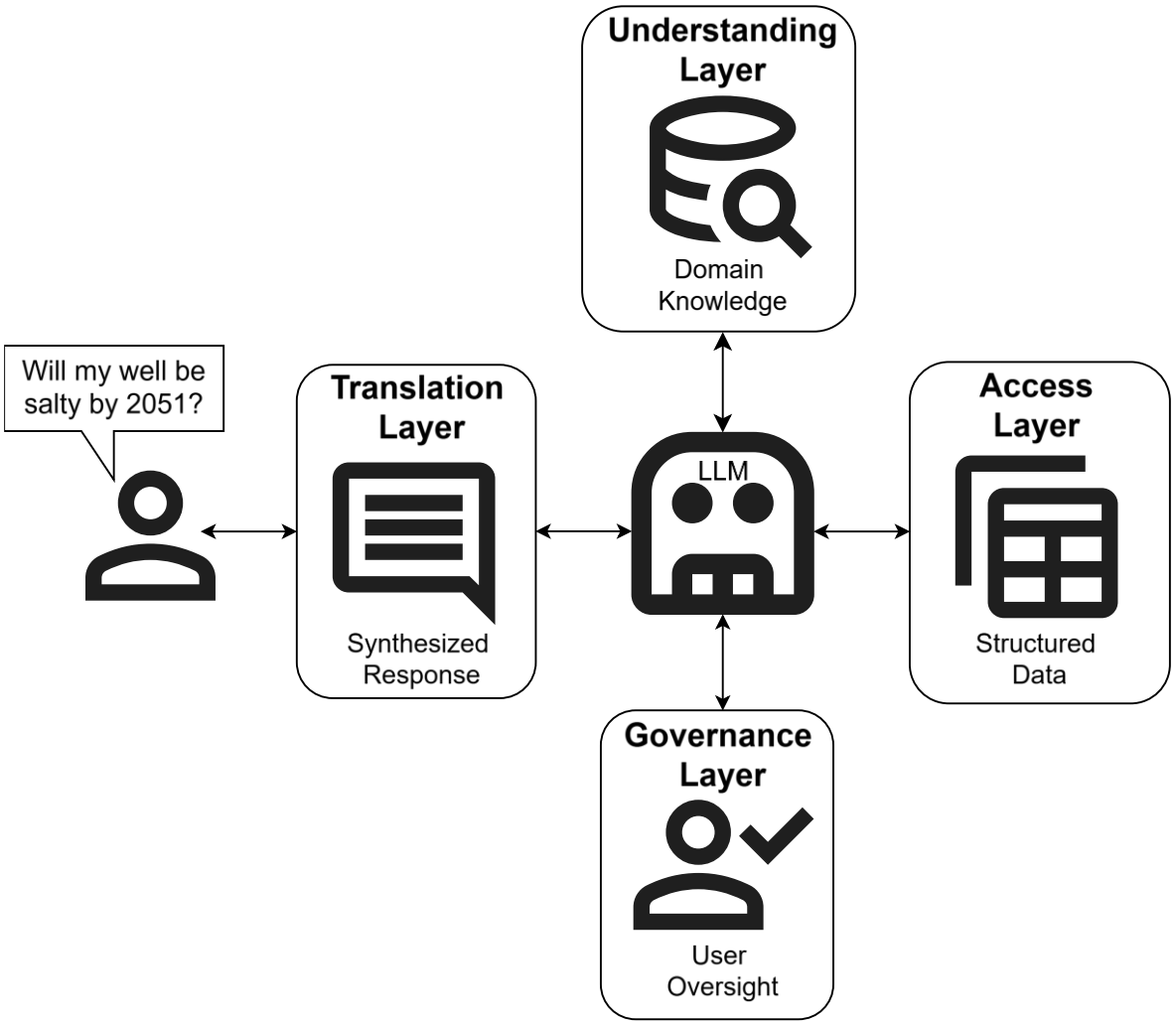

Figure 1. Conceptual Framework

2.1 Access Layer: Data Retrieval

The Access Layer is responsible for connecting the LLM to data that is external to its training distribution. A primary limitation of foundation models is that their knowledge is static and excludes specialized, real-time, or proprietary datasets of the kind that characterize operational environmental science (Xu, Wu, et al., 2025). The implemented systems described in the introduction address this in different ways, each granting the LLM a different degree of autonomy over the retrieval process. At the high-autonomy end, the LLM is given an unrestricted programming environment in which it can generate and execute arbitrary code

against environmental datasets, determining for itself what to extract and how to process it (L. Zhang et al., 2025), introducing a flexibility that, by design, imposes no constraints on what code is executed, leaving correctness, safety, and reproducibility entirely dependent on the model's own judgment. On the other hand, intermediate autonomy approaches share a different design philosophy; the LLM's data access is always mediated by a separate, researcher-defined execution component: whether parsing user intent into structured requests against federal data services (Ramirez & Demir, 2026), making constrained calls to a pre-specified simulation model (Nie & Liu, 2025), or delegating instructions to modular subcomponents across a multi-step workflow (He et al., 2025). At the low-autonomy end, the LLM receives a pre-computed summary of relevant data injected directly into its context, with no role in determining what is retrieved or how.

The degree of LLM autonomy in this layer has direct consequences for reliability. At the high-autonomy end, LLMs exhibit unreliable spatial reasoning in coordinate-based tasks, with accuracy degrading as spatial complexity increases (Yan et al., 2023), a limitation explicitly identified as critical for hydrological applications requiring precise geographic boundary specification, including watershed mapping and flood simulation (Kizilkaya, Sajja, et al., 2025). Even in intermediate approaches, this residual risk persists: the LLM's interpretation of spatial and temporal scope still shapes tool arguments and decides which data is ultimately retrieved, even if the data extraction itself is processed by a tool rather than the LLM itself. Intermediate approaches introduce a further failure mode: geospatial analysis inherently demands large, heterogeneous toolkits spanning data sources, spatial operations, and temporal queries, and as toolkit size grows, tool selection accuracy degrades markedly, a phenomenon Gan & Sun (2025) term prompt bloat, producing hallucinated and mis-parameterized function calls. These errors share a dangerous quality: they are subtle and locally plausible — a slightly misshapen geographic boundary, a salinity value within a credible range but wrong for the selected region — precisely the kind of errors a non-technical audience cannot detect (L. Zhang et al., 2025).

These failure modes motivate a low-autonomy access approach in which the LLM plays a diminished role in data retrieval. In this way the burden of making structured geospatial queries, which is a role LLMs are not particularly suited for, is delegated to a reliable deterministic data engine, letting the LLM act solely in the more natural role of data interpreter rather than data extractor, and ensuring that the data it interprets remains independently verifiable by the user. This is not merely a pragmatic engineering tradeoff; it reflects a principled stance on the primacy of safety in AI-assisted decision support. Rather than treating governance as a safeguard to be added after design decisions are made, a low-autonomy access design takes a proactive architectural approach, eliminating entire categories of failure before they can propagate into the outputs that decision makers rely on. This safety gain, however, is not

free: a low-autonomy design shifts the burden of specifying what data to retrieve — geographic extent, time period, and variable of interest — onto the user rather than the LLM. A well-designed interface can make this burden small (an interactive map and form controls, rather than raw dataset queries), but it remains a real usability cost relative to a system in which a user could simply state their question in natural language and have the LLM determine the relevant data automatically. Nonetheless, given the high-stakes nature of science communication in domains like coastal hydrology, where errors in salinity projections or flood risk estimates can influence infrastructure and water management decisions for communities, we hold that this proactive safety posture is the safer approach.

2.2 Understanding Layer: Domain-Adapted Reasoning

Even when the Access Layer provides the LLM with relevant quantitative data, a general-purpose model is often poorly equipped to interpret it correctly. Environmental science relies on specialized terminology, dataset-specific conventions, and interdisciplinary concepts (the relationship between sea-level rise and groundwater salinity, the distinction between emissions scenarios, the significance of specific hydrological variables) that most models have encountered only superficially during training (Kizilkaya, Sajja, et al., 2025; Kizilkaya, Sermet, et al., 2025; Xu, Li, et al., 2025; Xu, Wu, et al., 2025). The Understanding Layer addresses this by grounding the LLM in authoritative domain knowledge, so that its interpretations are scientifically informed and its claims are traceable to citable sources rather than the model's general training.

The primary technical approach for this is Retrieval-Augmented Generation (RAG), in which a curated library of relevant documents (technical reports, agency publications, scientific literature) is made searchable, and the passages most relevant to a user's query are provided to the LLM alongside it (Martelo et al., 2026; Xu, Wu, et al., 2025). Because each retrieved passage carries the title and page number of its source document, RAG enables the system to tell the user not just what it concluded, but where that conclusion came from. This source attribution is the technical foundation for our Human-on-the-Loop governance pattern: a non-specialist who sees a scientific claim accompanied by a specific document and page number can choose to verify it, question it, or flag it for expert review without needing to understand the underlying model. Without attribution, the user has no way to distinguish a claim grounded in authoritative literature from one the model generated from its general training, and the human oversight that HOTL requires becomes impossible in practice. RAG's primary limitations are retrieval errors — when the most relevant passages are not surfaced, the LLM may fill the gap with unsupported claims (Z. Ji et al., 2023) — and the need for active maintenance, since the knowledge base is only as current as the documents it contains.

Fine-tuning offers a complementary approach: rather than providing domain knowledge at query time, the model is retrained on domain-specific examples so that relevant understanding is built into the model itself (Y. Zhang et al., 2025). Work in hydrology has shown that relatively compact models trained on approximately 8,000 domain-specific examples can achieve meaningful gains in domain reasoning while remaining practical to deploy (Kizilkaya, Sermet, et al., 2025). The tradeoff is the inverse of RAG: fine-tuning improves the quality of the model's reasoning but does not produce citable outputs, and the knowledge it encodes is fixed at training time. Recent work combining both approaches suggests that hybrid systems can outperform either method alone, with fine-tuning helping the model make better use of retrieved passages (Zhou et al., 2026).

The key design consideration for this layer is ensuring that every domain knowledge claim the system makes is accompanied by a traceable source. The knowledge base should be built from documents at appropriate levels of specificity (primary model documentation, model-specific projection studies, regional scientific context, and recognized policy syntheses) and source and page information must be preserved and surfaced to the user alongside every claim. The knowledge base should be maintainable independently of the rest of the system so it can be updated as the science evolves. Fine-tuning on domain-specific material is recommended as a future direction, particularly where data privacy or API dependency make cloud-hosted models impractical.

2.3 Translation Layer: Natural Language Synthesis

The Translation Layer is where the LLM produces its response: synthesizing the verified data summary from the Access Layer and the retrieved domain knowledge from the Understanding Layer into a plain language explanation a non-specialist can act on. Comparable systems implement this layer in two broad modes: explanatory synthesis, where the LLM converts data into a contextually situated narrative interpretation (Koldunov & Jung, 2024; Martelo et al., 2026), and intent-to-action orchestration, where it decomposes user requests into executable data retrieval and analysis workflows (He et al., 2025; Ramirez & Demir, 2026). Koldunov & Jung (2024) describe the first mode as a "local climate service": a system that takes user queries alongside local climate model output and generates explanations connecting regional data to broader scientific framing.

The central failure mode at this layer is that the model's prior training can actively override provided context, producing responses that go beyond or misrepresent the injected data, particularly when a user's question invites speculation. Empirical benchmarking in water and environmental science confirms that this failure is systematic: general-purpose LLMs evaluated on hydrology and water management tasks exhibit error rates exceeding 30% and consistently underperform human domain experts even under controlled conditions (Xu, Li,

et al., 2025; Xu, Wu, et al., 2025). For non-specialist audiences, these errors are compounded by a documented gap between surface fluency and epistemological accuracy in LLM climate responses (Bulian et al., 2023), a risk made more acute by evidence that users tend to trust responses that repeat their own language back to them, even when the underlying information is unreliable (Chiesurin et al., 2023).

The appropriate design response is to ground the LLM's output in the verified inputs it has been provided, rather than allowing it to draw freely on prior training knowledge. This is typically implemented through system prompt instructions that direct the model to treat provided data as authoritative, attribute domain claims to retrieved sources, and decline to speculate beyond what those sources support. Prompting reduces but cannot eliminate this risk: prior training can override explicit instructions, producing responses that misrepresent or go beyond the provided inputs even when the model appears to comply (Longpre et al., 2021; Xie et al., 2023). Because these failures are not always detectable from the response itself, addressing them requires a complementary layer at the output, one that makes the LLM's grounded inputs visible to the user, so they can judge whether the response reflects those inputs or has departed from them.

2.4 Governance Layer: Ensuring Trust and Accountability

The preceding three layers each carry important governance implications. At the Access Layer, the primary governance decision is the choice of workflow autonomy: a low-autonomy, deterministic data pipeline eliminates whole categories of failure at the source rather than relying on downstream mechanisms to catch them. At the Understanding Layer, the governance requirement is provenance preservation: source attribution metadata must flow through the retrieval pipeline so that domain knowledge claims remain traceable. At the Translation Layer, two failure modes persist regardless of how carefully the layers above are designed: prior training can cause the LLM to go beyond its provided inputs, and surface fluency can mask when it does, and these cannot be eliminated by design, only made visible. The Governance Layer as such is interwoven through all the layers and encompasses all of these design considerations, as well as the interaction-level mechanism that makes Translation Layer failures visible to non-specialist users: Human-on-the-Loop (HOTL). According to this pattern the system continues to operate without mandatory approval but provides enough transparency for users to monitor outputs and intervene when needed (Baum & Laux, 2026; Parasuraman et al., 2000).

Making HOTL meaningful requires two things, each of which must be deliberately designed into the layers above. First, there must be a mechanism to provide transparency about the data the LLM is synthesizing. As such the data must be translated into a form that is displayable to the user alongside LLM responses, in such a way that any factual claim can be cross-

referenced without requiring the user to understand the underlying dataset. This is only possible if the Access Layer was designed to produce a verifiable, human-readable output rather than embedding retrieved data opaquely in the model's context. Second, the provenance of any domain knowledge claim must be surfaced in the interface, so the user can identify which source each claim derived from. This is only possible if the Understanding Layer preserved attribution metadata through the retrieval pipeline. The Translation Layer must then be designed to keep these two sources distinct in its outputs, so the user can tell whether a given statement derives from the grounded data or from retrieved domain literature.

In low-stakes applications with expert users, relying on system prompts and user expertise to catch errors may be sufficient. In high-stakes scientific domains serving non-technical audiences, it is not. Errors in the interpretation of salinity projections, flood risk estimates, or groundwater recharge rates can propagate into infrastructure and water management decisions that affect entire communities (Camps-Valls et al., 2025; Karimanzira et al., 2025; Zajac et al., 2025). At scale, repeated inaccuracies in AI-mediated climate communication carry a second risk: LLM outputs that are plausible and confident but factually wrong pose cumulative, long-term risks to science communication and shared social truth (Nabavi et al., 2026; Wachter et al., 2024). In these contexts, a disclaimer urging users to verify outputs is not enough; these systems also need mechanisms that make verification practical for non-specialists. For AI-assisted decision support in high-stakes scientific domains, this governance-centric ordering of design priorities is a safety requirement.

3. Case Study

South Florida faces compounding water management pressures — saltwater intrusion threatening municipal wellfields, rising groundwater levels increasing urban and agricultural flooding, and freshwater flows critical to Everglades restoration, weakened by salinization and reduced recharge (Gebremedhin et al., 2026; Swain et al., 2019). These processes have direct implications for land use and water management policy (Elshall et al., 2020). The stakeholders who must act on this information are largely non-specialists in hydrological modeling, and the scientific projections that describe these risks are not currently accessible to them in a form they can directly use (Dosemagen & Williams, 2022; Elliott, 2022; Hewitt et al., 2017; Mabon, 2020), making this a meaningful opportunity to implement LLMs as part of a science dissemination strategy through an agent like our prototype.

Our case study is based on outputs from the Biscayne and Southern Everglades Coastal Transport (BISECT) model (Swain et al., 2019). BISECT is a density-dependent, integrated groundwater–surface water model developed by the U.S. Geological Survey to represent the hydrologic system of the South Florida peninsula. The model accounts for groundwater flow, surface-water routing, salinity transport, canal–aquifer exchange, coastal boundary forcing,

and wetland hydrologic processes. Its domain covers Miami-Dade County, Monroe County, and part of southern Collier County, capturing both the highly managed urban corridor of the Biscayne Aquifer and the largely undeveloped Everglades wetland system (Figure 2).

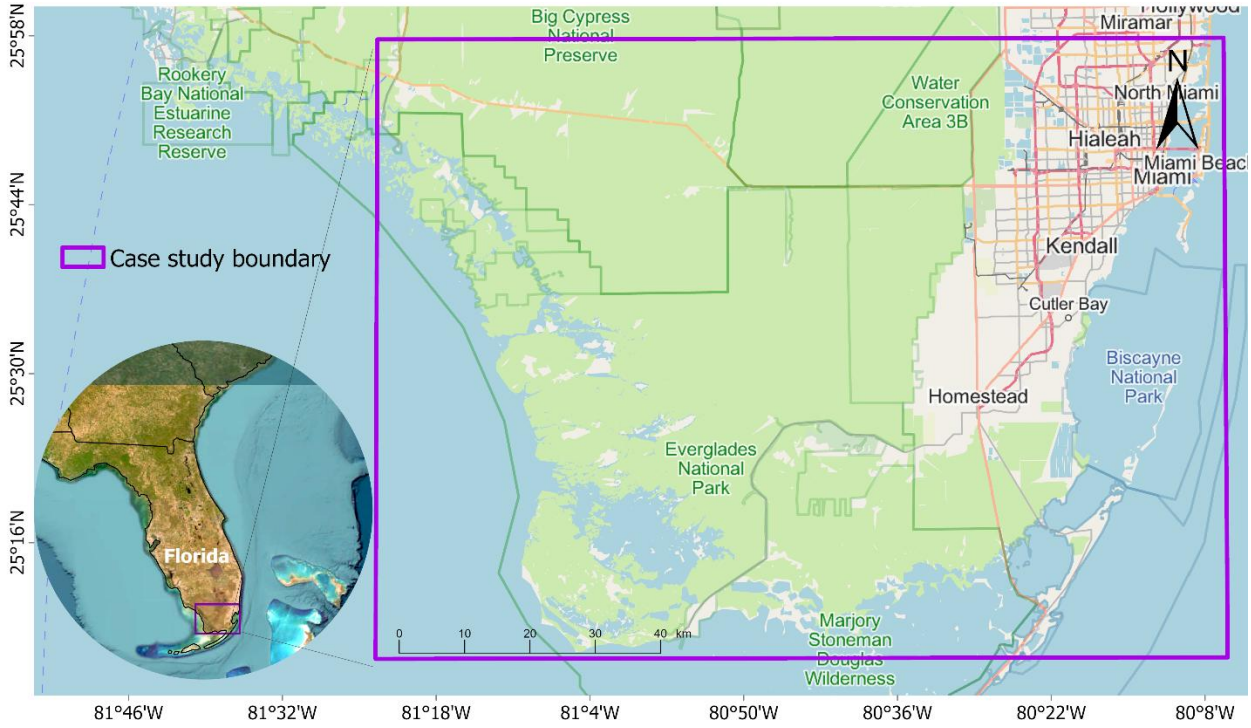

Figure 2. Location and spatial extent of the BISECT model domain in South Florida

The data used in this prototype are sourced from the climate-forced BISECT simulations presented by Gebremedhin et al. (2026). In that study, the calibrated model was validated against independent groundwater-level and surface-water stage observations, then forced with bias-corrected and statistically downscaled CMIP6 rainfall and sea-level rise projections from the NASA IPCC AR6 Sea Level Projection Tool. The simulations cover near-future (2051-2059) and far-future (2091-2099) periods under SSP2-4.5 and SSP5-8.5. The shared variables include groundwater levels, groundwater recharge, groundwater evapotranspiration, groundwater salinity, surface-water salinity, and surface-water temperature. These outputs vary across space and time, but they are not easy to analyze directly in their original format. Even expert users often need additional geospatial processing, statistical summarization, and hydrologic interpretation before the outputs can support decisions. The same study is also embedded directly in CLIO's Understanding Layer knowledge base (Section 4.2), so beyond grounding the case-study data itself, it is one of the sources the agent can retrieve from and cite when a user asks about the projection methodology behind the numbers they are viewing.

The climate forced BISECT simulations generate large, multidimensional spatiotemporal datasets representing hydrological conditions across the case study, over multiple simulation periods, and under different climate scenarios. Figure 3 illustrates surface-water salinity as an example of the spatially and temporally varying outputs generated by the climate-forced BISECT simulations. These outputs vary across both the model domain and the simulation periods, allowing changes in salinity patterns to be examined under different climate scenarios.

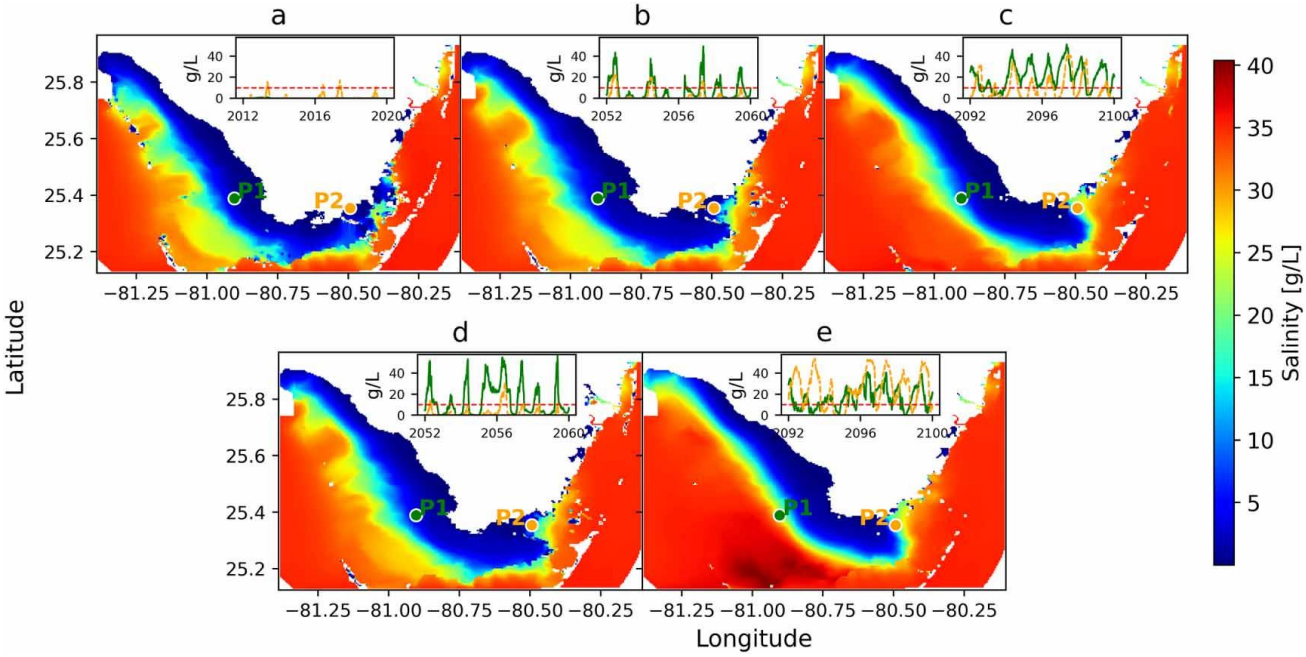

Figure 3. Simulated mean surface-water salinity for the baseline, near-future, and far-future periods under SSP2-4.5 and SSP5-8.5. Panels (a)–(e) correspond to the baseline, near-future SSP2-4.5, far-future SSP2-4.5, near-future SSP5-8.5, and far-future SSP5-8.5 simulations, respectively. The inset time series show simulated salinity at locations P1 and P2.(Gebremedhin et al., 2026)

In this prototype, the dashboard performs the data extraction and visualization, while the LLM translates selected model outputs into plain-language explanations tied to the underlying model-derived statistics. The BISECT dataset is used here as a case study to demonstrate the approach; the dashboard framework is not limited to this specific model or dataset and can be adapted to other environmental model outputs with similar spatial and temporal structure. These data in South Florida provides a particularly relevant setting for demonstrating the governance-first design introduced in Section 2 because of its technically complex model outputs, diverse non-specialist users, and consequential environmental management context increase the potential significance of errors in LLM interpretations.

4. Methods

This section describes the technical implementation of our hydrology agent CLIO as a concrete instantiation of the four-layer design pattern developed in Section 2. CLIO is built around Google's Gemini large language model (Gemini Team et al., 2023), orchestrated through the LangGraph framework (LangChain, Inc., 2023) as a stateful, graph-based agent, and served via a Flask backend to a single-page web frontend built on Leaflet. The following subsections describe each layer in turn.

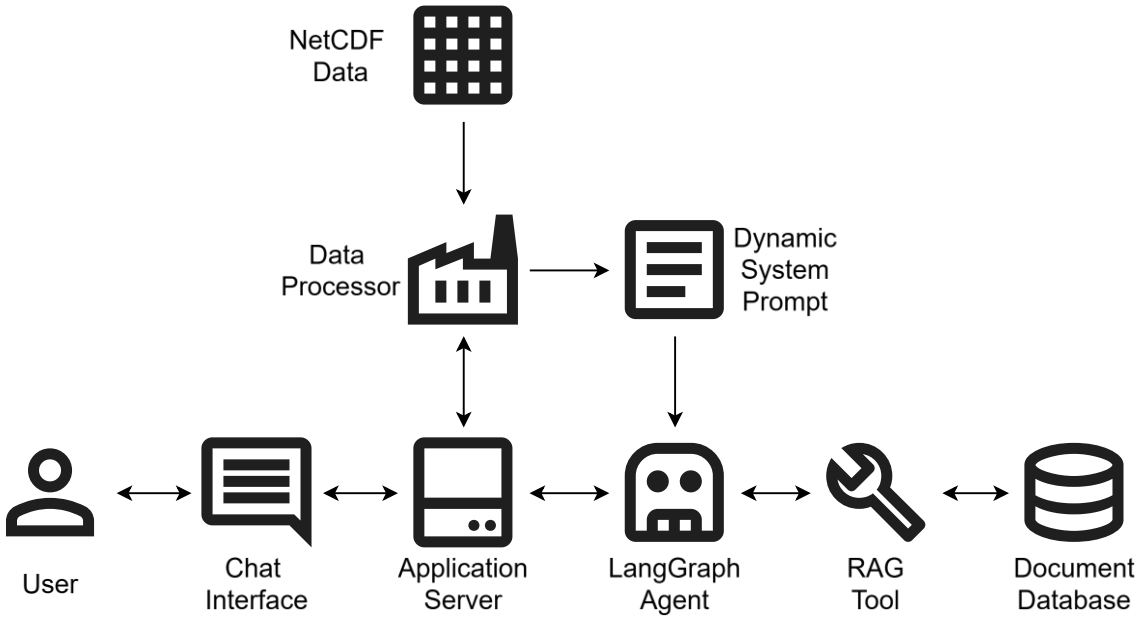

Figure 4. Implemented Architecture of the Case Study Prototype

4.1 Access Layer

The data processing pipeline starts with the source datasets, which exist in the form of CF-compliant NetCDF files (Hassell et al., 2017). It's important to note that the original BISECT projection outputs used in the case study below did not originate in this form: we converted them from their original heterogeneous data binaries into a standard format to make them interoperable with standard Python GIS workflows (Harris et al., 2020; Hoyer & Hamman, 2017; NSF Unidata, n.d.). The user triggers the data extraction through an interactive map on which they draw a rectangular region of interest, specify a time range, and select a variable of interest. The data engine computes a statistical summary (mean, maximum, minimum, standard deviation, and linear trend) for the selected region and period, alongside a

georeferenced heatmap, time series chart, and reverse-geocoded place names for geospatial context via the Nominatim geocoding service (OpenStreetMap contributors, n.d.). These outputs are computed deterministically (no gen-AI) and displayed directly to the user before the LLM is consulted. The statistical summary is then injected into the LLM's system prompt as a fixed, verified reference. As argued in Section 2.1, placing the LLM in control of data retrieval rather than interpretation introduces failure modes that are difficult to detect for a non-specialist audience; the design decision here is to remove the LLM from that role entirely, delegating spatial and statistical operations to a robust deterministic pipeline. This layer is designed to work with any CF-compliant (Climate and Forecast) dataset that provides a rectilinear horizontal grid and a single vertical level per variable, a category that includes the BISECT outputs used here and many single-level climate and hydrology products. From the six BISECT output variables introduced in Section 3, we select the two that fit this criterion for the purposes of our prototype: surface-water salinity and surface-water temperature.

4.2 Understanding Layer

The Understanding Layer is implemented as a RAG tool exposed to the LangGraph agent. When the agent determines that a user's question calls for domain knowledge beyond the injected statistics, it invokes the tool with a natural language query and receives the most relevant passages from the knowledge base, ranked by semantic similarity. Source document citation details (title, author, page number, and DOI) are preserved through the pipeline and returned alongside each passage, enabling the source attribution that the Governance Layer's provenance mechanism depends on. The agent's system prompt also dynamically lists the titles of every document currently in the knowledge base, so it can tell a user when a question falls outside what the system can retrieve rather than defaulting to silence or an ungrounded guess. This list is deliberately limited to titles: the prompt explicitly instructs the agent that recognizing a document's title from its own training is not equivalent to knowing its contents, and that any claim about a listed document must still be grounded in an actual retrieved passage. The knowledge base also needs a way to grow and stay current as the science evolves, so a separate command-line tool lets a user add new documents by confirming each one's citation metadata before it is embedded, and remembers those answers even if the database is later rebuilt — directly supporting the maintainability principle from Section 2.2.

The knowledge base is populated with four documents selected to cover the primary classes of questions a non-technical stakeholder is likely to ask when engaging with South Florida coastal hydrology data. The BISECT technical report (Swain et al., 2019) documents the development, calibration, and physical assumptions of the model that generated the dataset, grounding agent responses about variable definitions, spatial extent, and model behavior in

the authoritative primary source. Gebremedhin et al. (2026) documents the specific climate-forced projection study that produced the case-study dataset itself: the CMIP6 general circulation model selection and bias-correction method, and the SSP2-4.5/SSP5-8.5 near- and far-future simulation design, letting the agent answer methodological questions about the projections a user is actually viewing (which emissions scenarios were modeled, which climate model and downscaling approach were used) with a level of specificity the model documentation alone does not provide. NOAA Technical Report NOS CO-OPS 083 (Sweet et al., 2017) provides authoritative federal sea-level rise projections relevant to the South Florida coastal context, enabling the agent to situate output variables within the regional inundation and emissions assumptions stakeholders are most likely to ask about. The IPCC Sixth Assessment Report Synthesis Summary for Policymakers (Calvin et al., 2023) supplies the broader global climate science consensus and emissions scenario framing (SSP2-4.5, SSP5-8.5) in language already designed for non-technical audiences. Together, these documents represent a practical starting point for domain-adapted RAG knowledge bases in similar applications: primary model documentation, model-specific projection study, relevant regional climate context, and a globally recognized policy-facing synthesis.

4.3 Translation Layer

The Translation Layer is the conversational front end of the system, where the LLM translates the verified data snapshot into plain language responses for the user. At the start of each conversation turn, the model is provided with three pieces of context: a description of its role as a scientific interpreter for non-technical audiences; a summary of the active dataset's key characteristics drawn automatically from its metadata; and the statistical summary of the user's current map selection. The model is explicitly instructed to treat the injected statistical summary as its sole source of numerical claims and to attribute all domain knowledge to retrieved documents rather than general training. Armed with this context, the model can answer questions about the data, explain what the numbers mean, and retrieve supporting scientific literature when the question calls for it. Conversation history is carried forward across turns, so the user can ask follow-up questions naturally without repeating context, providing a recognizable interface for anyone already familiar with mainstream conversational AI systems.

4.4 Governance Layer

The Governance Layer operationalizes the Human-on-the-Loop (HOTL) pattern described in Section 2.4 through two complementary mechanisms that give non-specialist users the tools to independently verify the system's outputs. Prompt-level instructions alone — instructing the LLM to rely only on provided context — are insufficient (see Section 2.3), since they can be overridden by the model's prior training and provide the user with no

independent means of detecting when that occurs. The mechanisms described here address this by making the ground truth visible alongside every response.

The first mechanism is a statistical data panel that displays the exact numerical values computed by the deterministic data processing pipeline directly on screen alongside the LLM's response. Any numerical claim made by the LLM can be cross-referenced against this panel without requiring technical knowledge of the underlying dataset. The second mechanism is a collapsible provenance widget that accompanies every response in the chat interface, indicating whether the response was grounded in retrieved documents or based solely on the injected statistics. When document retrieval occurs, the widget displays the document title and page number of each retrieved chunk, allowing even non-specialist users to verify that citations are traceable to specific source pages in the underlying literature. When no retrieval occurs, the widget notifies the user, making them aware of potential cases where the agent might make claims beyond the actual ground truth knowledge it has available.

5. Prototype Demonstration and Discussion

This section presents the results of our implementation through a qualitative demonstration of the system operating on South Florida hydrological data, and then develops the implications, limitations, and future directions for the broader field of LLM-mediated environmental decision support.

5.1 Agent Demonstration

We demonstrate the system using BISECT projections over South Florida's coastal region under the far-future SSP5-8.5 scenario. The following walkthrough proceeds through each layer of the four-layer architecture in the order a user encounters it.

5.1.1 Setup and Region Selection (Access Layer)

The user is presented with an interactive web map centered on South Florida. Using the map's draw tool, the user selects a rectangular region of interest along with a time range and the surface salinity variable. Upon submitting the selection, the system generates a georeferenced heatmap overlay on the map and populates a time series chart and statistics panel (Figure 5), rendering each model grid cell at its true salinity value, with transparent pixels indicating land or areas outside the model domain.

Figure 5. Interactive Dashboard: Surface Salinity

These values are computed deterministically by the processing pipeline and displayed to the user before the LLM is ever consulted, establishing the verified numerical ground truth that grounds the responses in the exchanges that follow.

5.1.2 Data Interpretation (Translation Layer)

The two examples below demonstrate the Translation Layer's core function: synthesizing the loaded statistics with domain knowledge retrieved from the knowledge base into a single plain-language explanation. In this implementation, the Translation and Understanding Layers overlap within a single response, so what follows should be read as the two layers operating together, with the act of combining data and retrieved knowledge as the focus.

Using the same selection shown in Figure 5, the user asks for an interpretation of the implications of the loaded data. The agent's response (Figure 6) opens by recapping the region, period, trend, and extremes directly from the loaded statistics, then situates that trend in the domain knowledge retrieved from Gebremedhin et al. (2026), synthesizing an accurately cited narrative explaining the sea-level-rise mechanism driving the projected salinity increase and its ecological and water-supply implications. The provenance widget, as discussed earlier, is a central focus of the governance philosophy as it gives the user a mechanism to verify all citations provided by the agent.

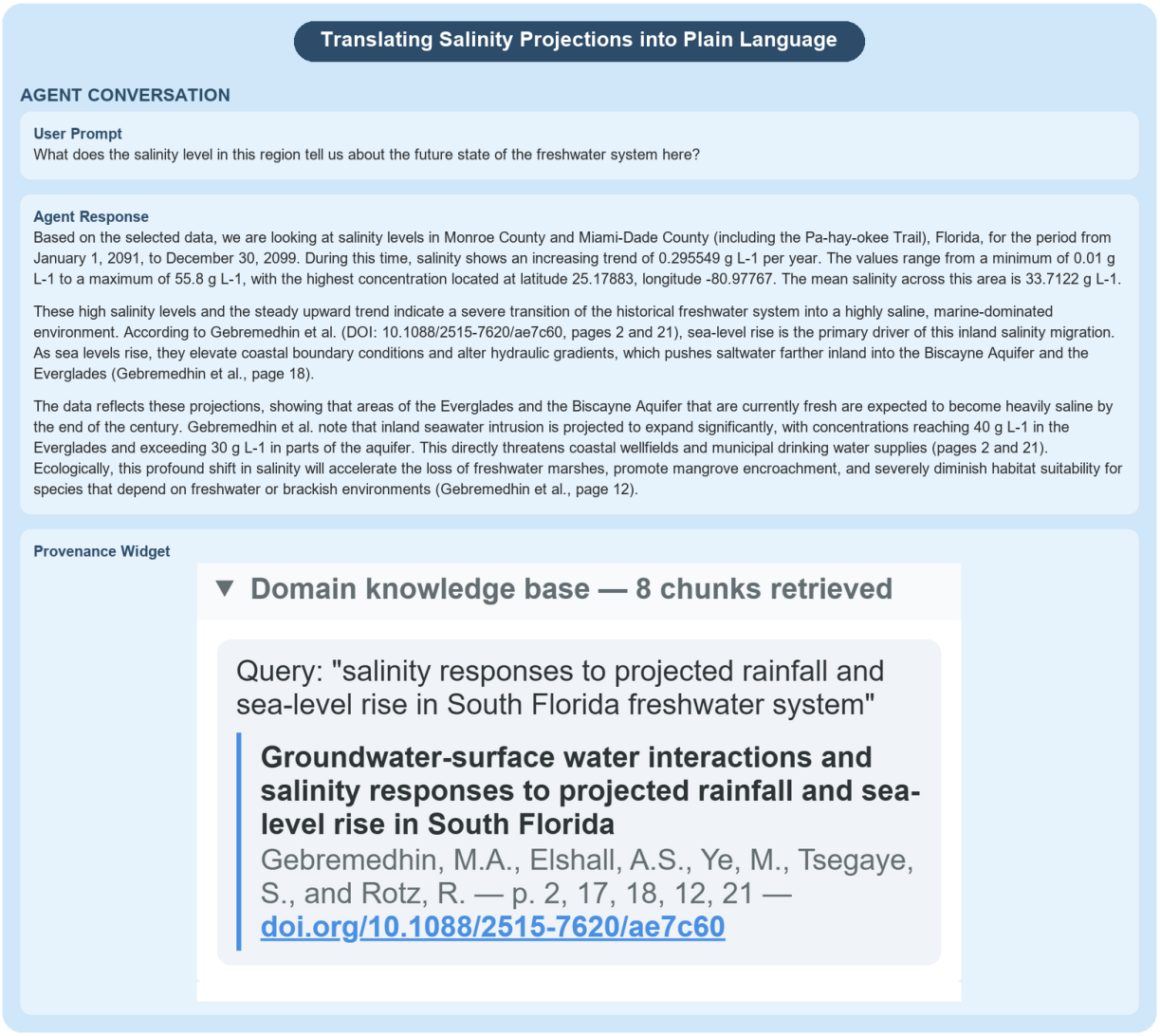

Figure 6. Translating Salinity Projections into Plain Language

For the same region and period, the user asks a question related to the surface-water temperature variable, prompting the agent to explain the physical process behind the observed phenomenon and its implications grounded in the retrieved literature.

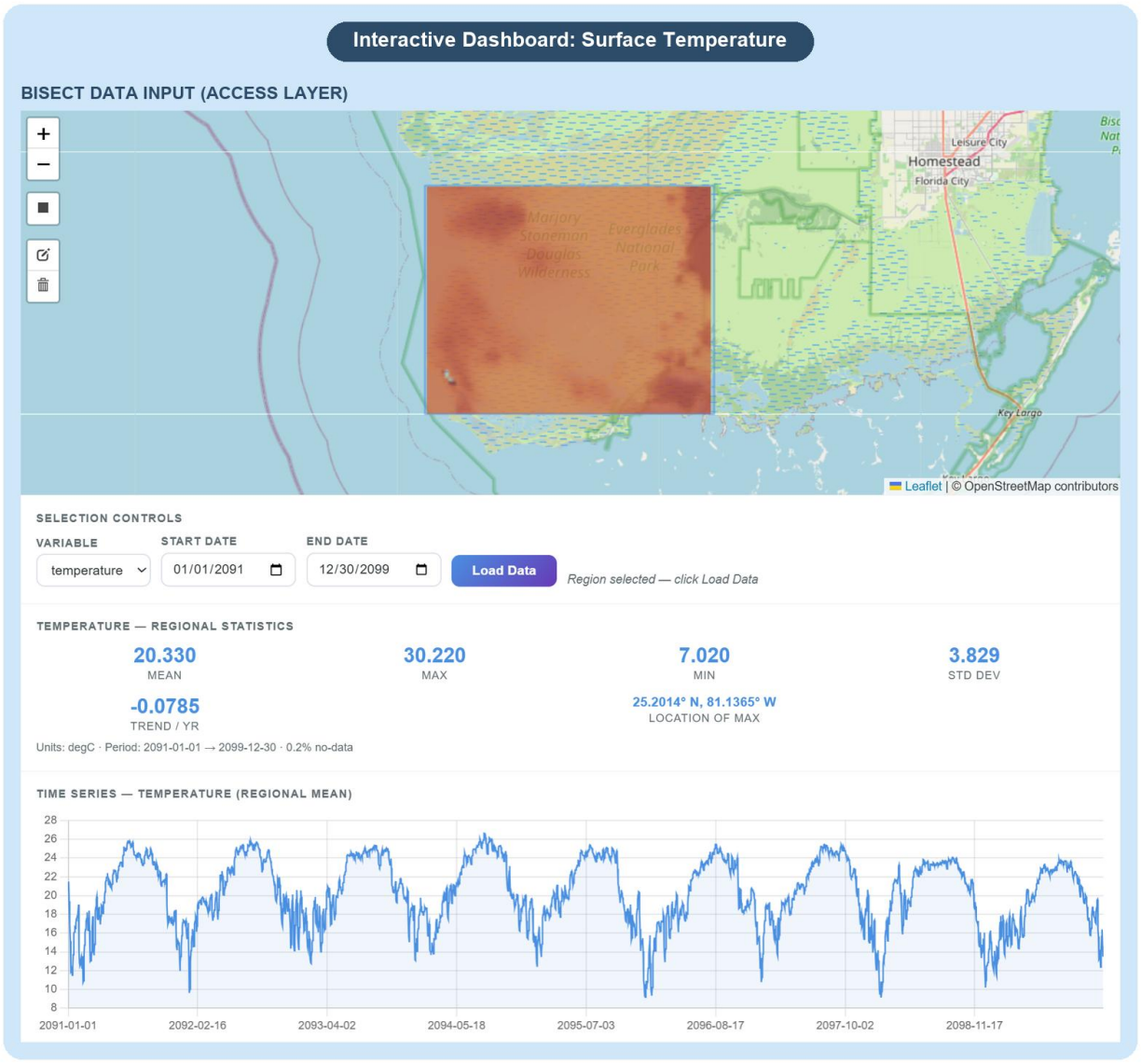

Figure 7. Interactive Dashboard: Surface Temperature

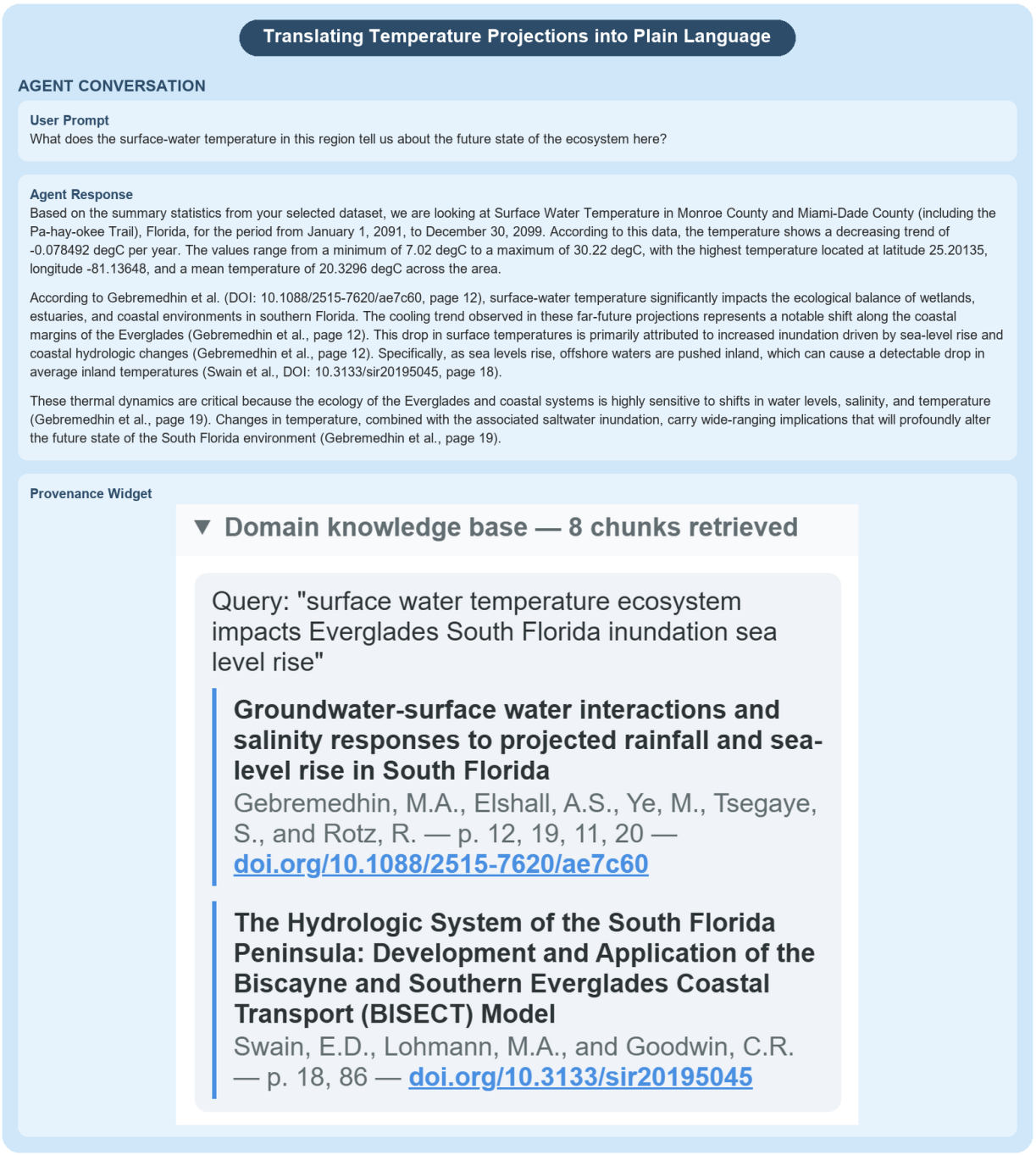

Figure 8. Translating Temperature Projections into Plain Language

5.1.3 Scientific Contextualization (Understanding Layer)

Using the same salinity selection shown in Figure 5, the user asks a question about model uncertainty. The agent's response explains what the model represents, the specific refinements that improve its accuracy, and the model's own documented limitations (sensitivity to tidal and frictional parameters, a likely overestimate of surface-water evaporation), before

closing on the source's own assessment that the model's broader spatial patterns and trends remain reliable despite those limitations.

Grounding Model Uncertainty in the Case-Study Literature

AGENT CONVERSATION

User Prompt
How confident can we be in this projection, and what is it actually based on?

Agent Response
Based on the selected data, we are looking at Salinity in Monroe County and Miami-Dade County (including the Pa-hay-okee Trail), Florida, for the period from January 1, 2091, to December 30, 2099. During this time, salinity shows an increasing trend of 0.295549 g L-1 per year. The values range from a minimum of 0.01 g L-1 to a maximum of 55.8 g L-1, with the highest concentration located at latitude 25.17883, longitude -80.97767. The mean salinity across this area is 33.7122 g L-1.

These projections are based on the Biscayne and Southern Everglades Coastal Transport (BISECT) model, developed by the U.S. Geological Survey (Swain et al., DOI: 10.3133/sir20195045, page 1). The BISECT model uses a numerical simulator to represent hydrodynamic surface-water flow and three-dimensional groundwater conditions. It explicitly accounts for the complex interactions in South Florida, including leakage between surface water and groundwater, water levels, streamflows, and both surface-water and groundwater salinity (Swain et al., page 23). The model was specifically designed to evaluate the effects of sea-level rise and future water-management practices on the region's wetlands and aquifers (Swain et al., pages 1 and 23).

Regarding our confidence in these projections, the BISECT model incorporates several refinements over previous models, including improved topographic representations, refined friction coefficients, and increased vertical aquifer discretization, which improve simulation results (Swain et al., page 1). However, there are inherent uncertainties. For instance, simulated salinity variations in coastal areas are highly sensitive to downstream parameters like tidal fluctuations, while upstream flows are sensitive to frictional resistance (Swain et al., page 10). Additionally, the model's representation of surface-water evaporation may be overestimated, which can introduce some uncertainty into the exact values of the simulations (Swain et al., page 25).

Despite these limitations, the model's overall spatial patterns and trends are considered reliable. Swain et al. conclude that BISECT is a useful and effective hydrologic tool for simulating regional future conditions, and its outputs, including salinity, provide essential input for ecological models and studies of how the South Florida system will respond to climatic changes (Swain et al., page 25).

Provenance Widget

▼ **Domain knowledge base — 16 chunks retrieved**

Query: "Biscayne and Southern Everglades Coastal Transport BISECT model confidence uncertainty temperature evaporation"

> **The Hydrologic System of the South Florida Peninsula: Development and Application of the Biscayne and Southern Everglades Coastal Transport (BISECT) Model**
> Swain, E.D., Lohmann, M.A., and Goodwin, C.R. — p. v, 25, 1, 2, 23, 86 — doi.org/10.3133/sir20195045

Query: "Biscayne and Southern Everglades Coastal Transport BISECT model confidence uncertainty salinity"

> **The Hydrologic System of the South Florida Peninsula: Development and Application of the Biscayne and Southern Everglades Coastal Transport (BISECT) Model**
> Swain, E.D., Lohmann, M.A., and Goodwin, C.R. — p. 2, 1, v, 10, 23 — doi.org/10.3133/sir20195045

Figure 9. Grounding Model Uncertainty in the Case-Study Literature

Using the same selection, the user asks a question about the emissions scenario underlying the projection, framed in terms of its broader global context rather than the loaded data itself. This pulls the agent's retrieval away from the case-study literature: the response (Figure 10) turns to the IPCC AR6 Summary for Policymakers (Calvin et al., 2023) to place the selection's underlying SSP5-8.5 scenario within the full range of Shared Socioeconomic Pathways before closing by connecting that global context back to the severity of the local salinity trend shown in the selection.

Figure 10. Grounding Emissions-Scenario Questions in Global Climate Policy

5.1.4 Governance Boundary (Governance Layer)

Besides the Human-on-the-Loop (HOTL) philosophy present in all agent interactions through the provenance widget and the dashboard, there are also specific situations that merit explicit governance precautions. For example, when a user asks a question outside the scope of the underlying dataset, the LLM is designed to decline any speculation beyond things it can glean directly from the data it has on hand, as seen in Figure 11.

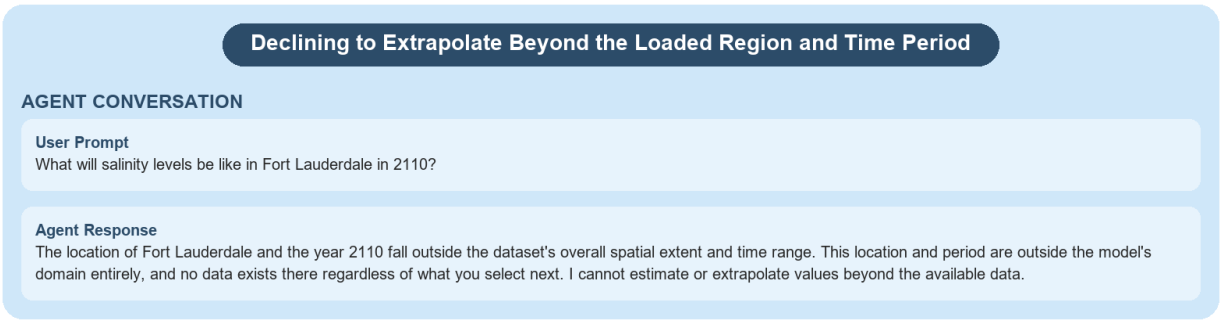

Figure 11. Declining to Extrapolate Beyond the Loaded Region and Time Period

5.1.5 Demonstration Conclusions

Taken together, this demonstration illustrates the system functioning as intended across its core use cases: data-grounded interpretation for non-technical audiences, citable scientific contextualization via RAG, and principled refusal to speculate beyond the verified data snapshot. The side-by-side layout of the data panel and chat interface ensures that users can cross-reference the agent's numerical claims against the source statistics at every point in the conversation, realizing the Human-on-the-Loop governance pattern.

5.2 Limitations and Future Directions

The following limitations of the implementation warrant discussion and point toward concrete directions for future work.

5.2.1 Formal Evaluation

This paper presents a framework and a proof-of-concept implementation; it does not include a quantitative evaluation of the system's output quality. Designing a meaningful benchmark for an LLM-mediated data interface requires domain expertise to construct ground-truth questions and verify citation accuracy, work that falls outside the scope of this contribution. Existing benchmarks evaluate factual accuracy on domain-specific questions in hydrology and water science (Kizilkaya, Sajja, et al., 2025; Xu, Li, et al., 2025; Xu, Wu, et al., 2025), but do not assess the class of properties most critical for this architecture. We note that a correct evaluation for the current architecture should focus on two properties: numerical faithfulness (whether the agent's responses stay within the bounds of the statistics it was given, rather than extrapolating beyond the injected snapshot) and citation accuracy (whether claims attributed to retrieved documents are actually supported by the cited source pages). Developing and applying such a domain-specific benchmark is the most important near-term research priority for establishing the trustworthiness of this class of system before deployment in real decision-support contexts.

5.2.2 Dataset Structural Constraints

The current Access Layer supports CF-compliant datasets with a rectilinear horizontal grid and a single vertical level per variable — the shape of the BISECT salinity outputs used in this case study. Datasets that diverge from this shape are not currently handled: curvilinear horizontal grids, common in ocean models such as several CMIP6 components, and variables carrying an additional vertical dimension, such as pressure-level CMIP6 fields or the multi-layer groundwater head output of the BISECT model itself. Supporting these datasets would instead require treating the vertical level as an explicit, user-specified selection, analogous to the existing variable selection. Doing so, however, would place a further manual burden on the user and would require thoughtful interface design. We leave multi-level variable support, and the design needed to keep that added burden manageable, as a direction for future work.

5.2.3 Multi-Dataset and Multi-Scenario Comparison

The current implementation loads a single NetCDF file at server startup. Many large-scale environmental datasets, including BISECT and CMIP6 outputs, are distributed as collections of files partitioned by time period or ensemble member, and supporting multi-file loading would extend the system's applicability to the full scope of these datasets' domains. More significantly, the system currently presents one scenario at a time. Extending it to load and compare multiple scenarios simultaneously (e.g., SSP2-4.5 versus SSP5-8.5) would directly address the most common class of stakeholder question: how outcomes differ under moderate versus high emissions. This would substantially strengthen the case for LLM-mediated access as a decision support tool.

5.2.4 Uncertainty Quantification

The summary statistics provided to the LLM represent a single model run. Meaningful uncertainty communication for climate decision support requires characterizing the spread across multiple ensemble members or scenarios, as the differences between projections are precisely the information stakeholders need to evaluate risk under alternative futures. Future implementations could accept multiple dataset paths, compute inter-ensemble statistics, and include scenario spread in the injected context block.

5.2.5 Arbitrary Region Selection

The current bounding box selection tool produces rectangular geographic extents. Real planning boundaries (watersheds, municipal boundaries, flood zones) are rarely rectangular. Support for arbitrary polygon selection would make the system more directly applicable to the administrative and hydrologic units that stakeholders already work with, though it would require masked rather than slice-based data extraction operations.

5.2.6 Data Source Coverage and API Integration

Beyond the single-file loading constraint, future versions should support direct integration with open data services such as the Copernicus Climate Data Store ("Climate Data Store," n.d.), enabling the system to query up-to-date observational and reanalysis products alongside model outputs.

5.2.7 Locally Hosted Fine-Tuned Model

Dependence on a commercial API introduces latency, cost, and data privacy concerns. A locally hosted model fine-tuned on hydrology and climate literature would improve performance on domain-specific queries (Kizilkaya, Sermet, et al., 2025; Y. Zhang et al., 2025) and suitability for deployment in sensitive or low-connectivity contexts.

5.3 Opportunities

The four-layer design pattern is not specific to hydrological modeling. Any domain where complex model outputs must reach non-specialist decision-makers presents a natural application context — only the Access Layer's data formats and the Understanding Layer's document library need to be adapted. The following examples illustrate this portability.

- **Public Health and Epidemiology:** The Access Layer could interface with real-time disease surveillance datasets, enabling an agent to synthesize epidemiological literature and provide context-sensitive guidance to public health officials and community stakeholders on emerging outbreak risks (Kaur & Butt, 2025; Kong et al., 2026).
- **Flood and Extreme Weather Management**: For flood forecasting or extreme weather monitoring, the Translation Layer could convert technical model outputs into plain-language summaries for emergency responders and community decision-makers, grounding communications in current observational data rather than generic preparedness messaging (Karimanzira et al., 2025).
- **Weather Forecasting and Agriculture:** The architecture can be adapted to interface with global atmospheric models such as Copernicus seasonal forecast products ("Climate Data Store," n.d.), allowing farmers and agricultural planners to query seasonal projections in plain language. LLM-based agricultural decision support of this kind is already an active area of development (Cantonjos & Biswas, 2025; Kandamali et al., 2025).

5.4 Challenges and Open Questions

Despite the promise demonstrated by our prototype and similar agents in this field, several fundamental challenges must be addressed before systems of this kind can be responsibly deployed at scale. We pose the following open questions to the AI and environmental science research communities.

- **How do non-technical users calibrate trust in AI-generated scientific interpretations?** The Human-on-the-Loop design assumes that users will engage with the provenance display and cross-check the agent's numerical claims against the data panel. Whether this assumption holds in practice is an empirical question that needs to be studied. LLMs in climate communication show a documented gap between surface fluency and epistemological accuracy (Bulian et al., 2023), and users have been shown to prefer and trust fluent-sounding responses that repeat their own words back to them, even when those responses are unfaithful to the underlying knowledge base (Chiesurin et al., 2023). Whether users will engage critically with provenance information when responses sound authoritative remains untested. Empirical research on trust calibration, including interface design interventions that promote appropriate skepticism, is needed before these systems can be deployed responsibly at scale.
- **Who bears accountability when an LLM-mediated system produces a misleading interpretation that influences a policy decision?** The field of AI-assisted environmental decision support lacks established norms for liability attribution. LLMs themselves cannot be held directly accountable for their outputs — responsibility is distributed across users, providers, and deployers without established legal frameworks to allocate it (Wachter et al., 2024) — and the question of whether responsibility lies with the model developer, the deploying agency, or the domain scientist who curated the knowledge base is unresolved.
- **Can LLM-mediated data interfaces serve the communities most affected by climate change?** These systems have the potential to democratize access to climate knowledge, but they also risk concentrating that access further if use depends on reliable internet connectivity, digital literacy, or fluency in high-resource languages. The communities most exposed to climate risk are frequently those with the least access to the infrastructure these systems require. Ensuring that LLM-mediated data interfaces deliver on their democratic promise requires deliberate attention to language support, offline capability, and co-design with target communities.
- **How should environmental decision-support systems manage the instability of foundation models?** The LLMs that power systems like the one described here are updated frequently and without notice by their developers. A deployment that behaves correctly under one model version may behave differently and dangerously after an upstream update, even with identical prompts and architecture. This creates a new class of maintenance burden for environmental agencies and research groups that deploy such systems: continuous re-validation across model versions, and clear protocols for suspending a system when its behavior can no longer be verified. The

field has not yet developed norms or standards for managing this instability in high-stakes scientific communication contexts.

- **Can LLMs reliably serve as data extractors?** The low-autonomy Access Layer design proposed in this paper (Section 2.1) has a corresponding cost: the user, not the system, must specify what data to retrieve — in the current implementation, manually drawing a bounding box, selecting a time range, and choosing a variable through the map interface before any question can be asked (Section 5.1.1). The interactive map and form controls reduce this to a small number of direct manipulations rather than requiring familiarity with NetCDF dimension names or coordinate systems, but it remains a meaningfully different interaction model than posing a question in natural language and trusting the system to determine the relevant data automatically, an issue that becomes more acute as the complexity of the shape of the underlying dataset grows. This usability cost is not merely inconvenience: it presumes users can translate their question into the system's structured vocabulary — a bounding box, a time range, a named variable — a translation step the non-technical audiences this system targets might not have. That cost may also not scale evenly across domains: a handful of form controls are manageable for a handful of variables across one dataset, but could become the dominant friction in a domain with dozens of datasets or variables to choose from. An alternative, increasingly explored in comparable systems, is allowing the LLM to drive data retrieval directly, translating natural language queries into structured geospatial operations, which would remove this manual burden and potentially make these systems even more accessible to all audiences. The Access Layer failure modes discussed in Section 2.1 make this difficult to achieve reliably in practice, but whether advances in model architecture, structured tool-use frameworks, or hybrid human-LLM workflows can eventually bring LLM-driven geospatial extraction to a standard of reliability appropriate for scientific decision support — closing the gap between reliability and the usability cost named above — remains an open question well worth investigating.

6. Conclusion

The development of LLM-mediated interfaces for environmental data represents a genuine opportunity to close the gap between open scientific datasets and the diverse publics who need to act on them. The FAIR data movement has substantially improved data findability, interoperability, and accessibility (Wilkinson et al., 2016); the remaining bottleneck is usability — the ability of non-specialist users to extract and understand decision-relevant information from technically demanding archives (Elliott, 2022). The framework and implementation presented in this paper address this bottleneck directly, using South Florida coastal hydrology as a case domain where the stakes (salinity intrusion, groundwater flooding,

freshwater sustainability under sea-level rise) are both technically complex and immediately consequential for community decision-making.

The four-layer design pattern described in this paper provides a modular blueprint that is not specific to the South Florida domain or to hydrological data. Any field in which a data-to-model-to-policy pipeline exists can instantiate this architecture by adapting the domain-specific data processing and knowledge base components while preserving the LLM's interpretive role. The central argument of this paper is that doing so responsibly requires treating governance not as a downstream addition but as the organizing constraint from which the remaining design decisions follow: what data the LLM is given access to, what knowledge it can draw on, and how its outputs are presented to users are all, in the end, governance decisions.

This contribution is best understood as complementary to, rather than competing with, the growing body of implemented LLM-mediated environmental systems it draws on. Voice-enabled retrieval platforms, GPT-4-powered flood-risk advisories, orchestrated water-resources workflows, and local climate services each demonstrate that LLM-mediated interpretation is feasible for a specific environmental use case; what they do not yet offer is a shared vocabulary for comparing the governance implications of their underlying design choices. The four-layer pattern developed here is intended to fill that gap as a lens through which their design choices, and the tradeoffs those choices carry, can be made explicit, compared, and evaluated by the water-resource-management and environmental-decision-support communities these systems are meant to serve.

The rapid advancement of foundation models increases pressure to automate more of the science communication pipeline from data extraction to interpretation to policy recommendation. While this trajectory offers real efficiency gains, it also raises the risk of progressively displacing the human judgment and domain expertise that give scientific communication its credibility and accountability. Our experience developing this system suggests that the more productive vision is not one in which LLMs replace human experts, but one in which they extend the reach of human expertise, making complex scientific data accessible to the stakeholders who need it, while keeping domain scientists, water managers, and community decision-makers in the roles of interpreter, validator, and final arbiter. The governance mechanisms described in this paper are designed with that vision in mind: to make the boundary between machine assistance and human judgment explicit, legible, and meaningful to the people the system is built to serve.

Acknowledgments

Generative AI Disclosure: Claude (Anthropic) was used in the development of this work, including software development, manuscript drafting and revision, and figure design and review. All AI-assisted content was reviewed and verified by the authors prior to inclusion. Claude was not used to generate, analyze, or interpret the underlying scientific data or results presented in this study.

(Placeholder for additional acknowledgments content)

Bibliography

Baum, K., & Laux, J. (2026, March 19). Constitutive vs. Corrective: A Causal Taxonomy of Human Runtime Involvement in AI Systems. arXiv. https://doi.org/10.48550/arXiv.2603.19213

Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, *624*(7992), 570–578. https://doi.org/10.1038/s41586-023-06792-0

Bulian, J., Schäfer, M. S., Amini, A., Lam, H., Ciaramita, M., Gaiarin, B., et al. (2023). Assessing Large Language Models on Climate Information (Version 2). https://doi.org/10.48550/ARXIV.2310.02932

Cabana, D., Rölfer, L., Evadzi, P., & Celliers, L. (2023). Enabling Climate Change Adaptation in Coastal Systems: A Systematic Literature Review. *Earth's Future*, *11*(8), e2023EF003713. https://doi.org/10.1029/2023EF003713

Calvin, K., Dasgupta, D., Krinner, G., Mukherji, A., Thorne, P. W., Trisos, C., et al. (2023). *IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.* (First). (P. Arias, M. Bustamante, I. Elgizouli, G. Flato, M. Howden, C.

Méndez-Vallejo, et al., Eds.). Intergovernmental Panel on Climate Change (IPCC). https://doi.org/10.59327/IPCC/AR6-9789291691647

Cammarano, D., Olesen, J. E., Helming, K., Foyer, C. H., Schönhart, M., Brunori, G., et al. (2023). Models can enhance science–policy–society alignments for climate change mitigation. *Nature Food*, *4*(8), 632–635. https://doi.org/10.1038/s43016-023-00807-9

Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K.-H., Höhl, A., Castelletti, A., Pacal, A., et al. (2025). Artificial intelligence for modeling and understanding extreme weather and climate events. *Nature Communications*, *16*(1), 1919. https://doi.org/10.1038/s41467-025-56573-8

Cantonjos, N. A., & Biswas, A. (2025). AgroAskAI: A Multi-Agentic AI Framework for Supporting Smallholder Farmers' Enquiries Globally (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2512.14910

Chiesurin, S., Dimakopoulos, D., Sobrevilla Cabezudo, M. A., Eshghi, A., Papaioannou, I., Rieser, V., & Konstas, I. (2023). The Dangers of trusting Stochastic Parrots: Faithfulness and Trust in Open-domain Conversational Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 947–959). Toronto, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.60

Climate Data Store. (n.d.). Retrieved May 22, 2026, from https://cds.climate.copernicus.eu/

Condon, L. E., Kollet, S., Bierkens, M. F. P., Fogg, G. E., Maxwell, R. M., Hill, M. C., et al. (2021). Global Groundwater Modeling and Monitoring: Opportunities and

Challenges. *Water Resources Research*, *57*(12), e2020WR029500. https://doi.org/10.1029/2020WR029500

Dosemagen, S., & Williams, E. (2022). Data Usability: The Forgotten Segment of Environmental Data Workflows. *Frontiers in Climate*, *4*, 785269. https://doi.org/10.3389/fclim.2022.785269

Elliott, K. C. (2022). Open Science for Non-Specialists: Making Open Science Meaningful Beyond the Scientific Community. *Philosophy of Science*, *89*(5), 1013–1023. https://doi.org/10.1017/psa.2022.36

Elshall, A. S., Arik, A. D., El-Kadi, A. I., Pierce, S., Ye, M., Burnett, K. M., et al. (2020). Groundwater sustainability: a review of the interactions between science and policy. *Environmental Research Letters*, *15*(9), 093004. https://doi.org/10.1088/1748-9326/ab8e8c

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Gan, T., & Sun, Q. (2025, May 6). RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation. arXiv. https://doi.org/10.48550/arXiv.2505.03275

Gebremedhin, M. A., Elshall, A. S., Ye, M., Tsegaye, S., & Rotz, R. (2026). Groundwater-surface water interactions and salinity responses to projected rainfall and sea-level rise

in South Florida. *Environmental Research Communications*, *8*(6), 065050. https://doi.org/10.1088/2515-7620/ae7c60

Gemini Team, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., et al. (2023). Gemini: A Family of Highly Capable Multimodal Models (Version 5). arXiv. https://doi.org/10.48550/ARXIV.2312.11805

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hassell, D., Gregory, J., Blower, J., Lawrence, B. N., & Taylor, K. E. (2017). A data model of the Climate and Forecast metadata conventions (CF-1.6) with a software implementation (cf-python v2.1). *Geoscientific Model Development*, *10*(12), 4619–4646. https://doi.org/10.5194/gmd-10-4619-2017

He, G., Luo, J., Luo, F., Yang, X., Jing, X., & Cui, H. (2025). IWMS-LLM: an intelligent water resources management system based on large language models. *Journal of Hydroinformatics*, *27*(11), 1685–1702. https://doi.org/10.2166/hydro.2025.130

Hewitt, C. D., Stone, R. C., & Tait, A. B. (2017). Improving the use of climate information in decision-making. *Nature Climate Change*, *7*(9), 614–616. https://doi.org/10.1038/nclimate3378

Hoyer, S., & Hamman, J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*(1), 10. https://doi.org/10.5334/jors.148

Huggins, X., Gleeson, T., Famiglietti, J. S., Reinecke, R., Zamrsky, D., Wagener, T., et al. (2025). A review of open data for studying global groundwater in social–ecological

systems. *Environmental Research Letters*, *20*(9), 093002. https://doi.org/10.1088/1748-9326/adf127

Iturbide, M., Fernández, J., Gutiérrez, J. M., Pirani, A., Huard, D., Al Khourdajie, A., et al. (2022). Implementation of FAIR principles in the IPCC: the WGI AR6 Atlas repository. *Scientific Data*, *9*(1), 629. https://doi.org/10.1038/s41597-022-01739-y

Ji, X., Wu, X., Deng, R., Yang, Y., Wang, A., & Zhu, Y. (2025). Utilizing large language models for identifying future research opportunities in environmental science. *Journal of Environmental Management*, *373*, 123667. https://doi.org/10.1016/j.jenvman.2024.123667

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kandamali, D. F., Porter, W. M., Porter, E., McLemore, A., & Rains, G. C. (2025). CottonBot: An AI-driven cotton farming assistant and irrigation advisor using LLM-RAG and agentic AI tools. *Smart Agricultural Technology*, *12*, 101640. https://doi.org/10.1016/j.atech.2025.101640

Karimanzira, D., Rauschenbach, T., Hellmund, T., & Ritzau, L. (2025). Improved Flood Management and Risk Communication Through Large Language Models. *Algorithms*, *18*(11), 713. https://doi.org/10.3390/a18110713

Kaur, J., & Butt, Z. A. (2025). AI-driven epidemic intelligence: the future of outbreak detection and response. *Frontiers in Artificial Intelligence*, *8*, 1645467. https://doi.org/10.3389/frai.2025.1645467

Kizilkaya, D., Sajja, R., Sermet, Y., & Demir, I. (2025). Toward HydroLLM: a benchmark dataset for hydrology-specific knowledge assessment for large language models. *Environmental Data Science*, *4*, e31. https://doi.org/10.1017/eds.2025.10006

Kizilkaya, D., Sermet, Y., & Demir, I. (2025). Towards HydroLLM: approaches for building a domain-specific language model for hydrology. *Journal of Hydroinformatics*, *27*(10), 1652–1666. https://doi.org/10.2166/hydro.2025.100

Koldasbayeva, D., Tregubova, P., Gasanov, M., Zaytsev, A., Petrovskaia, A., & Burnaev, E. (2024). Challenges in data-driven geospatial modeling for environmental research and practice. *Nature Communications*, *15*(1), 10700. https://doi.org/10.1038/s41467-024-55240-8

Koldunov, N., & Jung, T. (2024). Local climate services for all, courtesy of large language models. *Communications Earth & Environment*, *5*(1), 13. https://doi.org/10.1038/s43247-023-01199-1

Kong, J. D., Gillies, M., Gardner, E., & Bragazzi, N. L. (2026). Multi-model large-scale AI framework for avian influenza surveillance and preparedness: Harnessing large language models to enhance risk communication, real-time decision support, and public health response strategies. *One Health*, *22*, 101357. https://doi.org/10.1016/j.onehlt.2026.101357

LangChain, Inc. (2023, August 9). langchain-ai/langgraph: Build resilient agents. Retrieved June 26, 2026, from https://github.com/langchain-ai/langgraph

Longpre, S., Perisetla, K., Chen, A., Ramesh, N., DuBois, C., & Singh, S. (2021). Entity-Based Knowledge Conflicts in Question Answering. In *Proceedings of the 2021*

Conference on Empirical Methods in Natural Language Processing (pp. 7052–7063). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.565

Mabon, L. (2020). Making climate information services accessible to communities: What can we learn from environmental risk communication research? *Urban Climate*, *31*, 100537. https://doi.org/10.1016/j.uclim.2019.100537

Martelo, R., Ahmadiyehyazdi, K., & Wang, R.-Q. (2026). Towards democratized flood risk management: An advanced AI assistant enabled by GPT-4 for enhanced interpretability and public engagement. *Environmental Modelling & Software*, *197*, 106821. https://doi.org/10.1016/j.envsoft.2025.106821

Moeck, C., Grech-Cumbo, N., Podgorski, J., Bretzler, A., Gurdak, J. J., Berg, M., & Schirmer, M. (2020). A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships. *Science of The Total Environment*, *717*, 137042. https://doi.org/10.1016/j.scitotenv.2020.137042

Nabavi, E., Maier, H. R., Razavi, S., Hindes, A., Howden, M., Grant, W., & Raman, S. (2026). On using large language models to support social research for climate action. *Npj Climate Action*, *5*(1), 52. https://doi.org/10.1038/s44168-026-00375-1

Nie, Q., & Liu, T. (2025). Large language models: Tools for new environmental decision-making. *Journal of Environmental Management*, *375*, 124373. https://doi.org/10.1016/j.jenvman.2025.124373

NSF Unidata. (n.d.). NetCDF | NSF Unidata. Retrieved June 1, 2026, from https://www.unidata.ucar.edu/software/netcdf

OpenStreetMap contributors. (n.d.). OpenStreetMap. Retrieved July 23, 2026, from https://www.openstreetmap.org/

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

Ramirez, C. E., & Demir, I. (2026). AI-assisted voice enabled computing framework for hydrological analysis. *Environmental Modelling & Software*, *197*, 106833. https://doi.org/10.1016/j.envsoft.2025.106833

Reichert, P., Langhans, S. D., Lienert, J., & Schuwirth, N. (2015). The conceptual foundation of environmental decision support. *Journal of Environmental Management*, *154*, 316–332. https://doi.org/10.1016/j.jenvman.2015.01.053

Romera-Paredes, B., Barekatain, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., et al. (2024). Mathematical discoveries from program search with large language models. *Nature*, *625*(7995), 468–475. https://doi.org/10.1038/s41586-023-06924-6

Swain, E. D., Lohmann, M. A., & Goodwin, C. R. (2019). *The Hydrologic System of the South Florida Peninsula: Development and Application of the Biscayne and Southern Everglades Coastal Transport (BISECT) Model* (USGS Greater Everglades Priority Ecosystem Studies Initiative No. Scientific Investigations Report 2019-5045). Reston, Virginia: U.S. Geological Survey.

Sweet, W. V., Kopp, R., Weaver, C. P., Obeysekera, J., Horton, R. M., Thieler, E. R., & Zervas, C. E. (2017). Global and regional sea level rise scenarios for the United States. https://doi.org/10.7289/V5/TR-NOS-COOPS-083

Wachter, S., Mittelstadt, B., & Russell, C. (2024). Do large language models have a legal duty to tell the truth? *Royal Society Open Science*, *11*(8), 240197. https://doi.org/10.1098/rsos.240197

Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., et al. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, *620*(7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Xie, J., Zhang, K., Chen, J., Lou, R., & Su, Y. (2023). Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2305.13300

Xu, B., Li, Z., Yang, Y., Wu, G., Wang, C., Tang, X., et al. (2025). Evaluating and Advancing Large Language Models for Water Knowledge Tasks in Engineering and Research. *Environmental Science & Technology Letters*, *12*(3), 289–296. https://doi.org/10.1021/acs.estlett.5c00038

Xu, B., Wu, G., Li, Z., Xu, G., Zeng, H., Tong, R., & Ng, H. Y. (2025). Towards domain-adapted large language models for water and wastewater management: methods, datasets

and benchmarking. *Npj Clean Water*, *8*(1), 82. https://doi.org/10.1038/s41545-025-00509-8

Yan, H., Hu, X., Wan, X., Huang, C., Zou, K., & Xu, S. (2023). Inherent limitations of LLMs regarding spatial information (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2312.03042

Zajac, M., Kulawiak, C., Li, S., Erickson, C., Hubbell, N., & Gong, J. (2025). Unifying Flood-Risk Communication: Empowering Community Leaders Through AI-Enhanced, Contextualized Storytelling. *Hydrology*, *12*(8), 204. https://doi.org/10.3390/hydrology12080204

Zhang, L., Song, Y., Cui, H., Lu, M., Li, C., Yuan, B., et al. (2025). Foundation Models as Assistive Tools in Hydrometeorology: Opportunities, Challenges, and Perspectives. *Water Resources Research*, *61*(4), e2024WR039553. https://doi.org/10.1029/2024WR039553

Zhang, Y., Lin, S., Xiong, Y., Li, N., Zhong, L., Ding, L., & Hu, Q. (2025). Fine-tuning large language models for interdisciplinary environmental challenges. *Environmental Science and Ecotechnology*, *27*, 100608. https://doi.org/10.1016/j.ese.2025.100608

Zhou, Z., Jiang, T., & Jin, Q. (2026). Integrating fine-tuning and retrieval-augmented generation to address the application challenges of pollution control guidelines. *Journal of Environmental Management*, *400*, 128718. https://doi.org/10.1016/j.jenvman.2026.128718