arXiv:2508.02936v1 [cs.AI] 4 Aug 2025

AQUAH: Automatic Quantification and Unified Agent in Hydrology

Songkun Yan[1] Zhi Li[2] Siyu Zhu[1] Yixin Wen[3]
Mofan Zhang[4] Mengye Chen[1] Jie Cao[5] Yang Hong[1,*]

[1]School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, USA
[2]Department of Earth System Science, Stanford University, Stanford, CA, USA
[3]Department of Geography, University of Florida, Gainesville, FL, USA
[4]Civil and Environmental Engineering, Stanford University, Stanford, CA, USA
[5]School of Computer Science, University of Oklahoma, Norman, OK, USA

skyan@ou.edu li1995@stanford.edu Siyu.Zhu-1@ou.edu yixin.wen@ufl.edu
mofanz@stanford.edu mchen15@ou.edu jie.cao@ou.edu yanghong@ou.edu

Abstract

*We introduce **AQUAH**, the first end-to-end language-based agent designed specifically for hydrologic modelling. Starting from a simple natural-language prompt (e.g., "simulate floods for the Little Bighorn basin from 2020 to 2022"), AQUAH autonomously retrieves the required terrain, forcing, and gauge data; configures a hydrologic model; runs the simulation; and generates a self-contained PDF report. The workflow is driven by vision-enabled large-language models, which interpret maps and rasters on the fly and steer key decisions such as outlet selection, parameter initialisation, and uncertainty commentary. Initial experiments across a range of U.S. basins show that AQUAH can complete cold-start simulations and produce analyst-ready documentation without manual intervention—results that hydrologists judge as clear, transparent, and physically plausible. While further calibration and validation are still needed for operational deployment, these early outcomes highlight the promise of LLM-centred, vision-grounded agents to streamline complex environmental modelling and lower the barrier between Earth-observation data, physics-based tools, and decision makers.*

1. Introduction

Motivation. Hydrologic simulation and Earth observation analysis are indispensable for managing water resources in a changing climate[8, 14, 15]. Yet fragmented workflows, steep technical requirements, and lengthy model-setup times continue to restrict these capabilities—especially for non-experts and rapid-response applications[3, 5]. A genuinely transformative solution is therefore needed to dismantle these barriers and make advanced, equitable modeling tools broadly accessible.

Problem. Current hydrologic tools are not designed for accessibility or automation; setup and data processing are time-consuming [17]. Users must often manually download data, configure models, and interpret outputs, requiring both domain knowledge and programming skills. Additionally, interpreting the results generated by the model is a barrier and requires years of related domain experience. Although this is not the sole problem for hydrologic science, we, in this paper, propose to bridge the gap and enhance communication of hydrologic simulation.

Solution. We present **AQUAH**—*Automatic Quantification & Unified Agent in Hydrology*—a next-generation, vision-enhanced large-language-model (LLM) agent that converts free-form user prompts into end-to-end hydrologic simulations and narrative reports. Riding on the rapid advances in vision multimodal LLMs (VLMs), AQUAH leverages state-of-the-art vision capabilities to *interpret maps, rasters, and other geospatial imagery on the fly*, replacing several expert-driven decisions—such as outlet selection, parameter initialization—with reliable, data-driven automation that has already shown promising accuracy and consistency in our experiments.

Our agent AQUAH stitches together geospatial data retrieval, Earth-observation forcing data, hydrologic models (e.g., Coupled Routing and Excess Storage, CREST [24]), and automated visualization in a seamless workflow. Thanks to its *data-agnostic, model-agnostic, plug-and-play* design, AQUAH lowers the entry barrier for users without technical modeling backgrounds while still satisfying domain experts. By demonstrating how vision-enabled LLMs can assume formerly human-exclusive roles,

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/343019

AQUAH points the way toward fully autonomous hydrologic modeling agents.

2. Related Work

Multimodal and Tool-Augmented LLMs for Scientific Reasoning. Recent advances in LLMs have demonstrated remarkable capabilities in scientific reasoning when paired with external tools and multimodal inputs [12, 16, 20]. Frameworks such as ReAct, Toolformer, and HuggingGPT combine language understanding with programmatic control, enabling agents to interface with APIs, code environments, and databases [21, 26]. Emerging multimodal foundation models (e.g., GPT-4o [18], Gemini [1], Kosmos-2 [19]) have shown promise in parsing text, images, and structured data for scientific workflows [20, 23]. However, their application to Earth system science remains limited, especially for domain-specific physical modeling like hydrology. AQUAH builds on this foundation by integrating natural language processing with geospatial data tools, Earth observation inputs, and model execution capabilities.

Earth Observation + AI for Sustainability. The fusion of EO data and AI has advanced rapidly in applications like land cover classification, crop monitoring, and disaster mapping [4, 11, 25]. Vision-based foundation models (e.g., Segment Anything Model, SatMAE, Prithvi) have pushed the frontier in remote sensing understanding [22]. Yet, most works focus on static scene understanding rather than simulation-driven analysis. In hydrology, EO data like CHIRPS rainfall or MODIS evapotranspiration are used in modeling pipelines, but are rarely integrated via intelligent agents or prompted via natural language.

Automation in Hydrologic Modeling. Traditional hydrologic models (e.g., CREST, EF5, SWAT, HEC-HMS) are well-established for flood simulation and watershed analysis [9, 13, 24]. However, they require significant manual effort for setup, data integration, parameter calibration, and output interpretation. Recent efforts in workflow automation (e.g., RavenPy, RavenWPS) have improved usability [2], but these tools are not conversational, nor are they driven by natural language or LLMs. Our work fills this gap by combining the rigor of physics-based models with the accessibility of LLM agents, forming a bridge between EO, simulation, and narrative reporting.

3. AQUAH

3.1. System Architecture Overview

We design **AQUAH** as a modular language-agent framework that bridges natural language interaction with Earth observation data, geospatial processing, and hydrologic simulation tools. The architecture (Figure 1) consists of:

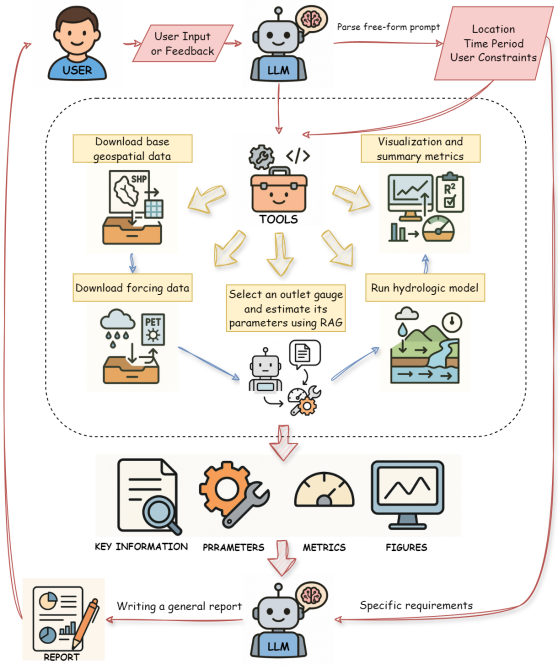


Figure 1. Overview of the AQUAH architecture showing key components and data flow.

- **LLM Interface:** Converts user-provided natural language inputs into structured simulation instructions specifying locations, time periods, and analytical goals.
- **Tool Executor Layer:** Manages and executes Python-based geospatial libraries, hydrologic model wrappers, visualization routines, and statistical summarization tools, orchestrating comprehensive data retrieval, simulation, and analysis workflows.
- **Dynamic Data Pipeline:** Automatically fetches essential hydrological data such as digital elevation models (DEM), precipitation, potential evapotranspiration (PET), and observed discharge datasets, based on user input.
- **Hydrologic Model Integration:** Implements the CREST model for hydrological simulations, utilizing dynamically obtained datasets and providing initial parameter estimates informed by Retrieval-Augmented Generation (RAG) and LLM reasoning.
- **Report Generation Engine:** Automatically compiles simulation outcomes, visualizations, and analytical summaries into structured, publication-quality Markdown or PDF reports.
- **Interactive Feedback Loop:** Allows users to refine simulations via natural-language feedback—e.g., selecting alternative gauges or adjusting parameters. The LLM parses these requests, updates model configurations

through the Tool Executor Layer, reruns the hydrologic simulation, and regenerates an updated report, enabling rapid, iterative scenario exploration.

AQUAH supports fully automated hydrologic simulations driven entirely by natural language requests, leveraging LLM-powered image interpretation and decision-making capabilities. It operates within data-available regions, particularly across the contiguous United States (CONUS), enabling both researchers and non-technical users to effortlessly conduct detailed hydrologic analyses.

3.2. Earth Observation

AQUAH automatically harvests the inputs required for hydrologic simulation: (i) basin outlines from the U.S. Geological Survey (USGS); (ii) terrain products—including a Digital Elevation Model, Drainage-Direction Map, and Flow-Accumulation Map—directly from the HydroSHEDS archive; (iii) precipitation forcing from the Multi-Radar/Multi-Sensor (MRMS) system and potential evapotranspiration (PET) fields from USGS, each clipped to the basin envelope; and (iv) in-situ discharge records served by USGS web APIs. Any missing files trigger fallback notifications and sensible default values, so the workflow remains robust across basins with heterogeneous data coverage. See Appendix A for more information.

3.3. Hydrologic Model

For runoff generation and routing, AQUAH employs the distributed CREST (Coupled Routing and Excess STorage) model [7, 24]. CREST solves basin water balance components—precipitation partitioning, infiltration, evapotranspiration, and subsurface exchange—and propagates the resulting flows using a kinematic-wave scheme. Model parameters are exposed for basin-specific calibration, with first-guess values supplied automatically by AQUAH's language-agent modules. Details are in Appendix B.

3.4. Multi-Agent Architecture

As shown in Figure 2, AQUAH is implemented as a *multi-agent system*: a collection of specialized, communicating agents—denoted $\mathcal{A}_*$—that transform a free-form hydrologic modeling request into reproducible simulations, diagnostics, and reports. Each agent owns a well-defined responsibility and passes structured artifacts to the next, enabling transparent reasoning, easier debugging, and seamless extensibility to additional Earth-system tasks.

- **Context Parser Agent** ($\mathcal{A}_{\mathrm{CP}}$) – Parses the user's natural-language request to extract structured simulation metadata—geographic extent, temporal window, and any special constraints—and forwards this normalized envelope to the Dataset Retriever $\mathcal{A}_{\mathrm{CP}}$, ensuring that all subsequent data acquisition is geographically and temporally aligned with the user's intent.

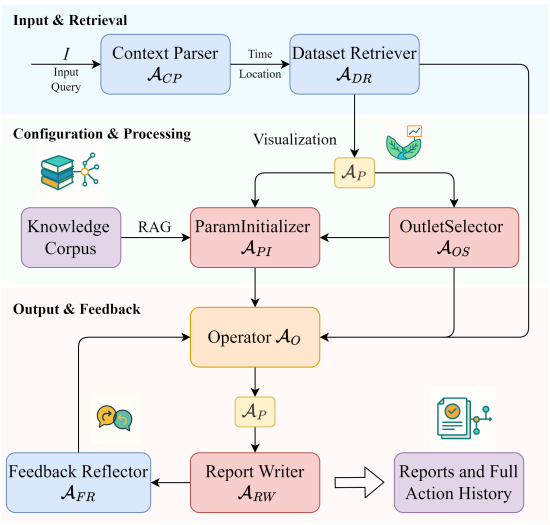


Figure 2. AQUAH multi-agent workflow across input–retrieval, configuration–processing, and output–feedback stages. Agents: $\mathcal{A}_{\mathrm{CP}}$, $\mathcal{A}_{\mathrm{DR}}$, $\mathcal{A}_{\mathrm{P}}$ (Perceptor), $\mathcal{A}_{\mathrm{OS}}$, $\mathcal{A}_{\mathrm{PI}}$, $\mathcal{A}_{\mathrm{O}}$, $\mathcal{A}_{\mathrm{RW}}$, and $\mathcal{A}_{\mathrm{FR}}$.

- **Dataset Retriever Agent** ($\mathcal{A}_{\mathrm{DR}}$) – Receives the spatio-temporal envelope emitted by $\mathcal{A}_{\mathrm{CP}}$; retrieves all mandatory forcing data and baseline geospatial layers (e.g., DEM, land cover, soil) intersecting that envelope; clips each layer to the provisional basin mask; generates quick-look visualisations that are forwarded to the Perceptor $\mathcal{A}_{\mathrm{P}}$ for morphological analysis feeding the OutletSelector $\mathcal{A}_{\mathrm{OS}}$ and ParamInitializer Agent $\mathcal{A}_{\mathrm{PI}}$; and converts the curated datasets (mentioned in §3.2) into the file formats and directory schema required by the Operator $\mathcal{A}_{\mathrm{O}}$.
- **Perceptor Agent** ($\mathcal{A}_{\mathrm{P}}$) – Serves as the vision-perception module: ingests the visual artefacts rendered by $\mathcal{A}_{\mathrm{DR}}$ (e.g., DEMs, flow-accumulation maps, preliminary hydrographs), employs a vision-augmented LLM to extract quantitative descriptors of basin morphology, drainage structure, and candidate gauge sites; supplies these descriptors to the OutletSelector Agent $\mathcal{A}_{\mathrm{OS}}$ and ParamInitializer Agent $\mathcal{A}_{\mathrm{PI}}$, and later interprets simulated versus observed hydrographs to deliver expert diagnostics for iterative refinement.
- **OutletSelector Agent** ($\mathcal{A}_{\mathrm{OS}}$) – Consumes the candidate gauge inventory and drainage descriptors extracted by $\mathcal{A}_{\mathrm{P}}$; applies hydrologic heuristics encoded in the system prompt—such as favouring the gauge closest to the pour point, with long, gap-free records and minimal upstream regulation—to rank the options and designate the optimal basin outlet; emits the selected gauge's identifier and coordinates to both the ParamInitializer $\mathcal{A}_{\mathrm{PI}}$ and Operator $\mathcal{A}_{\mathrm{O}}$ for subsequent simulation steps. The top example of Figure 3 further illustrates the input and output of this gauge selection process. We also provide a detailed case study for VLM-based gauge selection in Section 5.3.

- **ParamInitializer Agent** ($\mathcal{A}_{\mathrm{PI}}$) – Retrieves and parses domain manuals, peer-reviewed literature, and authoritative web resources via RAG to map each model parameter's physical meaning and admissible range; ingests the processed basin attributes from $\mathcal{A}_{\mathrm{P}}$, the forcing datasets prepared by $\mathcal{A}_{\mathrm{DR}}$, and the selected outlet metadata; integrates all evidence to generate a basin-specific, physically plausible initial parameter vector that will seed the Operator Agent's first simulation run.
- **Operator Agent** ($\mathcal{A}_{\mathrm{O}}$) – Ingests the forcing datasets (§3.2) and static rasters curated by $\mathcal{A}_{\mathrm{DR}}$, the outlet definition supplied by $\mathcal{A}_{\mathrm{OS}}$, and the initial parameter vector crafted by $\mathcal{A}_{\mathrm{PI}}$; configures the selected hydrologic models (§3.3) with these inputs; executes the simulation over the user-defined time window; captures full time-series outputs and performance diagnostics such as NSCE, RMSE, and bias; and packages these artefacts for downstream consumption by the Report Writer $\mathcal{A}_{\mathrm{RW}}$ and Feedback Reflector $\mathcal{A}_{\mathrm{FR}}$.
- **Report Writer Agent** ($\mathcal{A}_{\mathrm{RW}}$) – Consolidates the simulation outputs and diagnostics from $\mathcal{A}_{\mathrm{O}}$ with the hydrograph analyses and visual artefacts supplied by $\mathcal{A}_{\mathrm{P}}$; weaves in contextual metadata captured throughout the pipeline (basin description, forcing sources, parameter settings); and auto-compiles a structured, reader-friendly report enriched with maps, hydrographs, statistics, and explanatory narrative for delivery to the end user.
- **Feedback Reflector Agent** ($\mathcal{A}_{\mathrm{FR}}$) – Parses user commentary on the delivered report, updates the internal knowledge state, and, when revisions are warranted, transmits explicit re-run directives—updated parameters, alternative datasets, or extended periods—to the $\mathcal{A}_{\mathrm{O}}$, thereby closing the human-in-the-loop calibration loop.

Based on the above agents, AQUAH autonomously orchestrates the full hydrologic modelling pipeline—parsing user requests, retrieving and preparing geospatial and forcing data, extracting morphological descriptors via vision-augmented LLMs, selecting optimal gauge outlets, initializing model parameters, running simulations, generating diagnostic reports, and incorporating user feedback—without requiring domain expertise or manual intervention.

4. Design of VLM-based Agents

AQUAH embeds cross-modal reasoning by equipping three key agents with large-language-model *vision* capabilities:

$\mathcal{A}_{\mathrm{OS}}$. Guided by the *Outlet Gauge Selector* prompt, the vision–LLM pair first scans the basin map, DEM, and flow-accumulation layers to automatically list every *candidate* station and its attributes. It then applies the ordered rules embedded in the prompt—exclude gauges on reservoirs, favor the lowest-elevation pour point with the largest drainage area and longest gap-free USGS record, and perform a final upstream-reservoir sanity check—iterating until a single outlet meets all criteria and hydrologic common sense.

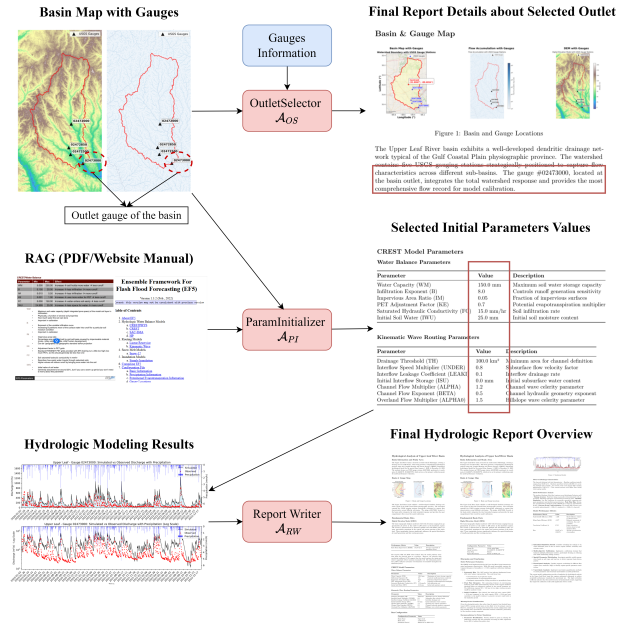


Figure 3. Vision–reasoning interplay within the AQUAH pipeline. The Perceptor Agent ($\mathcal{A}_{\mathrm{P}}$) transforms visual artefacts into structured cues that steer three downstream modules: (i) basin-map rasters yield gauge locations that the OutletSelector ($\mathcal{A}_{\mathrm{OS}}$) ranks to pick the outlet; (ii) geomorphic attributes extracted from the same imagery combine with RAG-sourced documentation to let the ParamInitializer ($\mathcal{A}_{\mathrm{PI}}$) populate an initial parameter table; and (iii) simulated hydrographs are interpreted to provide narrative context for the Report Writer ($\mathcal{A}_{\mathrm{RW}}$). The panels on the right illustrate how each branch of the vision workflow materialises in the final deliverable: outlet information, auto-selected parameter values, and a consolidated hydrologic analysis overview.

After ingesting the basin-with-gauges image (Fig. 3, upper-left), the agent returns two plain-text lines:

`Selected gauge: [ID]` and `Explanation: [brief justification]`, providing both a machine-readable choice and a concise, human-readable rationale.

Outlet Gauge Selector Prompt

You are a hydrologist who can interpret maps and select the most appropriate USGS gauge to represent the **natural** basin outlet. The user supplies: (1) a base-map with watershed boundary, (2) a DEM with gauges, and (3) a flow-accumulation map with gauges.
Apply the following ordered rules when selecting **ONE** gauge (earlier rules override later ones):
0) If the user's text clearly mentions or implies a specific gauge, city, or location, select it.

1) *Exclusion* – disqualify any gauge located downstream of, or directly on, a reservoir/lake.
2) From the remaining gauges, prefer the one at the lowest-elevation point on the basin boundary where flow naturally exits (use DEM).
3) Prefer gauges capturing the largest drainage area and highest flow-accumulation values.
4) Prefer gauges with extensive, reliable USGS discharge records.
5) *Second verification* – re-check that the chosen gauge is upstream of all reservoirs/lakes and sits at a natural outlet. If not, discard it and re-evaluate.
Return your response in this format:
`Selected gauge: [gauge ID number]`
`Explanation: [brief justification]`

$\mathcal{A}_{\text{PI}}$. The initializer prompt is fed by two descriptive inputs. *Basin description* (`{basin_desc}`) is produced by a VLM that "reads" the basin-map images (Fig. 3), then writes a short paragraph summarising key traits such as drainage area, relief, and dominant slope classes. *Parameter guide* (`{guide}`) comes from a RAG pipeline that scans PDF manuals and web pages, condensing each source into plain-language hints on plausible CREST parameter ranges. Given these two narrative snippets, $\mathcal{A}_{\text{PI}}$ returns a one-line JSON object: a full CREST parameter vector plus a brief justification for every value, providing both machine-ready inputs and transparent reasoning.

CREST Parameter Initializer Prompt

You are a hydrologist. Using the parameter guide and basin description, propose first-guess CREST parameters.
Basin description (from LLM *image interpretation*):`{basin_desc}`
Parameter guide (from LLM *document/web summarisation*):`{guide}`
Return exactly one line of JSON:
```
{"code":"crest_args =
types.SimpleNamespace(wm=<value>,
b=<value>, im=<value>, ...",
"explanation":"each param justified in
100-300 words"}
```
No Markdown, no extra keys.

$\mathcal{A}_{\text{RW}}$. Once the run finishes, $\mathcal{A}_{\text{RW}}$ fills the *Hydrological Report Writer* prompt with two text fragments: `summary`—an auto-generated paragraph that turns stored run metadata (basin name, simulation window, chosen gauge, key metrics) into plain language—and `figures_description`—sentences returned by the vision LLM after inspecting the maps and hydrographs. Guided by the prompt checklist, the agent assembles a Markdown file that includes (i) a title and basin locator map; (ii) cartographic layers for spatial context; (iii) rainfall-and-discharge plots; (iv) a table of CREST parameters and NSE, KGE, CC, bias, RMSE scores; and (v) a short discussion of results and recommended next steps. The Markdown is then rendered to PDF, giving users a compact, self-contained overview of basin features, model behaviour, and forecast quality (see report overview in Fig. 3,lower right).

Hydrological Report Writer Prompt

Description: Using the provided simulation metadata and results
– `summary: {summary}`
– `figures description: {figures_desc}`
generate a complete Markdown report containing:

1. **Title and Basin Information**
   – Level-1 heading with basin name.
   – Basin & gauge map, basic data and brief introduction.
2. **Analysis Sections**
   – Simulation vs observation comparison.
   – Model performance metrics.
   – CREST parameters.
   – Conclusion/Discussion.
3. **Required Images** (`![]()`):
   `combined_maps.png`, `results.png`
4. **Data Tables** (run arguments, metrics, parameters) – vertical listing.
5. **Discussion Points**
   – Model performance evaluation.
   – Warm-up period considerations if bias $< -90\%$.
   – Recommendations for future runs.

Expected output: a complete, publication-ready Markdown report (no extra text after the report).

5. Experiments

5.1. Settings

Hydrologic backbone and geospatial toolkit. AQUAH couples the distributed *CREST* model with standard open-source GIS utilities (GDAL, Rasterio, Shapely, and Folium) for raster reprojection, vector clipping, and mapping.

Large-language models. Three vision-capable LLMs are benchmarked: GPT-4o (OpenAI) [18], Claude-Sonnet-4 (Anthropic) [1], and Gemini-2.5-Flash (Google) [10]. Please refer to Appendix A for further details.

Earth-observation inputs. Daily precipitation forcing is provided by MRMS (1 km grid); potential evapotranspiration by USGS FEWS NET (1° grid); terrain layers—DEM, drainage-direction, and flow-accumulation—by HydroSHEDS at 3-arc-second (90 m) resolution; and

Model	Model Comp.	Sim. Results	Reasonableness	Clarity	Average
claude-4-opus	**7.51**	5.60	**6.97**	**7.95**	**7.01**
claude-4-sonnet	7.43	5.46	6.77	7.49	6.79
gpt-4o	6.74	**5.89**	6.06	6.51	6.30
o1	7.11	4.80	6.23	6.94	6.27
gemini-2.5-flash	6.91	4.68	6.28	6.57	6.11

Table 1. Quantitative evaluation of hydrological-report generation. Bold values mark the best score in each column.

discharge observations by USGS NWIS, which also supplies gauge metadata such as drainage area.

End-to-end automation. User prompts are decomposed by AQUAH's agent stack into structured tasks that autonomously handle data download, model runs, and post-processing. The workflow therefore spans geospatial preprocessing, Earth-observation ingestion, hydrologic simulation, and report generation without manual intervention.

5.2. Benchmark & Test Results

To quantify the quality of the hydrological–simulation reports produced by our *AQUAH*, we conducted a two–tier evaluation. All reports were anonymized and randomly ordered to ensure that evaluators were blind to the source model, thereby eliminating potential bias.

Domain–expert review. Several professional hydrologists were asked to score each report on a 10-point scale (1 = poor, 10 = excellent) along four facets that are critical for decision-grade hydrological studies: *Model Completeness*, *Simulation Results*, *Reasonableness*, and *Clarity*. Figure8 in Appendix C shows the detailed human judge interface.

LLM co-evaluation. In addition, we used the latest OpenAI `gpt-o3` model as an impartial, large-scale language model (LLM) judge. This hybrid protocol mitigates individual-expert variance while leveraging the consistency of an automated evaluator.

For each axis we take the arithmetic mean of the human and LLM scores; the Average column is the unweighted mean across the four axes. The numerical results are summarised in Table 1. Overall, `claude-4-opus` achieves the highest average score (7.01), leading or tying on three of the four criteria, and outperforming all other contenders by at least 0.22 points. Refer to Appendix C for further evaluation details.

The clear margin of `claude-4-opus` indicates that, for our task setup, higher model–completeness and more coherent reasoning translate directly into more actionable hydrological insights. Conversely, despite producing the strongest raw simulation summaries, `gpt-4o` lags on clarity, underscoring the need for balanced optimisation across all evaluation axes.

5.3. Selected Case Studies

LLM-Vision-driven outlet gauge selection. The outlet–selection agent $\mathcal{A}_{\mathrm{OS}}$ receives two georeferenced rasters—the flow-accumulation map and the DEM—together with vector layers of candidate gauges and their attributes (elevation, drainage area).

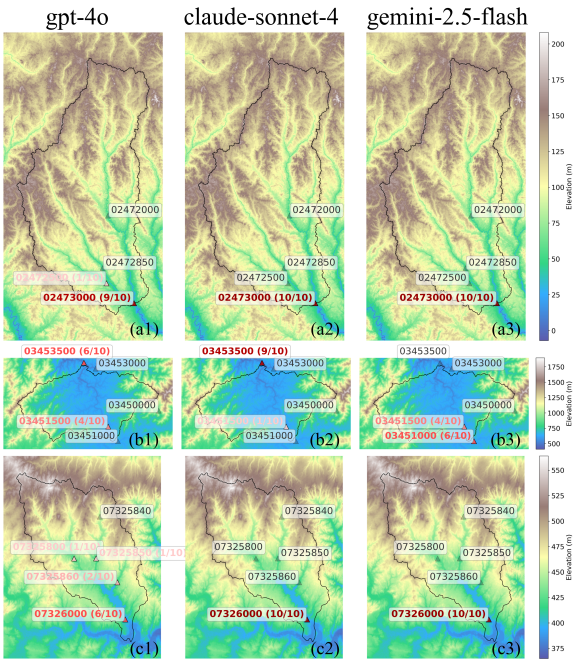


Figure 4. Gauge-selection frequency across three basins (rows a–c) and three LLMs (columns: GPT-4o, Claude-Sonnet-4, Gemini-2.5-Flash). Each basin–LLM combination was run ten times; gauge labels are shaded from black (never selected) to deep red (selected in all 10 runs).

At inference time a VLM (GPT-4o, Claude-Sonnet-4, or Gemini-2.5-Flash) is prompted with these layers plus a short ordered rule set: (i) respect any user-specified gauge; (ii) disregard stations situated on or below reservoirs/lakes; (iii) favour the lowest-elevation gauge on the

basin perimeter; (iv) prefer larger drainage areas and higher flow-accumulation values; and (v) break remaining ties with data-record quality. All gauges inside the watershed, along with those lying just downstream of the polygon, are considered so that official outlets positioned slightly outside the boundary are not overlooked.

Figure 4 summarises the agent's behaviour across three contrasting basins. In the simple, single-outlet catchment (row a) every LLM converges on the same gauge in almost every trial. The more dendritic basin in row b exposes subtle differences: GPT-4o and Claude-Sonnet-4 nearly always pick the hydrologically dominant tributary, whereas Gemini-2.5-Flash splits its choices between two interior stations, reflecting ambiguity in topographic cues. Row c highlights the importance of contextual constraints—when a major reservoir sits just upstream of the nominal outlet, all models occasionally retain the regulated gauge unless the prompt explicitly flags reservoir positions.

LLM-based first-guess parameterisation. Reliable calibration begins with a defensible *first guess*, yet selecting a plausible vector of hydrologic model parameters normally demands years of field intuition and many trial–error cycles. AQUAH tackles this bottleneck with the agent $\mathcal{A}_{\mathrm{PI}}$, which blends retrieval-augmented generation and vision reasoning. The agent consults CREST manuals(ingested as PDFs and web pages) to learn each parameter's physical role and admissible range, while simultaneously analysing basin rasters—DEM, FAM, DDM, and a land-cover basemap—to infer slope, drainage density, soil moisture capacity, and impervious fraction. Guided by this fused knowledge, the VLM proposes a basin-specific start vector (e.g. WM, B, KE, see Appendix B for more information), runs CREST once, and logs the resulting skill scores.

Figure 5b–c displays the distributions of CC and NSCE obtained from ten independent initialisations per LLM, while panels (d–f) show each model's best hydrograph and its associated parameter set. The text blocks underneath capture the LLM-generated rationale—for example, boosting WM in forest-dominated headwaters or lowering B over urban sub-catchments. Across the test basin the agent's proposals consistently land within physically reasonable bounds and, in several cases, achieve near-calibrated performance on the very first run. These results indicate that modern vision-LLMs already possess a rudimentary grasp of hydrologic parameter semantics, substantially shortening the path from "cold start" to productive calibration.

6. Discussion

Gauge–Outlet Selection (Fig. 4). Row (a) represents a straightforward basin; all three LLMs consistently identify the true outlet, confirming that the rule set is sufficient for simple landscapes. Row (b) introduces competing tributaries. The correct outlet is gauge `03453500`; GPT-4o selects it in 6/10 trials, Claude-Sonnet-4 achieves 9/10, whereas Gemini-2.5-Flash never resolves the ambiguity. Row (c) tests reservoir awareness: gauge `07326000` lies immediately downstream of a dam and should be rejected. All models struggle—most runs still choose the regulated site—although GPT-4o avoids it in 40 % of trials. These results highlight that discerning human regulation from natural flow remains challenging for current mainstream LLMs. Incorporating additional Earth-observation cues (e.g. reservoir masks) or employing stronger reasoning models with tailored prompts may mitigate this limitation.

Parameter-Initialisation Performance (Fig. 5). Across ten independent initialisations, the three LLMs exhibit distinct variance patterns. GPT-4o delivers the single highest CC/NSCE score but shows the widest spread, indicating strong stochasticity between runs. Claude-Sonnet-4 is the most consistent: its box-plots are narrow and uniformly positive, making it the most reliable performer for this basin despite not achieving the absolute best score. Gemini-2.5-Flash also displays high run-to-run variability but, unlike GPT-4o, its median skill is noticeably lower, leading to overall weaker performance. In sum, GPT-4o can produce outstanding results but requires multiple attempts; Claude offers dependable, high-quality starts.

Although none of the first-guess parameter sets is fully optimal, each agent delivers values of the correct order of magnitude—an outcome far superior to ad-hoc, manual guessing and crucial for a successful first run. Looking forward, the CREST outputs can be fed back to the *ParamInit Agent*, enabling RAG-guided, step-wise adjustments that iteratively refine parameters. Such a feedback loop would provide directionally consistent, interpretable calibration without exhaustive trial-and-error.

Limitation. Our study reused one prompt template—originally tuned for the OpenAI API—across all language-model back ends, which, while convenient for benchmarking, may not fully leverage each model's unique formatting or capabilities. The prototype also depends on publicly hosted data and inference services (e.g., USGS, MRMS, and a cloud LLM), so regional access limits or temporary outages could reduce functionality. Future work will investigate model-specific prompt tuning and local or cached data to ease these constraints.

7. Conclusion

We present **AQUAH**, the first end-to-end hydrologic-simulation *agent* that translates free-form language requests into physically consistent model runs and publication-ready reports. By coupling a large language model with vision

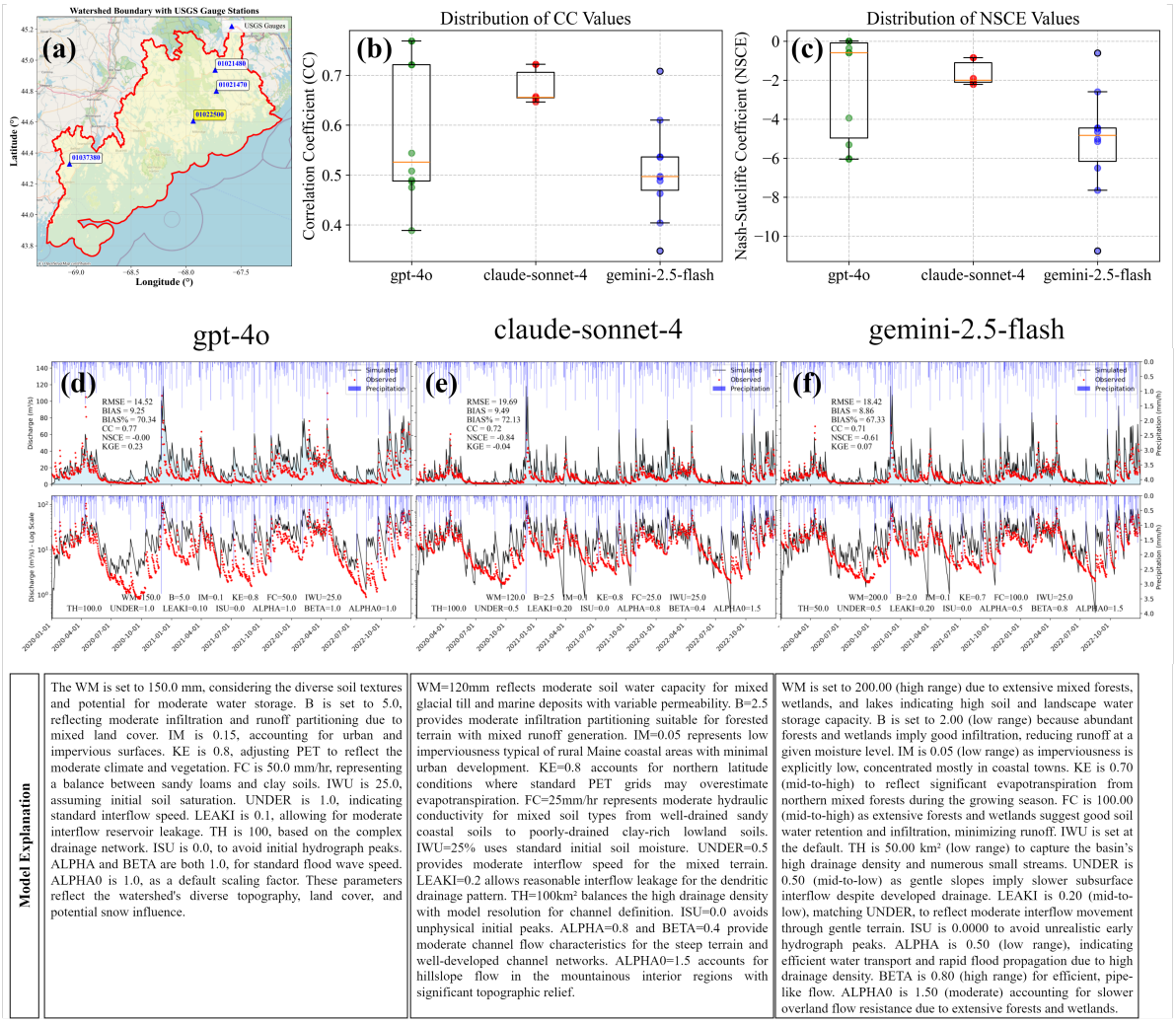


Figure 5. Performance summary for the Maine Coastal Basin. (a) Basin boundary and gauge locations. (b) Box-plot of CC from ten parameter-initialization runs for each LLM (GPT-4o, Claude-Sonnet-4, Gemini-2.5-Flash). (c) Corresponding NSCE box-plots. (d–f) Best-performing hydrographs for each LLM, annotated with their error metrics and calibrated CREST parameter sets. The lower text blocks provide the LLM-generated rationale for the chosen parameter values.

modules for DEM reading, gauge detection, and parameter inference, AQUAH automates the full workflow—from data acquisition through CREST simulation to visualization—requiring neither domain expertise nor coding. Benchmarks across several mainstream LLMs show AQUAH delivers decision-grade outputs with low entry barriers.

Beyond hydrology, the modular design illustrates how *LLM + CV* synergies can spawn specialised agents for other simulation-driven sciences across the globe. We argue that building task-aware toolchains around foundation models will become a core paradigm for next-generation platforms, enabling rapid, democratised access to complex numerical engines across Earth-system, engineering, biomedical, environmental, and climate domains.

Acknowledgments

We gratefully acknowledge the anonymous hydrologic experts who evaluated our agentic modeling framework and provided valuable feedback. This research was supported by NASA Award No.22-IDS22-0122 / 80NSSC24K0351, entitled "Predicting the Impact of Contemporary Climate Extremes on Nitrogen Flux along the Land-to-Ocean Continuum: Integrated Remote Sensing and Modeling Applied to the Mississippi River Basin."

References

[1] Anthropic. Claude sonnet 4: A hybrid-reasoning ai model. Technical Report / Model Card claude-sonnet-4@20250514, Anthropic, Inc., San Francisco, CA, 2025. Mid-sized model

in the Claude 4 family; 200 K token context, extended "thinking" mode, strong coding and reasoning capabilities, safety at AI Safety Level 2. 2, 5

[2] Richard Arsenault, David Huard, Jean-Luc Martel, Magali Troin, Juliane Mai, François Brissette, Christian Jauvin, Long Vu, James R Craig, Trevor J Smith, et al. The pavics-hydro platform: A virtual laboratory for hydroclimatic modelling and forecasting over north america. *Environmental Modelling & Software*, 168:105808, 2023. 2

[3] Abdolreza Bahremand. Hess opinions: Advocating process modeling and de-emphasizing parameter estimation. *Hydrology and Earth System Sciences*, 20(4):1433–1445, 2016. 1

[4] Benedikt Blumenstiel, Paolo Fraccaro, Valerio Marsocci, Johannes Jakubik, Stefano Maurogiovanni, Mikolaj Czerkawski, Rocco Sedona, Gabriele Cavallaro, Thomas Brunschwiler, Juan Bernabe Moreno, et al. Terramesh: A planetary mosaic of multimodal earth observation data. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2394–2402, 2025. 2

[5] J Stuart Carlton and Susan K Jacobson. Using expert and non-expert models of climate change to enhance communication. *Environmental Communication*, 10(1):1–24, 2016. 1

[6] Mengye Chen, Zhi Li, Humberto J Vergara, Jonathan J Gourley, Ming Xue, Yang Hong, Xiao-Ming Hu, Hector Mayol Novoa, Elinor R Martin, Renee A McPherson, et al. Conus-wide model calibration and validation for crestv3. 0–an improved coupled routing and excess storage distributed hydrological model. *Journal of Hydrology*, 626:130333, 2023. 11

[7] Zachary L Flamig, Humberto Vergara, and Jonathan J Gourley. The ensemble framework for flash flood forecasting (ef5) v1. 2: Description and case study. *Geoscientific Model Development*, 13(10):4943–4958, 2020. 3, 11

[8] Shang Gao, Zhi Li, Mengye Chen, Peirong Lin, Zhen Hong, Daniel Allen, Thomas Neeson, and Yang Hong. Spatiotemporal variability of global river extent and the natural driving factors revealed by decades of landsat observations, grace gravimetry observations, and land surface model simulations. *Remote Sensing of Environment*, 267:112725, 2021. 1

[9] Philip W Gassman, Ali M Sadeghi, and Raghavan Srinivasan. Applications of the swat model special section: overview and insights. *Journal of Environmental Quality*, 43(1):1–8, 2014. 2

[10] Google DeepMind. Gemini 2.5 flash: A hybrid reasoning ai model. Technical Report (Model Card) gemini-2.5-flash-preview-05-20, Google LLC, Mountain View, CA, 2025. Preview version available via Vertex AI / AI Studio; supports controllable reasoning with 1M token context and 64K output window. 5

[11] Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. *arXiv preprint arXiv:2504.11171*, 2025. 2

[12] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9, 2024. 2

[13] Zhi Li, Mengye Chen, Shang Gao, Xiangyu Luo, Jonathan J Gourley, Pierre Kirstetter, Tiantian Yang, Randall Kolar, Amy McGovern, Yixin Wen, et al. Crest-imap v1. 0: A fully coupled hydrologic-hydraulic modeling framework dedicated to flood inundation mapping and prediction. *Environmental Modelling & Software*, 141:105051, 2021. 2, 11

[14] Zhi Li, Shang Gao, Mengye Chen, Jonathan J Gourley, Changhai Liu, Andreas F Prein, and Yang Hong. The conterminous united states are projected to become more prone to flash floods in a high-end emissions scenario. *Communications Earth & Environment*, 3(1):86, 2022. 1

[15] Zhi Li, Xianwu Xue, Robert Clark, Humberto Vergara, Jonathan Gourley, Guoqiang Tang, Xinyi Shen, Guangyuan Kan, Ke Zhang, Jiahu Wang, et al. A decadal review of the crest model family: Developments, applications, and outlook. *Journal of Hydrology X*, 20:100159, 2023. 1, 11

[16] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2

[17] Marjolein Mens, Bennie Minnema, Koen Overmars, and Bart Van den Hurk. Dilemmas in developing models for long-term drought risk management: The case of the national water model of the netherlands. *Environmental Modelling & Software*, 143:105100, 2021. 1

[18] OpenAI. Gpt-4o system card, 2024. 2, 5

[19] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. Also described in Microsoft Research technical report. 2

[20] Katharine Sanderson. Gpt-4 is here: what scientists think. *Nature*, 615(7954):773, 2023. 2

[21] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36:38154–38180, 2023. 2

[22] Daniela Szwarcman, Sujit Roy, Paolo Fraccaro, Þorsteinn Elí Gíslason, Benedikt Blumenstiel, Rinki Ghosal, Pedro Henrique de Oliveira, Joao Lucas de Sousa Almeida, Rocco Sedona, Yanghui Kang, et al. Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications. *arXiv preprint arXiv:2412.02732*, 2024. 2

[23] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 2

[24] Jiahu Wang, Yang Hong, Li Li, Jonathan J Gourley, Sadiq I Khan, Koray K Yilmaz, Robert F Adler, Frederick S Policelli, Shahid Habib, Daniel Irwn, et al. The coupled routing and excess storage (crest) distributed hydrological model. *Hydrological sciences journal*, 56(1):84–98, 2011. 1, 2, 3

[25] Songkun Yan, Ziqiang Ma, Xiaoqing Li, Hao Hu, Jintao Xu, Qingwen Ji, and Fuzhong Weng. Pcssr-dnnwa: A physical

constraints based surface snowfall rate retrieval algorithm using deep neural networks with attention module. *Geophysical Research Letters*, 50(13):e2023GL103923, 2023. 2

[26] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[27] Siyu Zhu, Zhi Li, Mengye Chen, Yixin Wen, Shang Gao, Jiaqi Zhang, Jiao Wang, Yi Nan, Sebastian C Ferraro, Theresa E Tsoodle, et al. How has the latest imerg v07 improved the precipitation estimates and hydrologic utility over conus against imerg v06? *Journal of Hydrology*, 645: 132257, 2024. 11

A. Experimental Setup

Implementation framework. All agents are orchestrated with `crewAI` v0.75.0[1], which provides the task queue, tool interface, and inter-agent messaging used throughout AQUAH.

ParamInitializer workflow. For illustration we focus on the *ParamInitializer Agent*, whose logic is divided between two Python functions. `describe_basin_for_crest()` prompts a vision-enabled LLM to summarise basin physiography from DEM, flow-accumulation, drainage-direction rasters, and a locator map; `estimate_crest_args()` then launches a CrewAI agent that mines PDF manuals and websites to propose a physically plausible CREST parameter vector. A provider-agnostic wrapper converts images to the base-64 or `PIL.Image` formats required by OpenAI, Anthropic, or Gemini APIs; oversized payloads are iteratively down-scaled and JPEG-compressed to satisfy the strictest quota (5 MB for Claude).

Large-language models. Five mainstream models are queried via their native endpoints: GPT-4o (`gpt-4o`), Claude-4 Sonnet (`claude-4-sonnet-20250514`), GPT-o1 (`o1`), Claude-4 Opus (`claude-4-opus-20250514`), and Gemini-2.5 Flash (`gemini-2.5-flash-preview-05-20`). Text-only prompts use a deterministic temperature of 0, whereas vision prompts use 0.3.

Earth-observation data. Input layers are fetched on demand from public repositories: HydroSHEDS 90 m DEM, flow-accumulation, and drainage-direction rasters (https://hydrosheds.org/); USGS 3DEP high-resolution DEMs (https://apps.nationalmap.gov/downloader/); MRMS precipitation archives (https://mtarchive.geol.iastate.edu/); FEWS-NET potential-evapotranspiration grids (https://earlywarning.usgs.gov/fews/product/81); and USGS NWIS discharge records (https://waterdata.usgs.gov/nwis). All layers are clipped to the basin polygon produced by the CONTEXTPARSER agent and re-projected to a common grid before model execution.

B. CREST

EF5/CREST model description. The EF5/CREST (Coupled Routing and Excess STorage) hydrologic modelling framework—originating from the University of Oklahoma in collaboration with NASA—combines distributed water-balance calculations with kinematic-wave routing to deliver rapid, spatially explicit flood simulations. Over the past decade it has evolved into a versatile research and operational tool: CREST-iMAP couples hydrologic and hydraulic components for real-time inundation mapping [13]; continental-scale calibration and validation have demonstrated robust skill across the CONUS domain [6]; the framework has been leveraged to diagnose forcing uncertainties such as the impact of IMERG precipitation upgrades on streamflow prediction [27]; and a recent synthesis highlights continued advances and emerging applications across global flood forecasting, drought assessment, and land–surface interaction studies [15]. These studies underscore the model family's breadth and its suitability for the automated, agent-driven workflows pursued in AQUAH.

EF5/CREST Parameter Cheat-Sheet. The EF5/CREST hydrologic model framework separates calibration parameters into two broad blocks: (i) *runoff generation* governed by the CREST/Water-Balance scheme and (ii) *kinematic-wave routing* [7, 15]. Tables 2 and 3 list the key parameters, their recommended search ranges, and the qualitative hydrologic response when each value increases. This compact sheet is intended as a quick reference for modellers when setting up automatic or manual calibration routines.

C. Evaluation Criteria

The quality of each AQUAH-generated simulation is assessed through a two-tier protocol that combines *objective statistical metrics* and *human expert review*. The former quantify the numerical agreement between simulated and observed discharge, while the latter capture practitioner-oriented aspects such as interpretability and report readability.

Objective Verification Metrics Following established hydrological practice, five complementary statistics are evaluated over the full period (see Table 4). These are: the *Nash–Sutcliffe efficiency* (NSE, $-\infty$–1, ideal 1), which summarises overall predictive skill; the *Kling–Gupta efficiency* (KGE, ideal 1) that balances correlation, bias and variability; the *Pearson*

[1] https://github.com/crewAIInc/crewAI

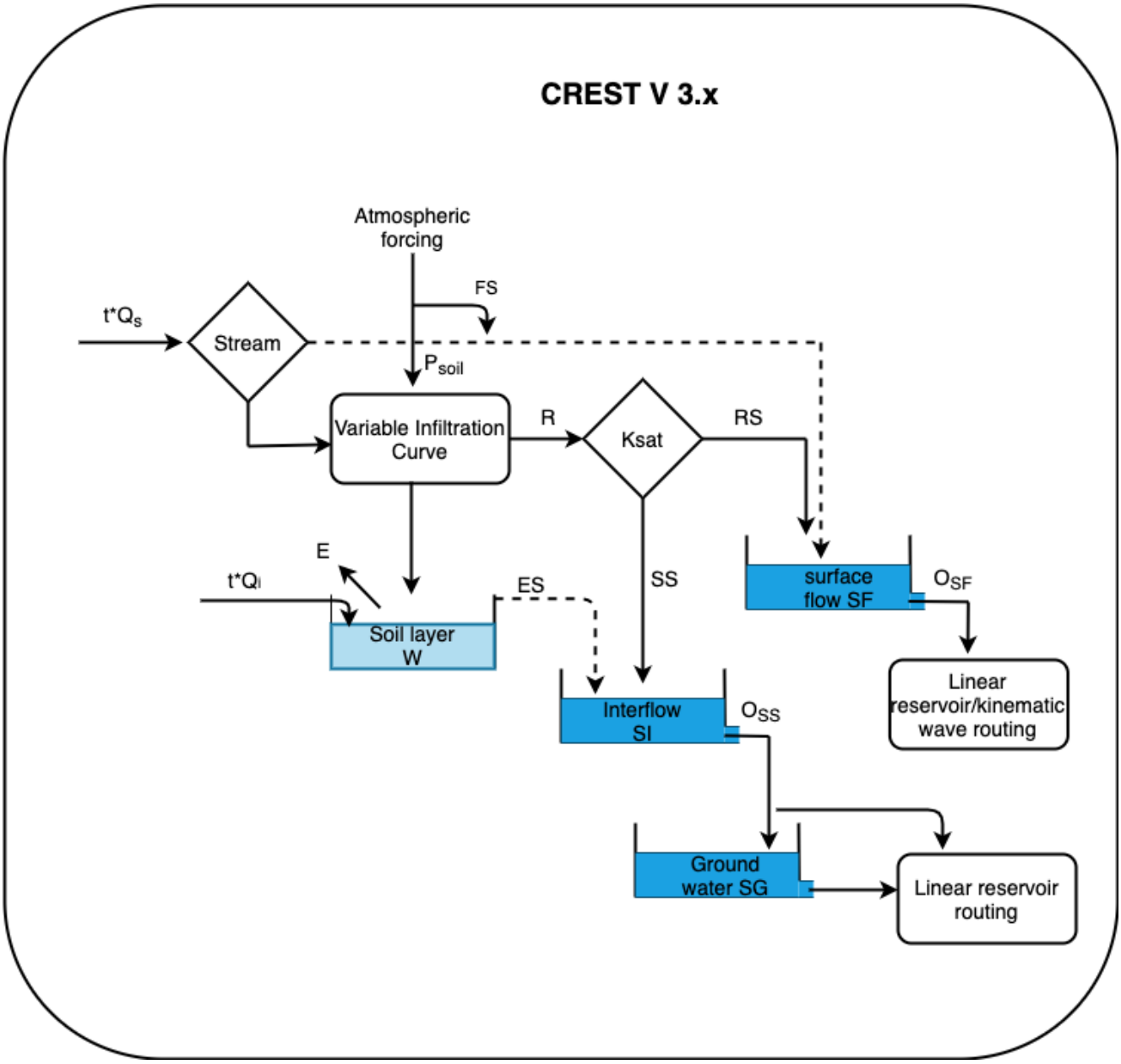


Figure 6. A schematic of the hydrologic processes represented by the latest EF5/CREST model

correlation coefficient (CC, ideal 1); the *root mean square error* (RMSE), where lower values indicate smaller deviations; and the *relative bias* (BIAS), whose optimum is 0. Together they diagnose both the accuracy and reliability of the CREST simulations across all flow regimes.

Final Report Evaluation Beyond the objective metrics, every report is first uploaded to the latest o3 large-language model for automated grading and then independently assessed—under blinded conditions—by a team of professional hydrologists, both parties applying the same four-axis rubric. *Model Completeness* gauges the suitability of data sources, openness of parameter disclosure, and overall workflow transparency; *Simulation Results* reflects the fidelity of hydrographs and accompanying statistics, including treatment of uncertainties; *Reasonableness* judges the physical plausibility of parameter choices, underlying assumptions, and recommended next steps; and *Clarity* measures readability, logical flow, figure and table quality,

Table 2. CREST / Water-Balance parameters

Parameter	Meaning	Range	Effect when value increases
WM	Maximum soil-water storage capacity (mm)	5–250	More storage ⇒ **less** direct runoff.
B	Infiltration curve exponent	0.1–20	Steeper curve ⇒ **more** surface runoff.
IM	Fraction of impervious area	0.01–0.50	Larger imperviousness ⇒ **more** runoff.
KE	PET utilisation / evapotranspiration coefficient	0.001–1.0	Higher ET loss ⇒ **less** runoff.
FC	Saturated hydraulic conductivity proxy (mm h^{-1})	0–150	Faster infiltration ⇒ **less** runoff.
IWU	Initial soil-water content (mm)	0–25	Wetter initial state ⇒ **higher** early runoff.

Table 3. Kinematic-wave routing parameters

Parameter	Meaning	Range	Effect when value increases
TH	Drainage-area threshold (km^2)	30–300	Smaller threshold ⇒ finer channel network.
UNDER	Interflow velocity multiplier (m s^{-1})	0.0001–3.0	Larger velocity ⇒ **quicker** runoff response.
LEAKI	Leakage factor from interflow layer	0.01–1.0	Higher leakage ⇒ faster hydrograph rise.
ISU	Initial subsurface storage unit	0–1 $\times 10^{-5}$	Non-zero may cause spurious early peak; keep near zero.
ALPHA	Muskingum–Cunge α for channel cells	0.01–3.0	Larger value **slows** flood-wave translation.
BETA	Muskingum–Cunge β for channel cells	0.01–1.0	Bigger β likewise slows and attenuates wave.
ALPHA0	α for overland/non-channel cells	0.01–5.0	Controls overland flow speed; β fixed at 0.6.

Table 4. Verification metrics used in this study. Q^t_{obs} (Q^t_{sim}) is the observed (simulated) discharge at time step t; $\overline{Q}_{\text{obs}}$ and $\overline{Q}_{\text{sim}}$ are their respective means; μ and σ are the mean and standard deviation; T is the total number of time steps. *CC* – Pearson correlation coefficient, *BIAS* – relative bias, *RMSE* – root mean square error, *NSE* – Nash–Sutcliffe efficiency, *KGE* – Kling–Gupta efficiency with $\alpha = \sigma_{\text{sim}}/\sigma_{\text{obs}}$ and $\beta = \mu_{\text{sim}}/\mu_{\text{obs}}$. The last column gives each metric's theoretical range and its perfect value (in parentheses).

Metric (abbr.)	Equation	Range (perfect)
Nash–Sutcliffe efficiency (NSE)	$NSE = 1 - \frac{\sum_{t=1}^{T}(Q^t_{\text{obs}} - Q^t_{\text{sim}})^2}{\sum_{t=1}^{T}(Q^t_{\text{obs}} - \overline{Q}_{\text{obs}})^2}$	$(-\infty, 1]$ (1)
Relative bias (BIAS)	$BIAS = \frac{1}{T}\sum_{t=1}^{T}(Q^t_{\text{sim}} - Q^t_{\text{obs}})$	$(-\infty, \infty)$ (0)
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(Q^t_{\text{sim}} - Q^t_{\text{obs}})^2}$	$[0, \infty)$ (0)
Correlation coefficient (CC)	$CC = \frac{\sum_{t=1}^{T}(Q^t_{\text{sim}} - \overline{Q}_{\text{sim}})(Q^t_{\text{obs}} - \overline{Q}_{\text{obs}})}{\sqrt{\sum_{t=1}^{T}(Q^t_{\text{sim}} - \overline{Q}_{\text{sim}})^2}\sqrt{\sum_{t=1}^{T}(Q^t_{\text{obs}} - \overline{Q}_{\text{obs}})^2}}$	$[-1, 1]$ (1)
Kling–Gupta efficiency (KGE)	$KGE = 1 - \sqrt{(CC-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$	$(-\infty, 1]$ (1)

and adherence to scientific-writing norms. Each axis is scored on an integer 0–10 scale by the expert panel and the LLM; the two values are averaged to obtain the axis score, and the unweighted mean across the four axes yields an overall quality index (see the UI mock-up in Fig. 7).

Example Analysis Figure 8 contrasts two reports generated from the identical prompt "*I want to simulate the streamflow of the Mad–Redwood basin from 2020 to 2022*." Panel (a) shows *B5_030.pdf*, produced by the `gemini-2.5-flash` agent, while panel (b) shows *B5_223.pdf* from `gpt-o1`. Although both agents follow the same workflow, their outputs diverge noticeably: `B5_030` omits several key figures, lowering its *Model Completeness* score, and its poor NSE drags down the *Simulation Results*. In contrast, `B5_223` includes all requisite graphics and attains a substantially better NSE (0.578), which,

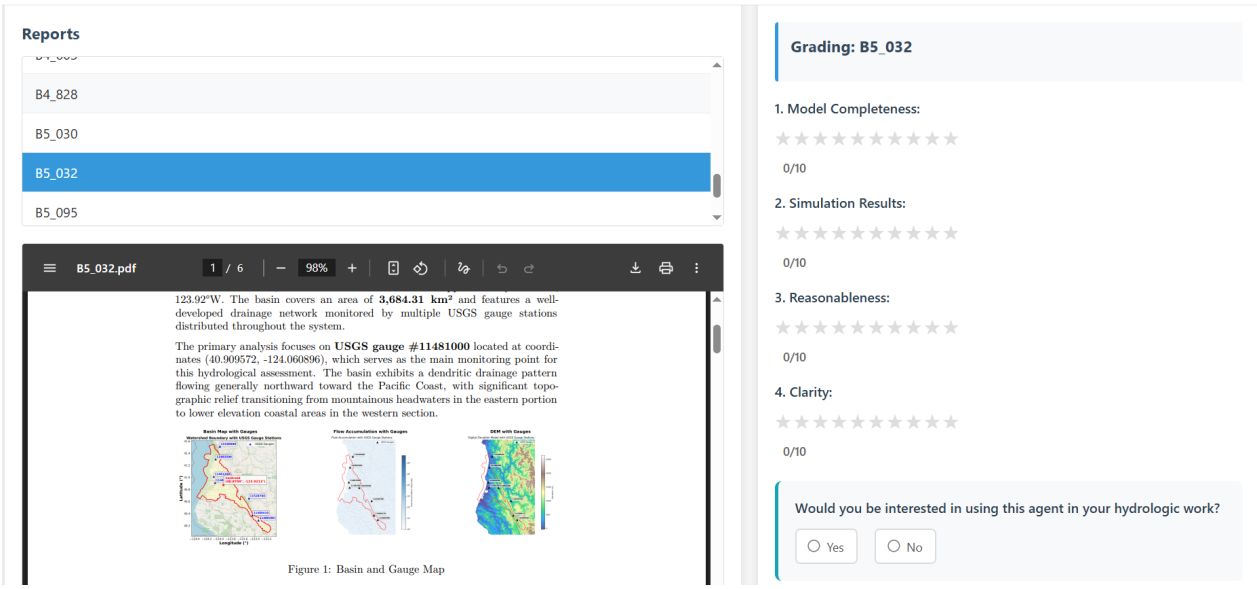


Figure 7. Human-grading interface used in this study. Experts (and an LLM co-evaluator) assign 0–10 star scores on four axes—Model Completeness, Simulation Results, Reasonableness, and Clarity—and record whether they would adopt the agent in professional hydrologic work.

together with clearer recommendations, yields higher marks across all four grading axes and a superior overall index. This example underscores how agent choice can strongly influence both the technical fidelity and presentation quality of first-pass hydrologic simulations.

(a)

Mad-Redwood Basin Hydrological Analysis

1. Introduction to the Mad-Redwood Basin

The Mad-Redwood basin, encompassing an area of 3684.31 km², is a coastal watershed located in northern California. This report presents a hydrological analysis of the basin, focusing on the period from 2020-01-01 to 2022-12-31. The analysis utilizes the CREST hydrological model, with a specific focus on USGS gauge #11481000 (40.909572, -124.060896) as the primary point of evaluation. This report details the model setup, simulation results, performance metrics, and recommendations for future improvements.

2. Basin & Gauge Map

The basin exhibits a distinct north-south orientation along the Pacific coastline. Key USGS gauge locations are strategically positioned throughout the watershed, providing valuable streamflow data. The gauge network includes several key stations. From upstream to downstream, the apparent main branch gauges are approximately: 11530500, 11482500, 11481200, 11481000, 11528700, 11480410, and 11480390. Gauge 11481000 appears to be located near the centroid (40.8709°, -123.9213°). The topography indicates steeper slopes in the northern parts of the basin, potentially influencing runoff dynamics.

Basin & Gauge Map

3. Fundamental Basin Data

The Digital Elevation Model (DEM) reveals significant topographic relief within the Mad-Redwood basin, ranging from near sea level to elevations exceeding 2500 meters. The higher elevations are concentrated in the northern and eastern regions, likely leading to orographic precipitation patterns. The Flow Accumulation Map (FAM) clearly delineates a primary river channel and its associated tributary network converging downstream. The logarithmic scale of the FAM accentuates the increasing flow accumulation along the main channel. The Drainage Direction Map (DDM) illustrates the dominant flow pathways, with a consistent pattern of water converging into the main channel and its tributaries. The relatively high drainage density evident in the DDM suggests a well-developed stream network, which can contribute to a rapid hydrologic response to precipitation events.

Fundamental Basin Data

4. Simulation vs Observation

The following figure compares the simulated and observed discharge at USGS gauge #11481000, along with precipitation data. The simulated discharge generally captures the overall trend of the observed discharge, including the timing of

(b)

Hydrological Simulation Report for Mad-Redwood

1. Title and Basin Information

Basin & Gauge Map



Figure 1: Basin & Gauge Map

Basin Overview

The Mad-Redwood basin is a coastal watershed encompassing 3684.31 km². Steep slopes in the interior transition rapidly toward lower coastal elevations, creating flashier responses to rainfall events. The basin's centroid is near (40.8709°, -123.9213°). The USGS gauge #11481000 used for this simulation is located at (40.909572, -124.060896).

Fundamental Basin Data

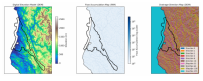


Figure 2: Fundamental Basin Data

Elevation in this region ranges from sea level near the coast to well over 2500 m in the mountainous interior. Flow-accumulation patterns reveal narrow valleys and a pronounced main stem, confirming the basin's steep terrain and efficient drainage networks.

Basin Details

several peak flow events. However, the model tends to overestimate the magnitude of peak flows, particularly during smaller events. There are also instances where the simulated peak discharge occurs slightly earlier than the observed peak, indicating a potential lag in the model's response. The simulation of baseflow appears to be reasonably accurate. Discrepancies in peak magnitude and timing suggest that further calibration is needed to enhance the model's accuracy, especially in capturing peak discharge events.

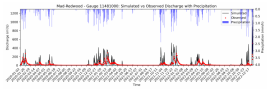


Figure 1: Simulation vs Observation

5. Model Performance Metrics

The following table summarizes the key performance metrics for the CREST model simulation.

Metric	Value
Nash-Sutcliffe Efficiency (NSCE)	-1.162
Kling-Gupta Efficiency (KGE)	-0.116
Correlation Coefficient (r)	0.696
Bias (m^3/s)	10.32
Bias (%)	48.8
Root Mean Square Error (RMSE) (m^3/s)	62.90

The negative NSCE and KGE values indicate that the model's performance is worse than simply using the mean of the observed data as a predictor. While the correlation coefficient shows a moderate positive correlation, the high bias and RMSE values suggest significant discrepancies between the simulated and observed streamflow.

6. CREST Model Parameters

The following table lists the key parameters used in the CREST model simulation.

Variable	Value
Basin Name	Mad-Redwood
Basin Area (km²)	3684.31
Simulation Period	2020-01-01 to 2022-12-31
USGS Gauge ID	11481000
Gauge Coordinates (Lat, Long)	(40.909572, -124.060896)

2. Analysis Sections

2.1 Simulation vs. Observation Comparison

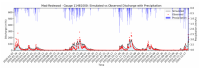


Figure 3: Simulation vs Observation Results

A comparison of observed and simulated streamflow for the period 2020–2022 highlights the model's ability to follow seasonal patterns and capture major runoff events. Precipitation inputs (depicted as blue bars) drive distinct winter peaks, and the observed discharge (red points) exhibits sharp rises during storm events. The simulated discharge (black line) generally aligns with observed spikes, though discrepancies in peak magnitude and timing are noted.

2.2 Model Performance Metrics

Metric	Value
Nash-Sutcliffe Efficiency (NSCE)	0.578
Kling-Gupta Efficiency (KGE)	0.515
Correlation Coefficient	0.788
Bias (m^3/s, % relative)	8.62 (40.8%)
RMSE (m^3/s)	27.80

The NSCE of 0.578 indicates moderate agreement between simulated and observed flows. The KGE of 0.515 provides a balanced measure of correlation, bias, and variability, demonstrating a need for refinement but an overall acceptable model performance. A bias of 8.62 m^3/s suggests the model tends to overestimate flow, while the correlation coefficient (0.788) indicates a strong temporal alignment of events. The RMSE (27.80 m^3/s) captures the residual scatter around observed flows.

Parameter	Value	Description
Water Balance Parameters		
Water capacity ratio (WM)	200.0	Maximum soil water capacity (mm)
Infiltration curve exponent (B)	10.0	Controls water partitioning to runoff
Impervious area ratio (IM)	0.05	Represents urbanized areas
PET adjustment factor (KE)	0.9	Affects potential evapotranspiration
Soil saturated hydraulic conductivity (FC)	75.0	Rate at which water enters soil (mm/hr)
Initial soil water value (IWU)	25.0	Initial soil moisture (mm)
Kinematic Wave (Routing) Parameters		
Drainage threshold (TH)	50.0	Defines river cells based on flow accumulation (km^2)
Interflow speed multiplier (UNDER)	2.0	Accelerates subsurface flow
Interflow reservoir leakage coefficient (LEAKI)	0.7	Increases interflow drainage rate
Initial interflow reservoir value (ISU)	0.0	Initial subsurface water
Channel flow multiplier (ALPHA)	0.5	Affects wave propagation speed in channels ($Q = A^{\cdot}$)
Channel flow exponent (BETA)	0.6	Affects wave propagation speed in channels ($Q = A^{\cdot}$)
Overland flow multiplier (ALPHA0)	0.8	Affects overland flow speed

7. Run Arguments (Basin Details)

While specific run arguments beyond parameters were not provided, the following details are relevant to the simulation setup:

Detail	Value
Basin Area	3684.31 km^2
Simulation Start Date	2020-01-01
Simulation End Date	2022-12-31

2.3 CREST Parameters

Water Balance Parameters

Parameter	Value	Description
WM (mm)	200.0	Max soil water capacity. Higher = more water storage, reducing runoff.
B	10.0	Infiltration curve exponent. Higher = less infiltration, more runoff.
IM	0.1	Impervious area ratio. Higher = larger urban areas, more direct runoff.
KE	0.8	PET adjustment factor. Higher = more evapotranspiration, reducing runoff.
FC (mm/hr)	50.0	Soil saturated hydraulic conductivity. Higher = easier water entry into soil, less runoff.
IWU (mm)	25.0	Initial soil water value. Higher = less available storage, leading to greater runoff initially.

Kinematic Wave (Routing) Parameters

Parameter	Value	Description
TH (km^2)	100.0	Drainage threshold. Higher = fewer cells classified as river channels.
UNDER	1.0	Interflow speed multiplier. Higher = faster subsurface flow.
LEAKI	0.05	Interflow reservoir leakage coefficient. Higher = faster interflow drainage.
ISU (mm)	0.0	Initial subsurface water. Higher = more immediate contribution to flow.
ALPHA	1.5	Channel flow multiplier ($Q = A^{\cdot}$). Higher = slower wave transmission in channels.
BETA	0.6	Channel flow exponent ($Q = A^{\cdot}$). Higher = slower wave propagation.
ALPHA0	1.0	Overland flow multiplier. Higher = slower overland flow.

3. Discussion

3.1 Model Performance Evaluation

The simulation captures seasonal flow variability and timing of major events well. However, the bias and moderate efficiency values suggest potential for improvement in peak flow estimation. Adjustments to the infiltration curve exponent (B) or saturated conductivity (FC) might address the high runoff bias.

Detail	Value
Target USGS Gauge	#11481000
Gauge Location (Lat, Lon)	(40.909572, -124.060896)

8. Conclusion/Discussion

The CREST model simulation for the Mad-Redwood basin, while capturing some aspects of the observed hydrograph, exhibits significant limitations in accurately predicting streamflow, as indicated by the poor performance metrics (NSCE and KGE). The model's tendency to overestimate peak flows and the presence of a timing lag suggest that further calibration is necessary.

Given the negative NSCE and KGE values, a warm-up period is not the primary concern. The fundamental model structure and parameterization require attention. The bias of 48.8% indicates a systematic overestimation of streamflow.

Recommendations:

- **Parameter Calibration:** Conduct a thorough calibration of the CREST model parameters, focusing on parameters that influence peak flow generation and timing, such as the infiltration curve exponent (B), soil saturated hydraulic conductivity (FC), and channel flow parameters (ALPHA, BETA). Consider using optimization algorithms to find parameter sets that minimize the discrepancies between simulated and observed streamflow.
- **Data Quality Assessment:** Evaluate the quality and completeness of the observed precipitation and streamflow data. Erroneous or missing data can negatively impact model performance.
- **Model Structure Evaluation:** Consider refining the model structure to better represent the hydrologic processes within the Mad-Redwood basin. This could involve incorporating additional components, such as groundwater storage, or using a more sophisticated routing scheme.
- **Spatial Resolution:** Investigate the impact of spatial resolution on model performance. Increasing the resolution of the input data (e.g., DEM, land cover) may improve the model's ability to capture the spatial variability of hydrologic processes.
- **Simulation Period:** Extend the simulation period to include a wider range of hydrologic conditions. This will provide a more robust assessment of the model's performance.
- **Evapotranspiration:** Evaluate the accuracy of the potential evapotranspiration (PET) estimates used in the model. Consider using alternative PET methods or calibrating the PET adjustment factor (KE).

By addressing these recommendations, the accuracy and reliability of the CREST model simulation for the Mad-Redwood basin can be significantly improved. Future work should focus on refining the model to better represent the complex hydrologic dynamics of this important coastal watershed.

3.2 Warmup Period Considerations

Although not explicitly analyzed here, a sufficient model warmup may help reduce initial condition errors, especially if the bias were below −90%—a signal of major systematic offset in the early period. This is not the case here, but a 3–6-month warmup could still improve early simulation alignment.

3.3 Recommendations and Next Steps

1. Further calibration focusing on reducing bias, particularly during high-flow events.
2. Evaluation of spatial variation in soil properties or imperviousness for improved runoff response.
3. Examination of seasonal PET adjustment factors (KE) to capture variable evapotranspiration rates throughout the year.
4. Testing longer warmup periods and refined initial soil water (IWU) settings to improve early-season match.

In conclusion, while the model shows moderate skill in replicating observed flows and timing, targeted parameter refinements and thorough calibration may enhance performance, particularly for flood peaks and sustained high flows.

Figure 8. Side-by-side grading example for two hydrological reports generated by different LLM agents. Panel (a) shows the report B5_030.pdf produced by `gemini-2.5-flash`, while panel (b) displays B5_223.pdf from `gpt-o1`. Both were created from the same prompt, "*I want to simulate the streamflow of the Mad–Redwood basin from 2020 to 2022*." The table underneath presents the averaged human + LLM scores on the four-axis rubric. Owing to missing figures, B5_030 lags in *Model Completeness*; its poorer NSE also lowers the *Simulation Results* score. In contrast, B5_223 achieves notably higher marks across all axes, leading to a superior overall quality index.