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`

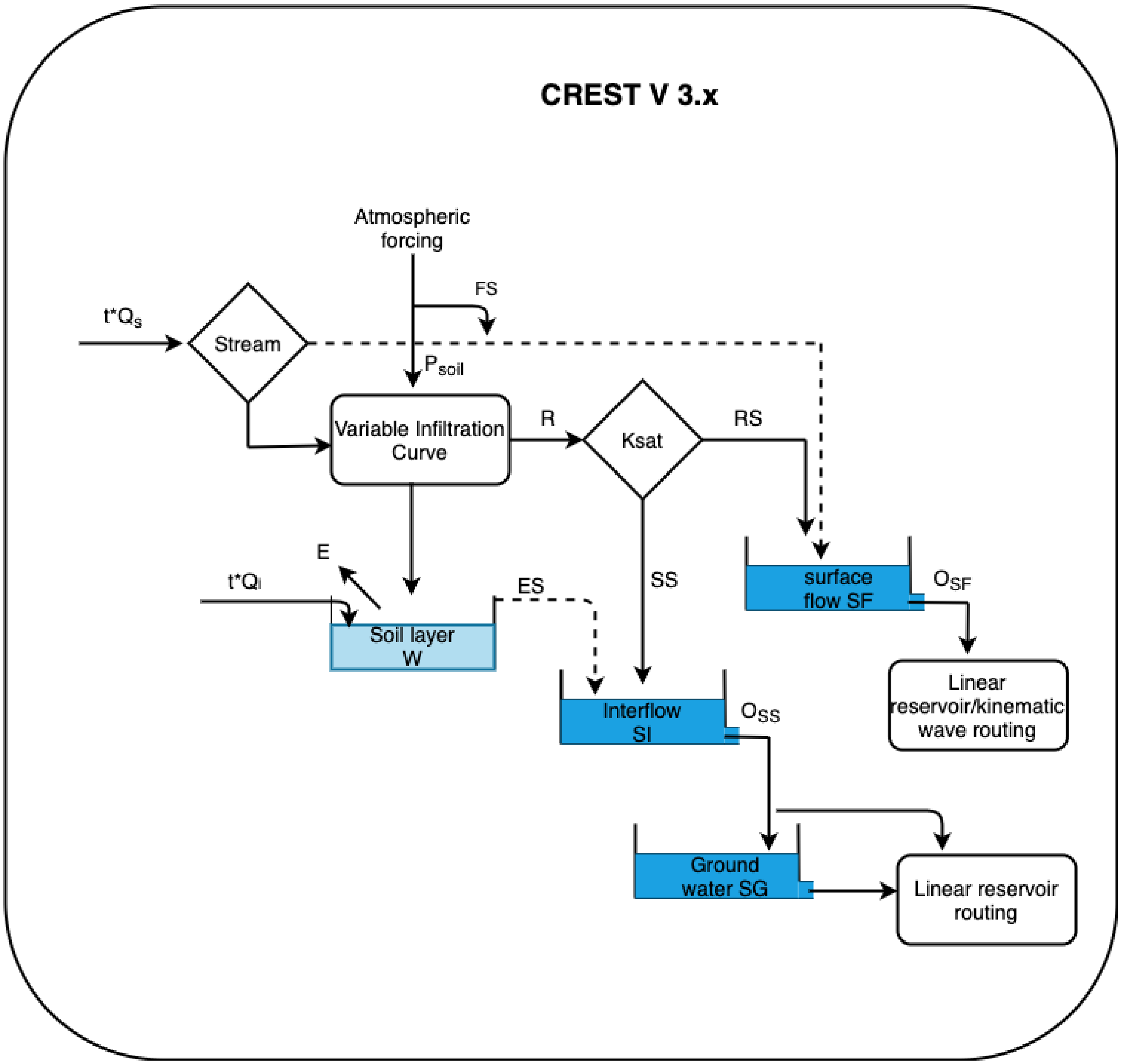

Figure 6. A schematic of the hydrologic processes represented by the latest EF5/CREST model

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)	$BIAS = \frac{1}{T}\sum_{t=1}^{T}\left(Q^t_{\text{sim}} - Q^t_{\text{obs}}\right)$	$(-\infty, \infty)$ (0)
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(Q^t_{\text{sim}} - Q^t_{\text{obs}}\right)^2}$	$[0, \infty)$ (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,

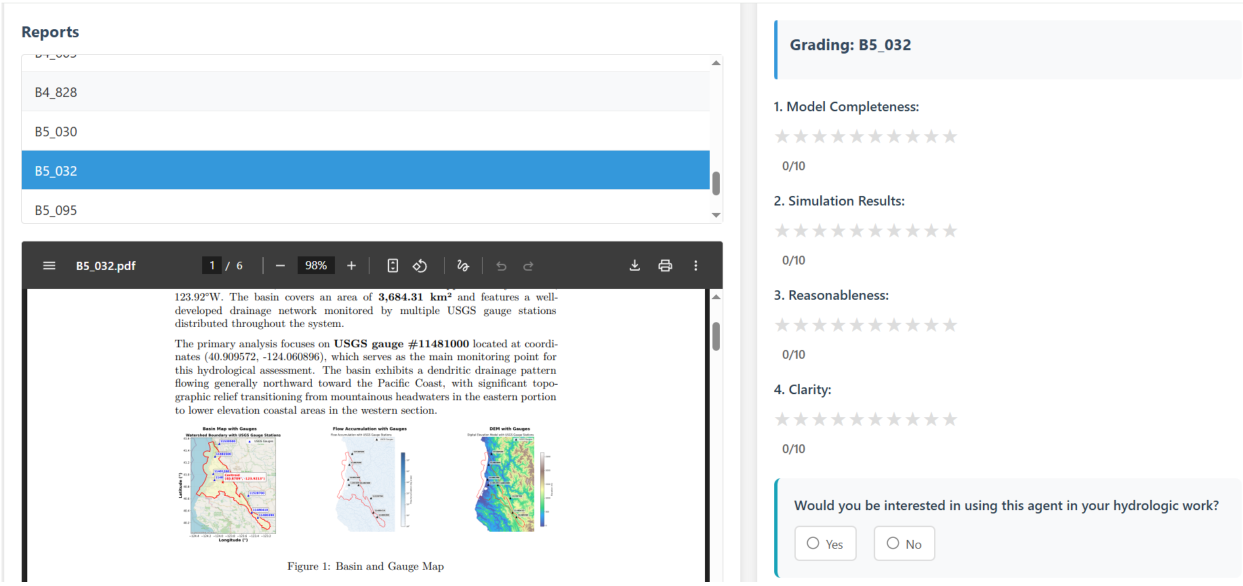

Figure 7. Human-grading interface used in this study. Experts (and an LLM co-evaluator) assign 0–10 star scores on four axes—Model Completeness, Simulation Results, Reasonableness, and Clarity—and record whether they would adopt the agent in professional hydrologic work.

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.

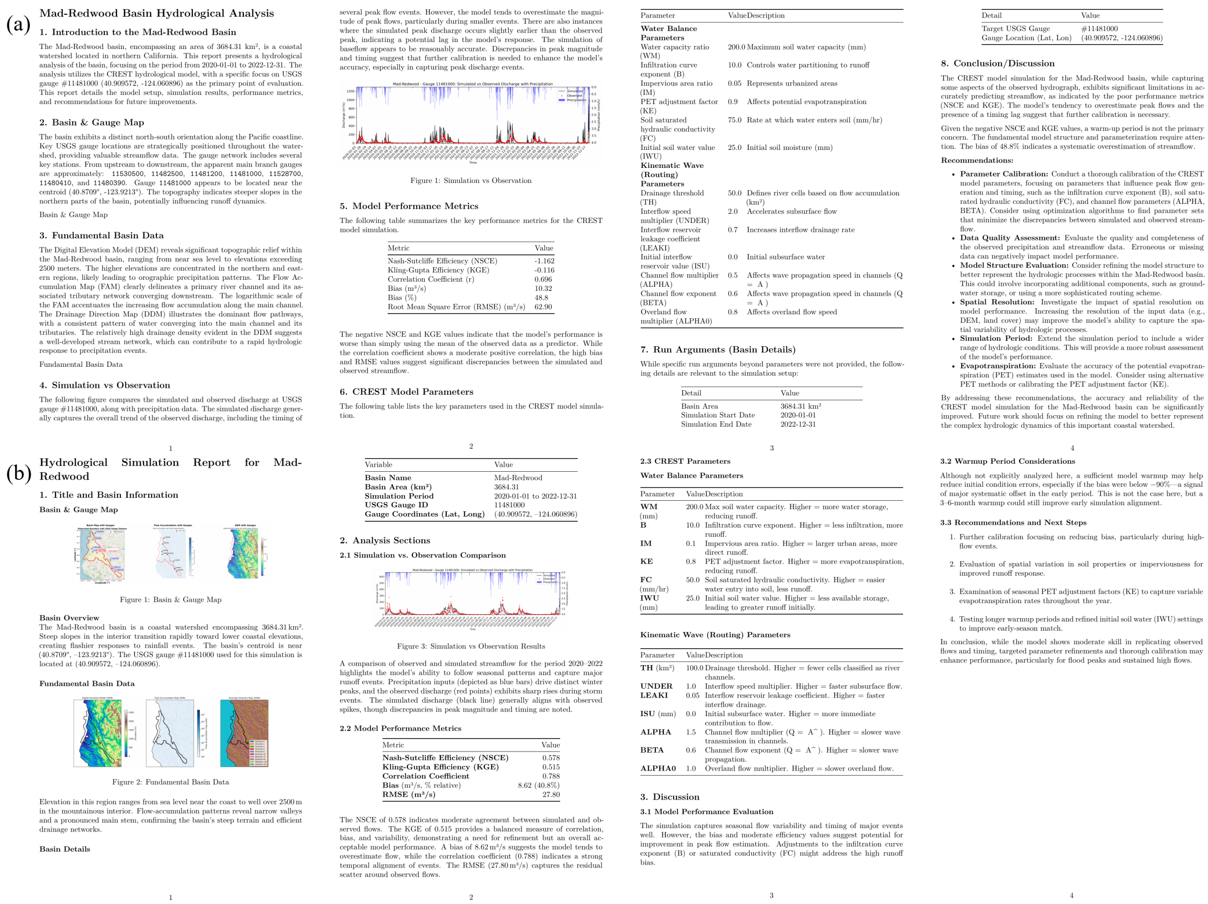

Figure 8. Side-by-side grading example for two hydrological reports generated by different LLM agents. Panel (a) shows the report B5_030.pdf produced by `gemini-2.5-flash`, while panel (b) displays B5_223.pdf from `gpt-o1`. Both were created from the same prompt, "*I want to simulate the streamflow of 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.