

A. Experimental Setup

Implementation framework. All agents are orchestrated with `crewAI` v0.75.0¹, 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 `CONTEXT_PARSER` 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>

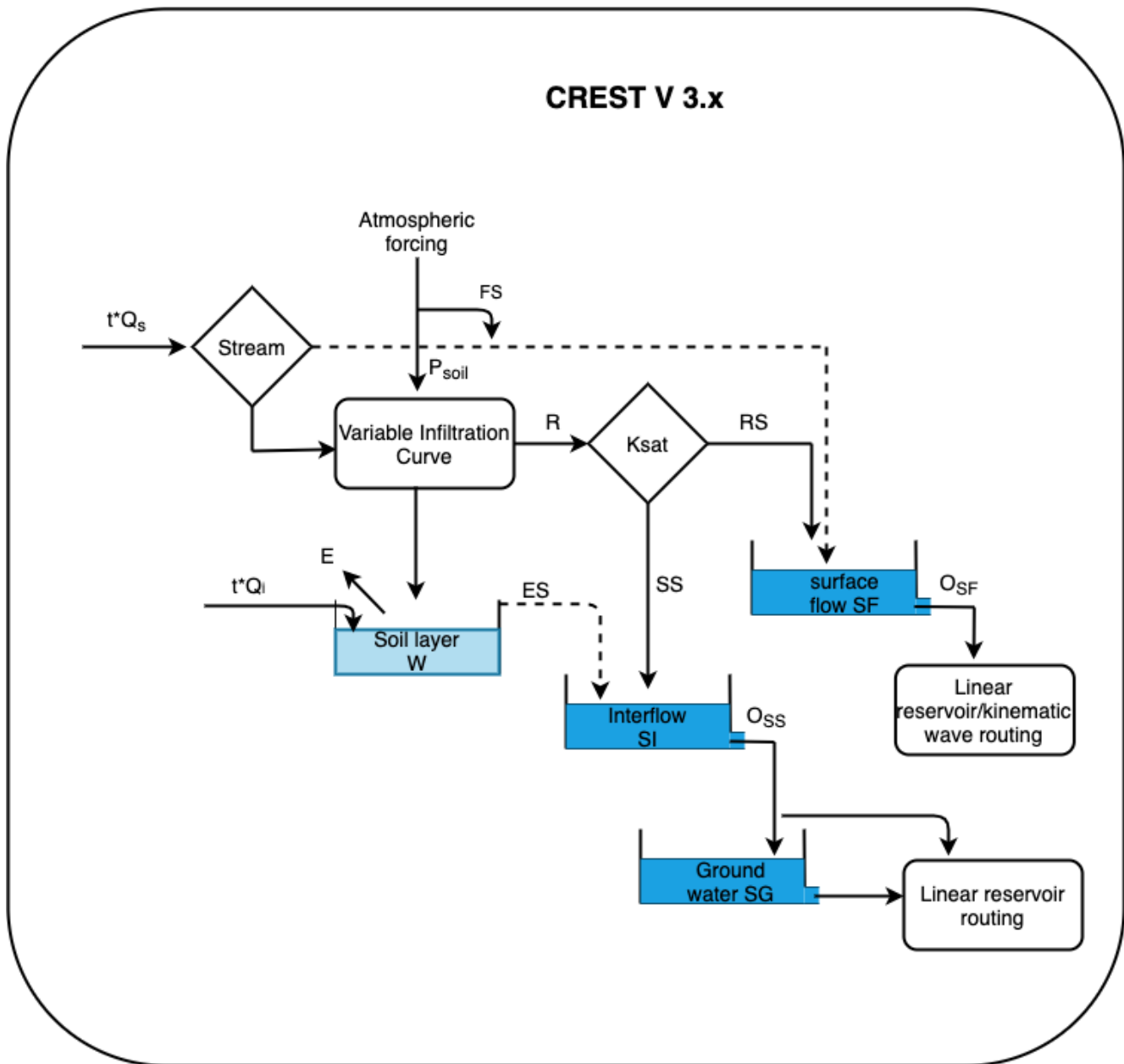


Figure 6. A schematic of the hydrologic processes represented by the latest EF5/CREST model

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 $\alpha 3$ 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 $\Rightarrow$ less direct runoff.
B	Infiltration curve exponent	0.1–20	Steeper curve $\Rightarrow$ more surface runoff.
IM	Fraction of impervious area	0.01–0.50	Larger imperviousness $\Rightarrow$ more runoff.
KE	PET utilisation / evapotranspiration coefficient	0.001–1.0	Higher ET loss $\Rightarrow$ less runoff.
FC	Saturated hydraulic conductivity proxy (mm h ⁻¹)	0–150	Faster infiltration $\Rightarrow$ less runoff.
IWU	Initial soil-water content (mm)	0–25	Wetter initial state $\Rightarrow$ 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 $\Rightarrow$ finer channel network.
UNDER	Interflow velocity multiplier (m s ⁻¹)	0.0001–3.0	Larger velocity $\Rightarrow$ quicker runoff response.
LEAKI	Leakage factor from interflow layer	0.01–1.0	Higher leakage $\Rightarrow$ 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_{obs}^t (Q_{sim}^t) is the observed (simulated) discharge at time step t ; $\bar{Q}_{\text{obs}}$ and $\bar{Q}_{\text{sim}}$ 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)	$NSE = 1 - \frac{\sum_{t=1}^T (Q_{\text{obs}}^t - Q_{\text{sim}}^t)^2}{\sum_{t=1}^T (Q_{\text{obs}}^t - \bar{Q}_{\text{obs}})^2}$	$(-\infty, 1]$ (1)
Relative bias (BIAS)	$BIAS = \frac{1}{T} \sum_{t=1}^T (Q_{\text{sim}}^t - Q_{\text{obs}}^t)$	$(-\infty, \infty)$ (0)
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (Q_{\text{sim}}^t - Q_{\text{obs}}^t)^2}$	$[0, \infty)$ (0)
Correlation coefficient (CC)	$CC = \frac{\sum_{t=1}^T (Q_{\text{sim}}^t - \bar{Q}_{\text{sim}})(Q_{\text{obs}}^t - \bar{Q}_{\text{obs}})}{\sqrt{\sum_{t=1}^T (Q_{\text{sim}}^t - \bar{Q}_{\text{sim}})^2} \sqrt{\sum_{t=1}^T (Q_{\text{obs}}^t - \bar{Q}_{\text{obs}})^2}}$	$[-1, 1]$ (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,

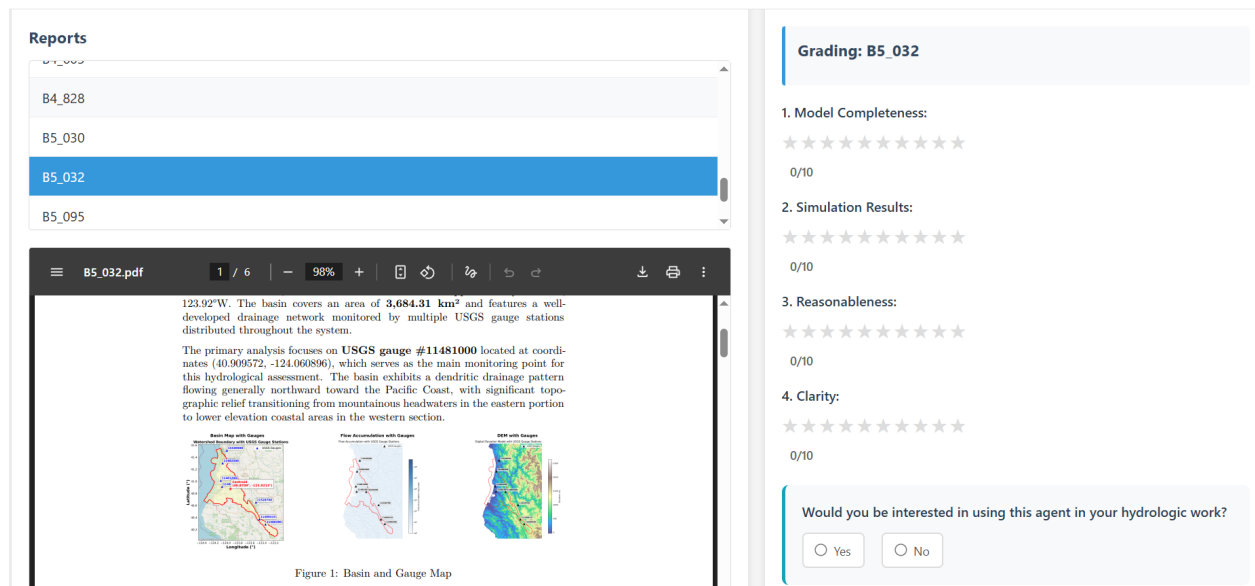


Figure 7. Human-grading interface used in this study. Experts (and an LLM co-evaluator) assign 0–10 star scores on four axes—Model Completeness, Simulation Results, Reasonableness, and Clarity—and record whether they would adopt the agent in professional hydrologic work.

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 (40.90572, -121.06086) 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 segment main branch gauges are approximately: 11535050, 11482500, 11481200, 11481000, 11528700, 11485450, and 11485280. Gauge 11481000 appears to be located near the confluence (40.8709° -121.9213°). 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 highest 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

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.

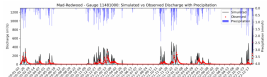


Figure 1: Simulation vs Observation

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)	0.096
Bias (m³/s)	10.82
Bias (%)	8.8%
Root Mean Square Error (RMSE) (m³/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)	(40.90572, -121.06086)

2. Analysis Sections

2.1 Simulation vs Observation Comparison

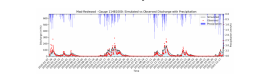


Figure 2: Simulation vs Observation Results

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 water 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 (m³/s, % relative)	8.62 (8.8%)
RMSE (m³/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 m³/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 m³/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 (D)	10.0 Controls water partitioning to runoff
Impermeable 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 (km²)
Interflow speed multiplier (UNDER)	2.0 Accelerates subsurface flow
Interflow storage lag coefficient (LEAK)	0.7 Increases interflow drainage rate
Initial interflow reserved water (ISU)	0.0 Initial subsurface water
Channel flow multiplier (ALPHA)	0.5 Affects wave propagation speed in channels (Q = A)
Channel flow exponent (BETA)	0.6 Affects wave propagation speed in channels (Q = A)
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 km²
Simulation Start Date	2020-01-01
Simulation End Date	2022-12-31

2.3 CREST Parameters

Parameter	Value/Description
WM	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 Impermeable area ratio. Higher = larger urban areas, more direct runoff.
KE	0.8 PET adjustment factor. Higher = more evapotranspiration, reducing runoff.
FC	50.0 Soil saturated hydraulic conductivity. Higher = easier water entry into soil, less runoff.
IWU	25.0 Initial soil water value. Higher = less available storage, leading to greater runoff initially.

Kinematic Wave (Routing) Parameters

Parameter	Value/Description
TH (km²)	100.0 Drainage threshold. Higher = fewer cells classified as river channels.
UNDER	1.0 Interflow speed multiplier. Higher = faster subsurface flow.
LEAK	0.05 Interflow storage lag 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 ^{1.5}). Higher = slower wave propagation.
BETA	0.6 Channel flow exponent (Q = A ^{0.6}). 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 (D) or saturated conductivity (FC) might address the high runoff bias.

Detail	Value
Target USGS Gauge	#11481000
Gauge Location (Lat, Lon)	(40.90572, -121.06086)

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 -50% - 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 responses.
- 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.