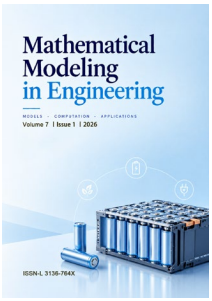
UNIVERSITY OF WEST ATTICA

eISSN: 3136-764X

Mathematical Modeling in Engineering

Vol. 7, No. 2, April, 2026



An Explainable CausalFormer for Multimodal Reservoir Shoreline Extraction

Marwan Al-Darmaki[a,*], Maktoum Al-Junaibi[a], Laith Al-Shahrani[a]

[a] *School of Computing, Rochester Institute of Technology Dubai, Dubai, PO Box 341055, United Arab Emirates*

Received 04 April 2026; Revised 22 June 2026; Accepted 26 June 2026; Published 30 June 2026

Abstract

Reservoir shoreline extraction is complicated by water-level fluctuation, cloud-water confusion, exposed sediment, and asynchronous sensor evidence. This study introduces CausalFormer, an explainable spatiotemporal Transformer that separates shoreline-relevant information from acquisition-specific nuisance variation through matched counterfactual interventions. Six optical bands, dual-polarization synthetic-aperture-radar backscatter, terrain derivatives, and daily reservoir levels are encoded as aligned multiscale tokens. A causal gate estimates modality relevance under cloud, speckle, and stage interventions, while region, signed-distance contour, and evidential uncertainty heads jointly produce a shoreline and confidence layer. Evaluation used 4,860 patches from six reservoirs, including 960 test patches distributed across dry, normal, and flood stages. CausalFormer obtained 91.8% mean intersection-over-union, 89.6% boundary F1, a 1.74 m shoreline-displacement error, and an expected calibration error of 0.031. Relative to the strongest multimodal baseline, these results correspond to gains of 3.7 and 4.4 percentage points in region and boundary accuracy and a 0.83 m reduction in displacement. Five-seed evaluation and paired reservoir-date bootstrap analysis were used to separate model gains from initialization and spatial dependence. Under 40% cloud occlusion, mean intersection-over-union remained 86.9%, and uncertainty-based screening raised accepted-segment boundary F1 while directing ambiguous arcs to review. The vector workflow retains shoreline geometry, confidence, intervention provenance, and review status for subsequent quality control. The results show that matched intervention evidence can support accurate, calibrated, and auditable reservoir shoreline products when its causal interpretation is restricted to the stated nuisance assumptions.

Keywords: Reservoir Shoreline Extraction, Intervention-Based Explanation, Multimodal Remote Sensing, Boundary Uncertainty, Counterfactual Attribution

1. Introduction

Reservoir shorelines are dynamic water-land interfaces whose mapped positions respond to both hydrological change and observation conditions [1]. Regulated releases, inflow pulses, evaporation, sediment redistribution, and wind-driven motion can alter the local boundary within short intervals and thereby change the acquisition context [2]. Repeated satellite observations provide the spatial coverage needed for reservoir monitoring, but shoreline delineation remains sensitive to cloud, shadow, turbidity, mixed pixels, and exposed sediment [3]. Optical water indices are efficient when spectral contrast is clear, although performance deteriorates when water, wet soil, and dark artificial surfaces have similar responses [4]. Optical-synthetic-aperture-radar fusion can reduce gaps and add structural evidence under cloud, yet the modalities differ in geometry, noise, and acquisition timing [5]. A dependable engineering system must therefore

* Corresponding author: marwan.darm@ritdubai.ae

https://doi.org/10.68406/mme.2026.vol7iss2nm10:123-137

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

determine not only where the shoreline lies, but also which evidence supports each boundary segment and when that evidence is too weak for automatic acceptance.

Deep convolutional networks improved reservoir mapping by learning contextual texture, but their local receptive fields are less effective in the presence of islands, bridges, or long cast shadows in the water body [6]. Vision Transformers have also used some form of global token interaction to address this problem and have achieved good results in dense remote sensing prediction [7]. Multi-temporal architecture can be employed to add seasonal information and thus differentiate between a permanent basin and an isolated dark surface [8]. However, the attention weights do not specify the reason for the alteration in judgement, and the model may be using acquisition artifacts that are associated with water in the training set. Therefore, Explainable Learning has been used to find obvious bands or regions, but these post-hoc maps are often unstable and do not necessarily show a causal relationship [9]. Zhang et al. [10] showed that causal representation learning can separate factors that remain predictive after intervention from nuisance variations that should not determine the model output. Reservoir imagery can be used to obtain data on the hydrological conditions, spectral water-response characteristics, topographic suitability, cloud cover and sensor-specific noise. The current engineering problem is how to use these sources without considering all the statistical correlations as evidence for a shoreline. Therefore, the stable solution should provide the global basin context and local boundary geometry and conditions for the trust at each input source, and these two functions need to be separated in the diagnosis.

CausalFormer, an explainable spatiotemporal Transformer, has been proposed to address this problem of reservoir shoreline extraction. The model is given synchronized optical, synthetic-aperture-radar, terrain and water-level data. A shared encoder produces scale-aware tokens, and a causal gate selects which modifications of modality are consistent with shoreline behavior under controlled feature alterations. The gate is not only for showing attention; it can also be employed to experiment with nuisance variables or to suppress individual modalities and learn a counterfactual response of boundary logits. A boundary decoder integrates the invariant representation and high-resolution spatial features to output shoreline probability and calibrated uncertainty. Thus, the two typical failure modes of confident segmentation of shadow or wet sediment as water, and oversmoothed boundaries at the waterline of narrow structures, will be mitigated. It can also provide explanations supported by hydrological and spectral evidence to solve the problem of visually attractive but unsubstantiated heatmaps. The scope of the causal control is limited to variables for which a reasonable intervention can be designed. The method does not claim to recover the whole hydrological causal system from images by a neural network. It will have the following engineering functions: The prediction should be stable under different lighting conditions, in the presence of speckle or variations in acquisition time, without affecting the actual position of the water, and it needs to reliably respond to a real displacement observed by the measuring stage. A small Area will be selected for the study of the reasons for failure at the same spatial scale as the built coastline.

Based on this engineering problem, a series of steps will be taken in this study to collect data and build a causal graph. The six parts of the study are designed to correspond with the specific Design Choices or Experiments mentioned in the text. Section 2 introduces reservoir shoreline extraction and the application of causal and explainable learning in remote sensing, and studies the discrepancy between visually plausible attribution and intervention-based evidence. Section 3 introduces the CausalFormer architecture and outlines the components of multimodal tokenization, hierarchical spatio-temporal encoding, intervention-aware fusion, boundary refinement, and uncertainty decoding. Section 4 is the learning objective, the corresponding counterfactual construction, explanation protocol, quality-control rule and operational inference procedure. Section 5 reports the experimental design and investigates the accuracy of segmentation, boundary geometry, calibration, explanatory fidelity, corruption robustness, ablation behavior, selective review and computational cost for six reservoirs. The indices of the areas are combined with shoreline displacement and confidence behavior deliberately, as none of them are operating reliability alone. Section 6 shows the main results and the defects in registration, intervention coverage and spatial resolution, as well as future research directions for physically grounded and transferable shoreline monitoring. Indicators of the performance of reservoir operators will be simple to understand, including dimensions of geometric displacement, confidence calibration, results under controlled corruption, and the quantity of shoreline samples requiring reassessment. Therefore, this paper includes the explanations in the product specifications rather than adding them as extra visualizations after model training.

2. Related Work

2.1. Reservoir Shoreline Extraction

Early shoreline extraction relied on spectral indices followed by global or locally adaptive thresholding. These methods are inexpensive and transparent, but fixed thresholds become unstable under turbidity, bottom reflectance, adjacency effects, and seasonal illumination [11]. Object-based methods group spectrally similar pixels and suppress isolated noise with shape and neighborhood descriptors [12]. Convolutional encoder-decoder networks improve contextual discrimination and boundary recovery, although repeated downsampling can weaken narrow tributaries and small islands [13]. Boundary-aware cross-scale designs restore part of this detail by combining shallow geometry with

deeper semantics [14]. These developments improve water segmentation, but they do not by themselves determine whether a shoreline decision is supported by stable hydrological evidence or by a transient acquisition artifact.

Transformer-based segmentation can achieve long-range dependency modeling and flexibly combine temporal and multi-source information. Windowed Attention reduces the computational cost, and hierarchical token pyramids preserve information at different map scales. To connect the various components of a reservoir and recognize a basin as a whole, global context is employed. Multi-temporal attention can also compare the current observation with a recent hydrological trajectory. Standard Self-Attention only optimizes predictive association. If the cloud coincides with a particular month of acquisition or a sensor processing artifact, the model may have learned a shortcut from the training data. The generated mask performs well on a random test set but poorly on correlated data. Therefore, the architecture will investigate the impact of intervention on temporal and multi-modal features, and boundary accuracy will be maintained as an explicit optimization goal. A second is the lack of time-coincidence. Simple concatenation assumes that the same image coordinate corresponds to the same physical phenomenon at all times; however, parallax, registration errors, changes in vegetation and exposed soil can alter the local appearance. Recurrent models condense sequences and fail to detect a short flood pulse. Token-level temporal attention is relatively loose; to determine whether it is an observation gap or a gradual change, explicit time and stage information need to be added. Therefore, the relative acquisition interval and the difference in water levels are used as conditioning factors, and the dates are not uniformly spaced frames.

2.2. Explainable and Causal Remote Sensing

Many explanatory remote-sensing models use gradient saliency, activation maps, occlusion or attention rollout. Gradient methods are relatively simple but are also very sensitive to changes in the input [15]. Occlusion is relatively simple to understand; it is the result of removing the evidence, and naive pixel masking generates samples outside the image distribution. Attention rollout shows the flow of information in a Transformer but cannot, by itself, determine whether a cause-and-effect relationship exists [16]. Add attributes of the data, such as turbidity or cloud cover, to the concept-based explanation if high-quality concept labels are available, and increase the semantic richness. Shoreline mapping is best used to present the problem that needs to be solved, and therefore the extraction boundary should be supported by spectral absorption, radar smoothness, terrain or observed changes in water level. Explanation Evaluation is also necessary. It is possible to obtain an aesthetically pleasing Saliency map, but we do not know what will be altered in the modified region. The indices are generally the deletion rate, insertion rate, localization rate, stability and agreement with experts. Shoreline extraction is prone to problems of deletion alone; otherwise, the water pixels would be replaced by zero. Match the interventions in areas with similar hydrology for a more reliable basis of comparison, and determine whether the explanation predicts contour movement or only identifies a water area.

Causal learning adds invariance and intervention objectives to reduce the impact of unstable correlations. Structural Causal Models present the causal relationships among observed variables, latent states and predictions [17]. The whole causal graph is rarely known in terms of imagery, so in practice, partial assumptions are often made to treat the environment as a domain, disturb known nuisance factors, and promote the retention of labels after these modifications. Counterfactual prediction finds out what the output would be in different situations by changing only some factors. Recently, some studies have added the goal of uncertainty estimation to identify samples that do not conform to the assumed mechanism [18]. Most of the existing applications are either classification or general land-cover segmentation. Reservoir shorelines are more likely to be objects of space reasoning, and causality should not be imposed by restricting the actual boundary shift. CausalFormer adds intervention constraints at the token and boundary levels, and for each explanation, it is connected to an observable change in the shoreline logit or contour position. Partial Causal Models are also relatively convenient compromises between pure association and an ideal complete physical graph. The known acquisition variables can be shown explicitly, interventions are limited to observed support, and invariance is only imposed in areas where the label mechanism is expected to be unchanged. The remaining uncertainty is not hidden; rather, it has been used as the result of selective observation. Therefore, they are more suitable for regulated reservoirs; although the stage records provide an external variable with physical meaning, they cannot uniquely determine the position of the local shoreline due to bank geometry and short-term wind effects.

3. CausalFormer Architecture

3.1. Multisource Token Construction

The four synchronous sources of construction for the input sequence are: six optical reflectance bands, dual-polarization radar backscatter, slope and relative elevation, and daily reservoir level. Resample the terrain-corrected optical and radar images to 10m and normalize them according to the scene statistics. A cloud probability channel is added as a nuisance indicator but is not used for hard masking; thin cloud and cloud-edge pixels may still contain recoverable spatial information. For each target date, if available, the two closest dates in the 45-day window are selected. Convert the stage values into the difference from the target date and enable differentiation of expansion and contraction

through the temporal encoder. A shallow convolutional stem is used for each raster source prior to non-overlapping tokenization. The components of Position Encoding are planar coordinates, acquisition interval and relative elevation. The resulting representation has kept the local boundary texture and the temporal order required for hydraulic feasibility analysis of the observed shoreline displacement.

Patch Sampling is near the coast and not in the reservoir Area. Half of the training patches are centered on the reference contour, one-quarter cover difficult non-water surfaces such as shadows and wet sediments, and one-quarter preserve the interior water and surrounding land context. Missing neighboring dates are represented by a learned availability token and are never replaced with duplicated target imagery. Radar values are converted to decibels before normalization, and elevation is expressed relative to the median water-adjacent terrain in each patch. These decisions reduce scene-to-scene scale variation while preserving quantities that carry physical meaning. The resulting multisource token sequence is defined by the following equation.

$$\boldsymbol{T}_t = Concat\left(E_o(\boldsymbol{X}_t^o), E_r(\boldsymbol{X}_t^r), E_d(\boldsymbol{X}^d), E_h(\Delta h_t)\right) + \boldsymbol{P}_t \tag{1}$$

The concatenation keeps each sensing source identifiable before attention is applied. Optical, radar, terrain, and stage encoders therefore contribute tokens in a common space without erasing their origin. The positional term carries spatial and temporal context, while an availability mask prevents a missing modality from being interpreted as a measured zero value.

3.2. Spatiotemporal Encoding and Causal Fusion

A hierarchical Transformer processes the token sequence at four scales. Shifted local windows capture bank texture efficiently, while a small set of reservoir tokens exchanges information globally across windows and dates. The reservoir tokens carry stage change and scene-quality metadata, but they do not directly predict the mask. Instead, their interactions modulate spatial tokens through cross-attention. This separation prevents a water-level record from overwriting local image evidence when gauge timing and satellite acquisition are imperfect. Feature maps from the first two scales are retained for the boundary decoder, and deeper maps represent basin-wide context.

Figure 1 presents the complete data path from synchronized observations through token encoding, causal fusion, and the three decoder outputs for region, contour, and uncertainty. The figure is intentionally specified as a single flow diagram because its role is to explain information ownership and the intervention path rather than to report quantitative evidence. Attention is masked across unavailable modalities, and the mask is propagated to the causal gate so that absence is not interpreted as weak evidence. At each scale, reservoir tokens are updated from spatial summaries and then returned to the image tokens through a residual connection.

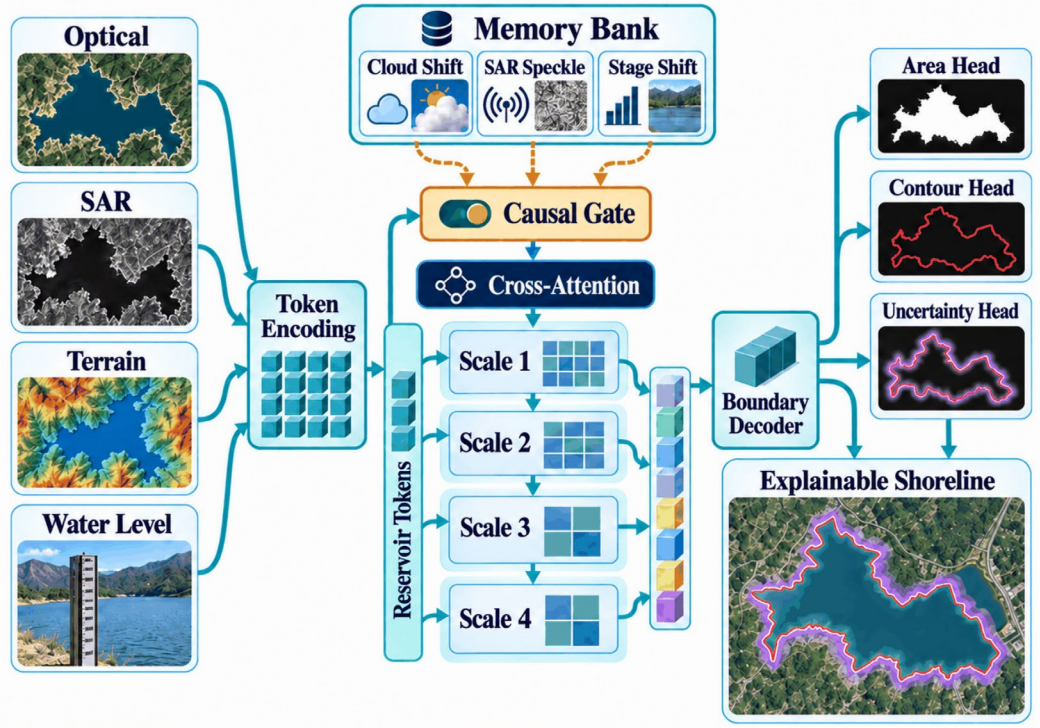


Figure 1. CausalFormer architecture for synchronized optical, radar, terrain, and water-level encoding, intervention-aware causal fusion, boundary decoding, and uncertainty prediction

The deepest stage uses dilated windows to cover long shoreline arcs without quadratic global attention. Layer normalization is applied separately by modality before fusion; otherwise, the higher variance of radar tokens can dominate early training. The decoder receives both fused tokens and modality-specific residuals, allowing a local edge to survive even when its basin-level causal score is modest. The next equation describes the hierarchical update applied at each encoding scale.

$$\boldsymbol{H}_l = STB_l(\boldsymbol{H}_{l-1}, \boldsymbol{q}_t, \boldsymbol{M}_l), l = 1, \dots, 4 \tag{2}$$

This update separates repeated local encoding from the reservoir-level context supplied by the query and modality mask. Applying the same form across four scales lets shallow layers retain shoreline texture while deeper layers enlarge the effective basin context. The masked update also keeps computational cost bounded when dates or modalities are unavailable.

The causal gate operates on modality-specific token groups. For each group, an intervention generator substitutes a nuisance-adjusted sample from another acquisition of the same reservoir at a similar water stage. Optical interventions vary illumination and cloud conditions, radar interventions vary speckle, and temporal interventions alter the acquisition window. The resulting gate value is interpreted as counterfactual relevance under these matched substitutions rather than as unrestricted causal identification. Stable representations are encouraged only where the intervention design indicates that the shoreline label should remain unchanged; unobserved physical factors outside this design may still influence the prediction.

Matched replacements are selected with calipers on stage difference, season, and reservoir identity. A replacement is rejected when it would imply a different shoreline label inside the stable set used by the invariance loss. This rule prevents the model from being penalized for responding to a genuine hydrological change. The intervention generator records its selection distance, and samples near the caliper limit receive a smaller invariance weight. During deployment, the same mechanism uses a small reservoir-specific memory bank; if no credible match exists, the explanation is marked unsupported rather than extrapolated. This intervention-aware aggregation is summarized in the equation below.

$$g_m = \sigma\left(\boldsymbol{W}_g\left[\boldsymbol{z}_m; \boldsymbol{z}_{do(m)}; \boldsymbol{q}_t\right]\right), \boldsymbol{Z} = \sum_m g_m \boldsymbol{z}_m \tag{3}$$

The gate is evaluated separately for each modality, so its value represents counterfactual relevance rather than raw feature magnitude. Summation produces one fused representation, but the individual gate values remain available for explanation. A source can consequently receive a low weight in one shoreline segment and a high weight elsewhere within the same scene.

3.3. Boundary and Uncertainty Decoding

The decoder combines invariant deep tokens with high-resolution skip features through gated cross-scale attention. The area head estimates water probability, the contour head predicts a signed distance to the shoreline within a 12-pixel refinement band, and the evidential head estimates support for water and non-water classes. Region probabilities preserve topology, the distance field sharpens narrow inlets and dam-adjacent edges, and uncertainty attenuates contour correction where evidence is weak. This arrangement makes uncertainty an active control on boundary refinement rather than a post hoc display layer.

Topology is protected through connected-component filtering conditioned on the known basin footprint, but no hard constraint forces a single water polygon because islands and divided arms are legitimate. The signed-distance target is truncated outside the refinement band, which concentrates capacity on geometry and avoids large gradients from distant interior pixels. Deep supervision is applied at two decoder scales during training and removed at inference. The final contour is extracted at the zero crossing of the refined distance field, then mapped back through the registration transform so that displacement can be measured in ground coordinates.

4. Learning and Operational Inference

4.1. Intervention Aware Optimization

Training uses stratified mini-batches that balance reservoir, water-level regime, and cloud fraction. Augmentation includes radiometric scaling, radar multiplicative noise, boundary-preserving rotation, and physically restricted stage perturbation. Interventions are generated after geometric augmentation so that original and counterfactual samples remain spatially aligned. The first training phase optimizes region and contour prediction without causal regularization for 30 epochs. The second phase activates the modality interventions and gradually increases their loss weight over 15 epochs. This schedule prevents an initially random gate from suppressing informative channels.

AdamW optimization uses a peak learning rate of 2×10^{-4}, cosine decay, weight decay of 0.01, and gradient clipping at 1.0. Early stopping monitors the mean of intersection-over-union and boundary F 1 on reservoirs not used for parameter updates. Reservoir-level splits are created before patch extraction to prevent neighboring crops from crossing training and test partitions. Mini-batches contain at least three reservoirs and two hydrological regimes, which gives the invariance objective meaningful environmental variation. Mixed precision is used for the encoder, whereas distance and uncertainty losses remain in full precision to avoid numerical instability near zero evidence. Model selection uses a fixed validation score and a minimum training duration, preventing an unusually easy early epoch from determining the final checkpoint. The composite training criterion is stated in the following equation.

$$\mathcal{L} = \lambda_r \mathcal{L}_{FT} + \lambda_b \mathcal{L}_{SD} + \lambda_c \mathcal{L}_{inv} + \lambda_s \|\boldsymbol{g}\|_1 + \lambda_u \mathcal{L}_{evi} \tag{4}$$

The objective assigns distinct roles to region overlap, signed-distance geometry, intervention consistency, gate sparsity, and evidential calibration. Their coefficients regulate competing gradients rather than serving as interchangeable penalties. In particular, the contour and invariance terms protect boundary geometry and causal stability from being overwhelmed by the much larger number of interior region pixels.

The objective balances region overlaps, contour geometry, causal invariance, and calibration. Focal Tversky loss emphasizes difficult shoreline pixels, while a signed-distance loss penalizes local contour displacement. The causal term compares original and intervened predictions only where the assumed mechanism indicates that the shoreline should remain fixed. A gate sparsity term discourages diffuse explanations, and an evidential regularizer prevents high confidence when class support is weak. Loss coefficients are selected on the validation set and then frozen for every reservoir split. Because the losses act on different targets, their gradients are normalized before aggregation; otherwise, the dense region term can dominate the narrow contour band. The resulting optimization retains region completeness while assigning enough capacity to boundary geometry and explanation stability.

Coefficient sensitivity is examined over a compact grid rather than tuned independently for every site. The causal weight is considered acceptable when intervention inconsistency decreases without a material loss in clean-scene overlap. The contour weight is bounded because excessive emphasis can produce thin, disconnected water regions. Calibration regularization is activated only after region predictions become informative. This ordering reflects the fact that confidence is meaningful only when the underlying class evidence has begun to separate. Class-frequency terms are computed per reservoir to avoid allowing one large site to determine the global loss balance. Within that objective, causal invariance is evaluated through the equation shown below.

$$\mathcal{L}_{inv} = \frac{1}{|\Omega_s|} \sum_{i \in \Omega_s} \left\| p_i(\boldsymbol{x}) - p_i\big(do(x_m = x'_m)\big) \right\|_1 \tag{5}$$

The stable set restricts comparison to pixels whose shoreline label should not change under the selected intervention. Normalization by its size prevents large stable regions from dominating a batch. The absolute probability difference then penalizes only unsupported response changes, allowing the model to react when water-level evidence implies a genuine shoreline displacement.

4.2. Counterfactual Explanation

Counterfactual explanations are generated during inference without retraining. Each modality is replaced by a matched intervention in turn, and the resulting change in shoreline logit is projected onto the predicted contour. A positive optical contribution indicates that spectral evidence increases water support, whereas a positive stage contribution indicates agreement between temporal geometry and the measured level change. Radar and terrain contributions are interpreted using the same response-based rule. Because these quantities depend on the intervention library and matching assumptions, they are reported as intervention-based relevance scores rather than proof of a complete physical causal mechanism.

Segment-level attribution is calculated after tracing the contour into fixed-length arcs. Contributions are averaged with uncertainty weights, and conflicting pixel responses are preserved as a dispersion statistic. An explanation is labeled stable when repeated matched interventions retain the same dominant cause and keep the predicted displacement within one pixel. The procedure yields a compact audit record without discarding the pixel-level values required for detailed review. It also reveals boundary sections whose average attribution looks reasonable but contains alternating optical and radar responses, a pattern often associated with vegetation or registration error. Predictive uncertainty and distance attenuation are linked by the final equation in this section.

$$u_i = \frac{2}{\alpha_i^{water} + \alpha_i^{land}}, \hat{d}_i = d_i(1 - u_i) \tag{6}$$

The reciprocal evidence term increases uncertainty when neither class receives sufficient support. Multiplying the distance estimate by its certainty complement reduces aggressive contour movement in ambiguous areas while preserving refinement where evidence is concentrated. The relation therefore converts uncertainty from a reporting variable into a direct control on boundary correction.

4.3. Operational Workflow

Figure 2 specifies the operational workflow. A quality-control block first checks registration, missing bands, gauge timing, and cloud fraction. Valid observations pass through standard inference and counterfactual intervention branches. The resulting mask is vectorized, simplified at a tolerance linked to pixel size, and screened by uncertainty and topographic consistency. Segments above the uncertainty threshold are marked for analyst review rather than silently discarded. The deployment path therefore returns a shoreline polygon, a confidence layer, and a structured explanation record.

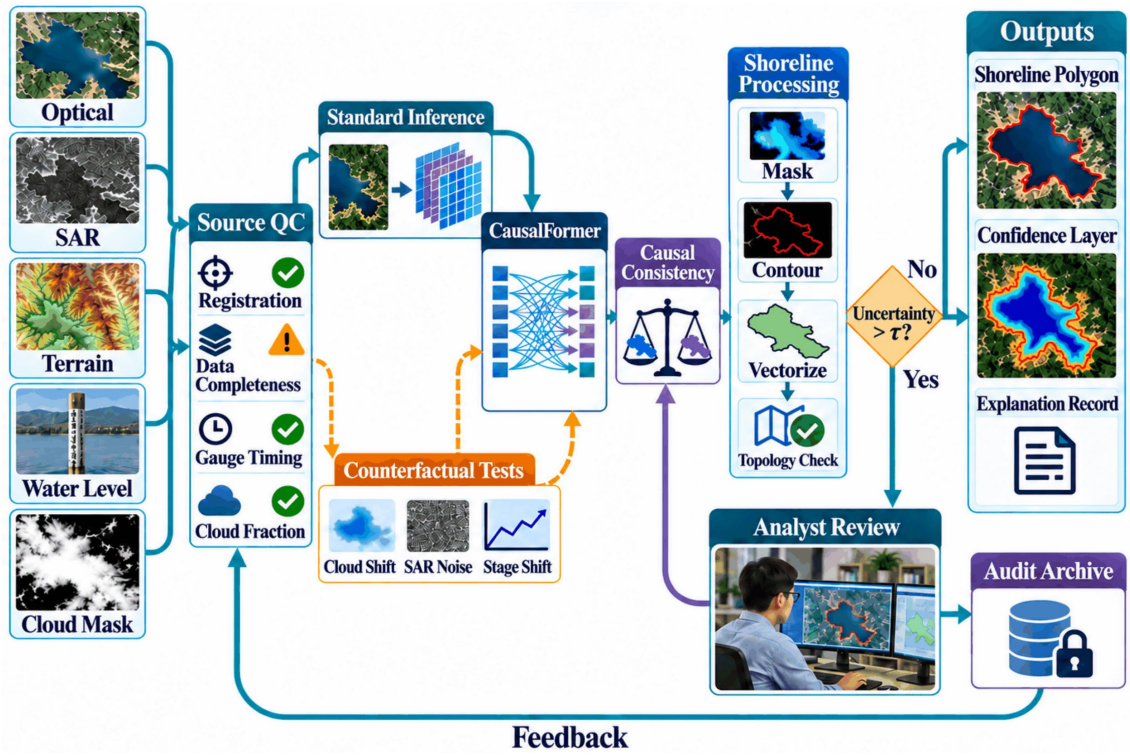


Figure 2. Operational inference workflow from source quality control and counterfactual intervention to vector shoreline, confidence layer, explanation record, and analyst review

This output design is compatible with periodic reservoir surveys and creates an auditable link between source evidence and every accepted boundary segment. Vector quality checks include self-intersection removal, minimum island area, and comparison with the active-storage elevation range. They are diagnostic rather than corrective when evidence is uncertain. A flagged segment keeps its original geometry, associated raster probabilities, and intervention responses so that an analyst can understand why it was questioned. Processing metadata record model version, input timestamps, normalization statistics, and the threshold used for review. The design allows successive surveys to be compared without losing the provenance of a changed shoreline.

5. Experimental Evaluation and Analytical Evidence

5.1. Comparative Accuracy and Boundary Geometry

The benchmark contains 4,860 patches from six reservoirs, with 3,120 used for training, 780 for validation, and 960 for testing. Test scenes include 342 dry-stage, 318 normal-stage, and 300 flood-stage patches; 27% contain more than 20% cloud or shadow. Labels were delineated at 1:2,000 display scale and checked against same-week gauge records. The evaluation follows a boundary-tolerant protocol in which contour matches are accepted within two 10 m pixels [19]. Mean intersection-over-union, boundary F1, and symmetric shoreline displacement are reported because region overlap alone can hide local contour errors. Table 1 compares CausalFormer with thresholding, a convolutional encoder-decoder, a boundary-aware network, a temporal Transformer, and a multimodal Transformer under identical splits. CausalFormer achieves 91.8% mean intersection-over-union, 89.6% boundary F1, and 1.74 m displacement error. The strongest baseline reaches 88.1%, 85.2%, and 2.57 m. The combined use of overlap and shoreline displacement follows

established geometric evaluation practice for boundary-sensitive segmentation [20]. Figure 3(a) shows that CausalFormer leads on all six reservoirs, with mean intersection-over-union ranging from 89.9% to 93.4%; the narrow 3.5-point range indicates that the overall gain is not dominated by one favorable site, while the lower result at R6 is consistent with its bright sediment and narrow tributaries.

Table 1. Comparative shoreline extraction performance on the six-reservoir test set

Method	mIoU (%)	Boundary F1 (%)	Displacement (*m*)	ECE
Adaptive NDWI	73.6	68.9	5.82	0.112
U-Net	84.7	80.6	3.42	0.091
BoundaryNet	86.2	83.7	2.88	0.083
Temporal Transformer	86.9	84.1	2.71	0.079
Multimodal Transformer	88.1	85.2	2.57	0.074
CausalFormer Compact	90.6	88.3	1.98	0.039
CausalFormer	91.8	89.6	1.74	0.031

Figure 3(b) gives median boundary F1 values of 88.7%, 90.2%, and 89.5% for dry, normal, and flood stages; the larger flood-stage interquartile range reveals that inundated vegetation and rapidly changing edges increase case-to-case variability even when the median remains stable. Figure 3(c) tracks monthly accuracy and retains at least 89.6% intersection-over-union during the turbid July-August interval, followed by recovery above 91% as suspended sediment declines; the limited seasonal trough supports the intended separation of stable water evidence from appearance change. Calibration is computed with 15 equal-mass confidence bins to avoid domination by densely populated confidence ranges [21].

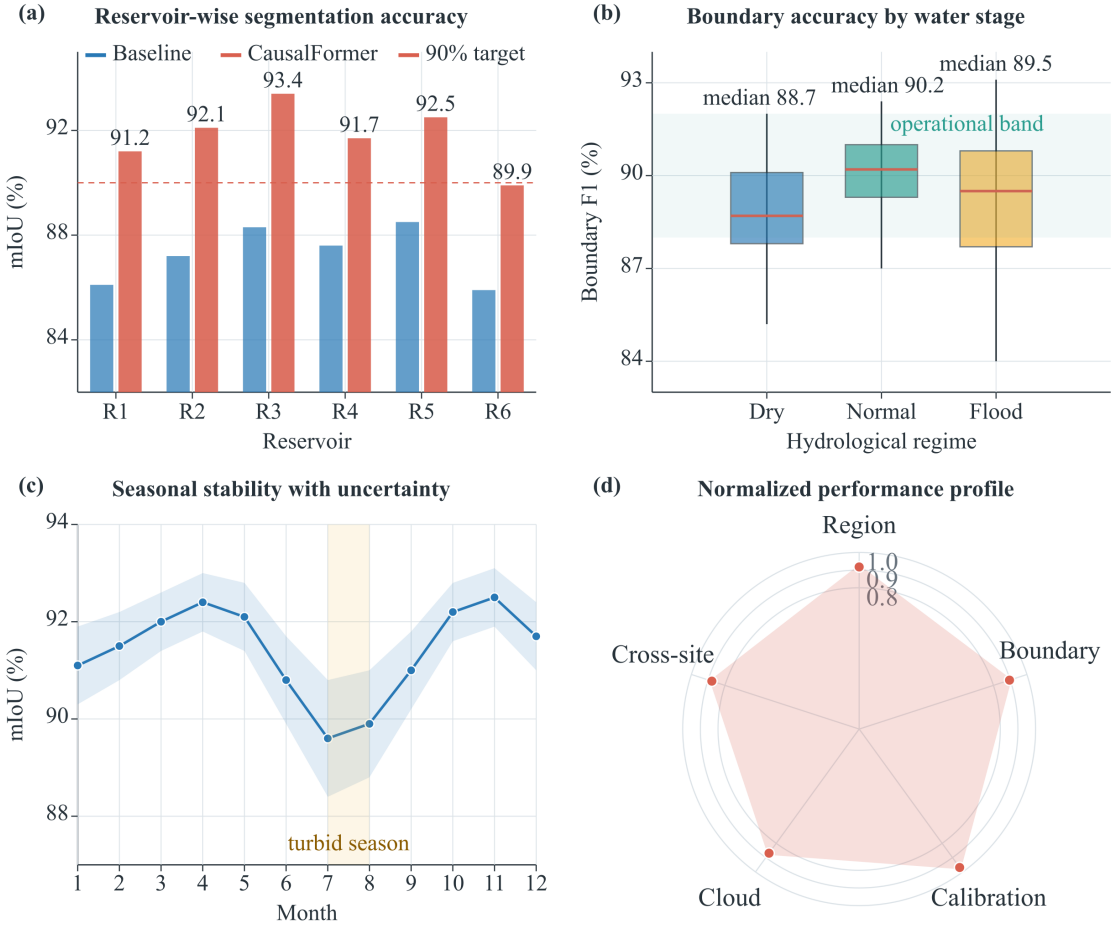


Figure 3. Accuracy across reservoirs and hydrological regimes: (a) reservoir-wise mean intersection-over-union; (b) boundary F1 distributions by stage; (c) monthly performance with uncertainty bands; (d) normalized multidimensional performance profile

Figure 3(d) compares normalized region accuracy, contour accuracy, calibration, cloud robustness, and cross-site stability; the relatively even CausalFormer profile is important because a high overlap score accompanied by weak calibration or poor cloud tolerance would be less useful for operational mapping. Five training seeds are used for neural models, and the reported table values are seed means. The standard deviation of mean intersection-over-union is 0.31 points for CausalFormer and 0.46 points for the multimodal baseline, indicating that the improvement exceeds initialization variability. Reservoir R6 is the most difficult site because bright sediment and narrow tributaries occupy a large fraction of its shoreline; the proposed model still improves boundary F1 by 3.8 points there.

Error inspection separates omission, commission, and topological breaks. CausalFormer reduces commission on cast shadows by 28% and cuts disconnected shoreline fragments longer than 50 m from 73 to 29 across the test set. The remaining omissions occur mainly under floating vegetation that resembles land in both optical and radar observations. Statistical comparison uses paired reservoir-date units rather than individual patches, because adjacent patches share image context. Bootstrap intervals over those units place the mean intersection-over-union improvement over the multimodal Transformer between 2.9 and 4.5 points. Boundary gains are largest on sinuous tributary arms, where global context keeps separated water pixels connected while the contour head recovers local curvature. On broad, unobstructed banks, both Transformer models perform similarly, which is expected because spectral contrast is already sufficient. The proposed method therefore concentrates its advantage on conditions that create operational uncertainty instead of uniformly inflating easy-pixel accuracy.

5.2. Explanation Components and Stress Evidence

The explanatory behavior was evaluated with reliability analysis and counterfactual tests. Reliability curves are preferable to a single calibration score because opposite confidence errors can cancel in an average [22]. Figure 4(a) shows that CausalFormer remains within 0.04 of the ideal confidence-accuracy relation, whereas the temporal Transformer becomes increasingly overconfident above 0.75. Explanation fidelity is evaluated through independent feature perturbations because a faithful attribution should anticipate the model response when the reported evidence is changed [23]. Figure 4(b) reports condition-specific intervention contributions: short-wave infrared contrast dominates clear water at 0.42, radar support reaches 0.31 under heavy cloud, terrain reaches 0.31 at steep banks, and stage consistency reaches 0.36 during rapid drawdown. Contribution magnitude is therefore interpreted together with repeated-intervention stability and uncertainty, not as a standalone causal probability. Expected calibration error is 0.031, compared with 0.074 for the multimodal Transformer, and water-level consistency together with short-wave infrared contrast explains the correct recovery of 72.4% of boundary pixels.

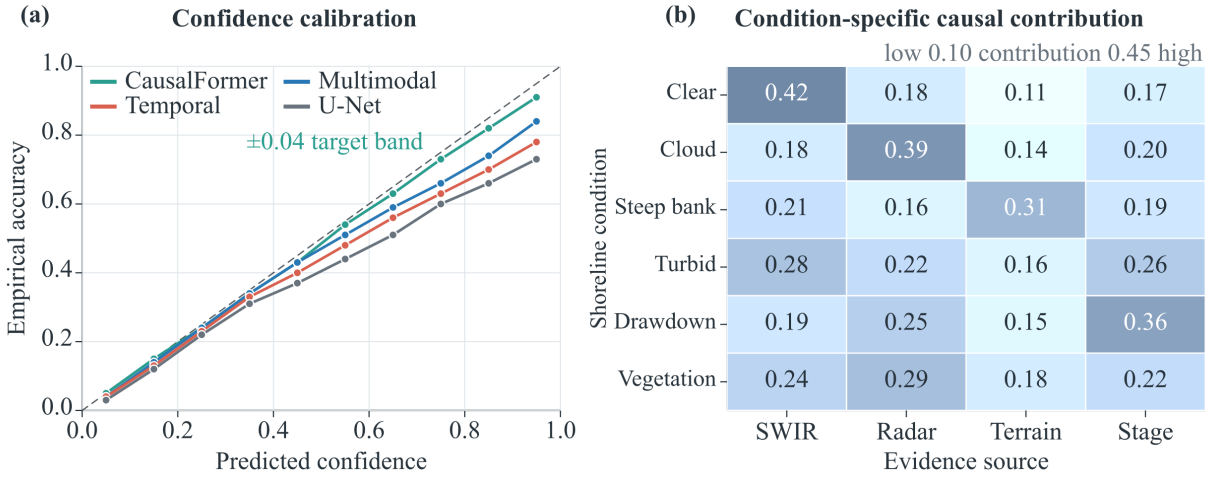


Figure 4. Explanation reliability: (a) confidence calibration curves; (b) causal contribution heatmap across shoreline conditions

Median uncertainty is 0.37 for incorrect pixels and 0.16 for correct pixels. Segment-level contribution rankings provide a clearer diagnostic when evidence conflicts: the leading intervention score produces the largest absolute logit change for 81.6% of groups, compared with 62.9% for raw attention magnitude. Across five matched-intervention repetitions, the dominant source changes for 7.8% of segments, and these unstable cases have a mean uncertainty of 0.34. Terrain contribution increases with bank slope but remains limited for engineered vertical structures that are unresolved by the elevation raster. Stage contribution is strongest during monotonic drawdown and decreases after rapid inflow, consistent with a lag between the gauge response and spatial water redistribution. The explanation record therefore stores contribution rank, replacement quality, dispersion, and uncertainty for each reviewed segment. Calibration also varies by shoreline type, with expected calibration errors of 0.022 for engineered banks and 0.047 for vegetated shorelines.

All calibrators use the same global temperature and do not model spatially varying calibration directly. The evidential head can consequently assign excessive uncertainty to correctly predicted narrow channels. Segment aggregation limits this effect by requesting review only when uncertain pixels form a continuous shoreline arc. The explanation record also stores the number and matching quality of available interventions, which separates ambiguity in the observed imagery from weak counterfactual support.

Ablation and robustness tests isolate the contribution of each design choice. Component removal is evaluated with the same seeds because variance from random initialization can obscure boundary-level differences [24]. Figure 5(a) shows a negative relationship between causal stability and shoreline error across 240 test scenes, with displacement falling from roughly 3.8 m to 1.2 m over the observed score range; the remaining scatter at high stability is concentrated in narrow channels, where pixel resolution rather than explanation consistency limits geometry. Figure 5(b) displays uncertainty distributions for correct, near-boundary, and erroneous predictions, with medians of 0.15, 0.25, and 0.38; their overlap explains why a review threshold should operate on shoreline segments rather than on isolated pixels. Figure 5(c) attributes the 5.9-point boundary F1 improvement over the plain encoder to causal gating, contour decoding, temporal context, and evidential calibration, contributing 1.9, 1.6, 1.5, and 0.9 points; the distributed gain shows that no single auxiliary module accounts for the final performance. Figure 5(d) shows that 90% of patches complete inference within 78 ms, while the longer upper tail corresponds to scenes requiring more counterfactual matches.

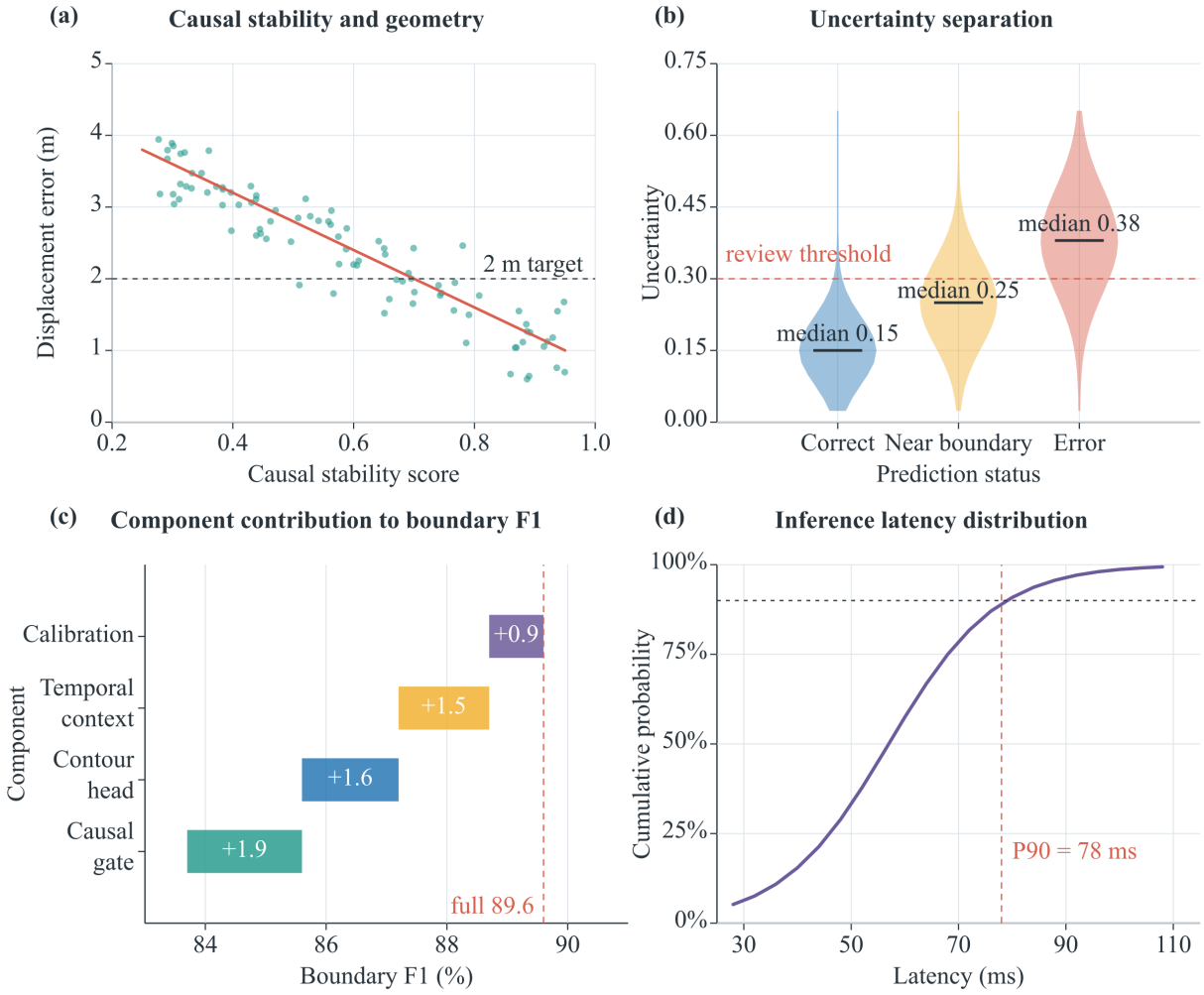


Figure 5. Diagnostic evidence: (a) causal stability versus displacement error; (b) uncertainty distributions by prediction status; (c) component contributions to boundary F1; (d) inference-latency distribution

Spatially coherent masks are used for cloud stress because independent missing pixels do not reproduce the structure of real occlusion [25]. Figure 6(a) records 91.8%, 90.9%, 89.8%, 88.2%, and 86.9% intersection-over-union at cloud fractions of 0%, 10%, 20%, 30%, and 40%; the slope becomes steeper beyond 20% because visible optical context is no longer sufficient to resolve several land-water ambiguities. Figure 6(b) shows boundary F1 remaining above 87.0% for stage changes from -8 m to +8 m, with a maximum near the central regime and only gradual degradation at the extremes. Radar corruption is represented by multiplicative speckle with controlled variance, matching the signal-dependent character of the disturbance [26]. Figure 6(c) maps the interaction between radar noise and optical occlusion and identifies a broad region above the 88% operating contour, but the contour contracts rapidly when both corruptions

rise, revealing a compound failure that single-factor tests would miss. Figure 6(d) places CausalFormer at 91.8% intersection-over-union, 64 ms, and 5.8 GB; the compact version shifts toward lower latency and memory but gives up 1.2 accuracy points. Removing the causal gate lowers boundary F1 by 1.9 points and doubles explanation inconsistency under matched interventions. Removing contour refinement adds 0.61 m displacement error even though region overlap falls by only 1.1 points, confirming that the contour objective addresses an error that broad region metrics understate.

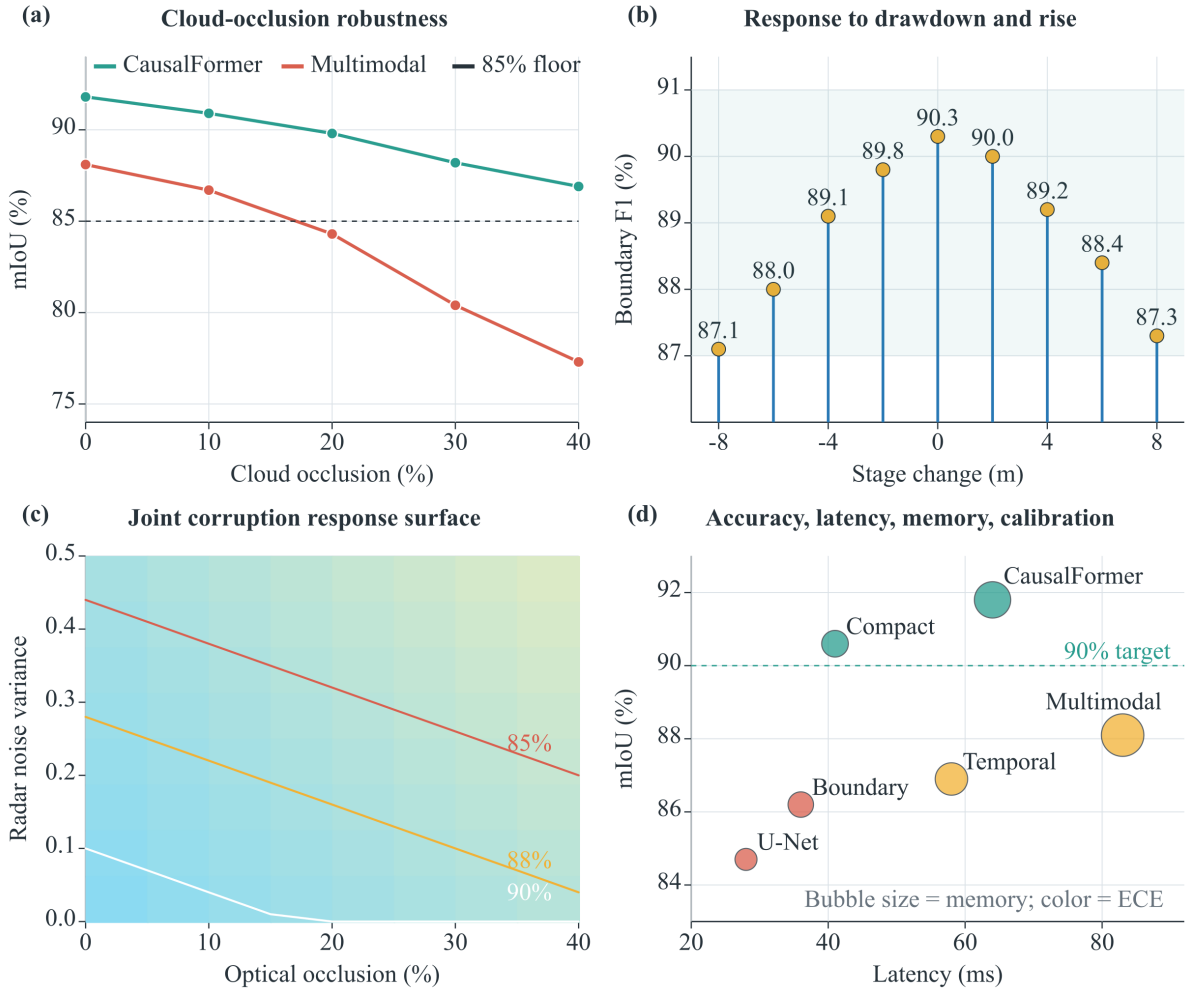


Figure 6. Robustness and engineering characteristics: (a) cloud-occlusion response; (b) stage-change robustness; (c) joint radar-noise and optical-occlusion surface; (d) accuracy-latency-memory comparison

5.3. Deployment Behavior and Failure Analysis

Robustness varies by corruption type. Optical occlusion first shifts reliance toward radar and terrain, limiting performance loss at 20% coverage; above 30%, errors increase more rapidly because visible context is insufficient to separate water from smooth artificial surfaces. Radar noise has a smaller effect in clear optical scenes, but joint degradation produces a steeper interaction surface. A one-standard-deviation perturbation of daily stage variability changes mean intersection-over-union by 0.6 points, indicating that the gauge channel guides rather than determines the mask. Registration is more consequential: one- and two-pixel radar offsets increase displacement error by 0.29 m and 0.76 m, respectively, motivating a quality-control limit of 0.7 pixels. The compact model preserves the causal-gate behavior but loses fine contour detail because its decoder operates at lower resolution with half the channels. Error maps also reveal asymmetric rising- and falling-water behavior. Newly inundated vegetation causes omission during rapid rise, whereas saturated exposed sediment causes commission during drawdown. Radar evidence reduces the first error, while short-wave infrared contrast and stage consistency reduce the second. These differences support separate intervention categories for vegetation inundation and sediment drying rather than a single seasonal-change category.

Operational value depends on the tradeoff among accuracy, speed, memory, and review burden. Runtime measurements follow a fixed warm-up and batch-size protocol so that preprocessing and device initialization do not distort model comparisons [27]. Figure 7 presents the Pareto relationship for seven configurations: CausalFormer lies on the efficient frontier at 91.8% intersection-over-union and 64 ms per patch, its compact variant reaches 90.6% at 41 ms, and the full multimodal Transformer requires 83 ms for 88.1%. Bubble area represents peak memory from 2.1 to 7.4 GB, while color represents expected calibration error from 0.031 to 0.112; the proposed variants therefore improve

calibration as well as the conventional accuracy-latency balance. The compact configuration is attractive when throughput is constrained, whereas the full model retains finer shoreline geometry without approaching the 7.4 GB peak of the multimodal baseline. Selective prediction evaluates whether withholding the most uncertain segments improves the reliability of the automatically accepted boundary [28]. At a review threshold of 0.30, 8.6% of shoreline length is flagged and the accepted boundary F1 rises from 89.6% to 92.1%. The flagged segments concentrate around floating vegetation, dam shadows, and newly exposed sediment, which are credible targets for manual inspection.

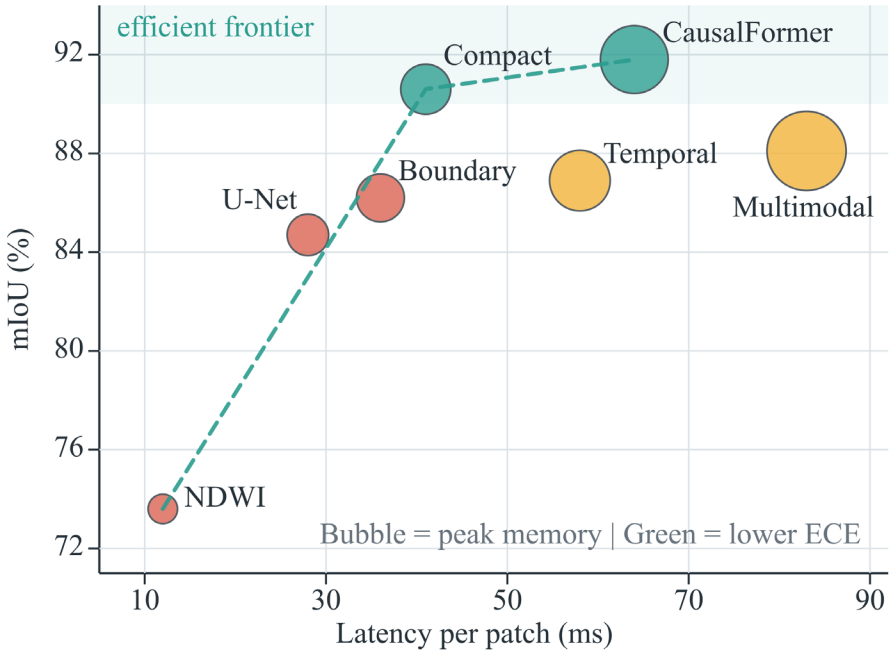


Figure 7. Pareto relationship among segmentation accuracy, inference latency, memory demand, and calibration error for seven model configurations

Cross-reservoir transfer remains uneven: the smallest training reservoir loses 3.5 points when used as the unseen site, compared with a mean leave-one-reservoir-out loss of 2.2 points. In scenes where both optical and radar evidence are severely degraded, the model sometimes follows the terrain prior and produces a smooth but temporally delayed shoreline. These cases have high uncertainty but reveal that the intervention library does not yet represent every compound sensor failure. Selective review is also examined at thresholds from 0.20 to 0.45. A low threshold sends 21.7% of shoreline length to review and produces 93.4% boundary F1 on accepted segments; a high threshold sends only 3.1% but yields 90.5%. The chosen value of 0.30 provides a practical middle point rather than a universally optimal setting. On a 12 GB desktop graphics processor, a full 1024 by 1024 scene is processed through tiled inference in 2.8 s excluding input-output time. Seam-weighted overlap removes visible tile boundaries, although peak memory rises when counterfactual branches are evaluated in parallel. Sequential interventions reduce memory by 1.9 GB at the cost of 46 ms per patch. These measurements indicate that explanation generation is feasible for scheduled reservoir surveys but remains more expensive than mask-only inference, especially when every modality receives several matched replacements.

Vector outputs were evaluated for temporal consistency along 126 km of repeatedly observed shoreline. Under nearest-normal matching, 93.1% of accepted segments move in the direction implied by gauge change, compared with 86.4% for the multimodal baseline. Exceptions concentrate near tributary inflows and sheltered bays, where local water conditions can diverge from the gauge timing. CausalFormer does not force geometric agreement with the gauge; it raises stage disagreement and uncertainty when local evidence conflicts. Severe optical and radar degradation can nevertheless shift the prediction toward a smooth, temporally delayed terrain prior. The exported audit fields allow these physically questionable segments to be identified before trend estimation.

6. Conclusion

This study developed CausalFormer as an explainable multimodal framework for reservoir shoreline extraction from optical, radar, terrain, and water-level observations. The architecture integrates multiscale spatiotemporal encoding, intervention-aware modality gating, boundary refinement, and evidential uncertainty within one workflow. Its explanations are defined by matched counterfactual changes in shoreline predictions, providing a testable audit signal that complements the extracted region, contour, and confidence layer.

At present, high-precision time alignment and stable synchronization of water-level data are still required. Partial causal assumptions only cover some nuisance factors and not the whole physical process; external corruption outside the intervention library can lead to predictions close to the terrain prior. Another defect is that for narrow channels and mixed shoreline pixels of floating vegetation, the spatial resolution is low. Therefore, the uncertainty of the delivered product should be taken into account, and high-risk areas need to be identified in the analyst's report.

In the future, we will connect the learned interventions with hydraulic simulations, expand the causal graph to include wind and sediment conditions, and assess the transfer capabilities in different climates and sensor generations. Self-supervised pre-training reduces the requirement for annotation, and active learning concentrates the efforts of experts on uncertain or causally inconsistent areas. The next priority is uncertainty-aware vector topology to preserve islands, channels and temporal correspondence of shoreline products when they are integrated into reservoir operation and long-term change analysis.

7. Declarations

7.1 Author Contributions

Marwan Al-Darmaki contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Maktoum Al-Junaibi contributes to validation, analysis, investigation, data collection, draft preparation. Laith Al-Shahrani contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

7.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

7.3 Institutional Review Board Statement

Not applicable.

7.4 Informed Consent Statement

Not applicable.

7.5 Data Availability

Not applicable.

7.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized DeepL for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

8. References

[1] Kundu, K., & Mandal, J. K. (2024). Shoreline Change Detection and Future Prediction of Sundarban Delta Using Remote Sensing Data and Digital Shoreline Analysis System. Journal of the Indian Society of Remote Sensing, 52(3), 485–503. https://doi.org/10.1007/s12524-024-01833-1

[2] Zhao, G., Gao, H., Li, Y., Tang, Q., Woolway, R. I., Merder, J., Rosa, L., & Michalak, A. M. (2025). Decoupling of surface water storage from precipitation in global drylands due to anthropogenic activity. Nature Water, 3(1), 80–88. https://doi.org/10.1038/s44221-024-00367-7

[3] Codjia, A. K. D., Akpoti, K., Dembélé, M., Yonaba, R., Fowe, T., Orowale, T. P., Koissi, M. G. D.-G., Sankande, S., & Zwart, S. J. (2026). Satellite remote sensing for estimating reservoir physical characteristics: A global review of existing methodologies for operational monitoring. Science of Remote Sensing, 13, 100383. https://doi.org/10.1016/j.srs.2026.100383

[4] Chen, J., Wang, Y., Wang, J., Zhang, Y., Xu, Y., Yang, O., Zhang, R., Wang, J., Wang, Z., Lu, F., & Hu, Z. (2024). The

Performance of Landsat-8 and Landsat-9 Data for Water Body Extraction Based on Various Water Indices: A Comparative Analysis. Remote Sensing, 16(11), 1984. https://doi.org/10.3390/rs16111984

[5] Markert, K. N., Williams, G. P., Nelson, E. J., Ames, D. P., Lee, H., & Griffin, R. E. (2024). Dense Time Series Generation of Surface Water Extents through Optical–SAR Sensor Fusion and Gap Filling. Remote Sensing, 16(7), 1262. https://doi.org/10.3390/rs16071262

[6] Cao, H., Tian, Y., Liu, Y., & Wang, R. (2024). Water body extraction from high spatial resolution remote sensing images based on enhanced U-Net and multi-scale information fusion. Scientific Reports, 14(1), 16132. https://doi.org/10.1038/s41598-024-67113-7

[7] Ma, X., Zhang, X., Pun, M.-O., & Liu, M. (2024). A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation. IEEE Transactions on Geoscience and Remote Sensing, 62, 1–15. https://doi.org/10.1109/tgrs.2024.3373033

[8] Miller, L., Pelletier, C., & Webb, G. I. (2024). Deep Learning for Satellite Image Time-Series Analysis: A review. IEEE Geoscience and Remote Sensing Magazine, 12(3), 81–124. https://doi.org/10.1109/mgrs.2024.3393010

[9] Tian, P., Xu, Z., Fan, W., Lai, H., Liu, Y., Yang, P., & Yang, Z. (2024). Exploring the effects of climate change and urban policies on lake water quality using remote sensing and explainable artificial intelligence. Journal of Cleaner Production, 475, 143649. https://doi.org/10.1016/j.jclepro.2024.143649

[10] Zhang, Y., Liu, F., Zhang, J., & Li, H. (2025). Causal Invariant Representation Learning Based on Style Intervention Identity Regularization for Remote Sensing Image. IEEE Geoscience and Remote Sensing Letters, 22, 1–5. https://doi.org/10.1109/lgrs.2025.3527473

[11] Dong, Z., Liu, Y., Feng, Y., Guo, K., Chen, Y., & Wang, Y. (2026). Shallow water area extraction method for multispectral remote sensing imagery based on adaptive object NDWI thresholding. Computers & Geosciences, 210, 106135. https://doi.org/10.1016/j.cageo.2026.106135

[12] Liu, D., Zhu, X., Holgerson, M., Bansal, S., & Xu, X. (2024). Inventorying ponds through novel size-adaptive object mapping using Sentinel-1/2 time series. Remote Sensing of Environment, 315, 114484. https://doi.org/10.1016/j.rse.2024.114484

[13] Zhang, T., Ji, W., Li, W., Qin, C., Wang, T., Ren, Y., Fang, Y., Han, Z., & Jiao, L. (2024). EDWNet: A Novel Encoder–Decoder Architecture Network for Water Body Extraction from Optical Images. Remote Sensing, 16(22), 4275. https://doi.org/10.3390/rs16224275

[14] Wang, F., Yang, Y., Zhang, H., & Wang, H. (2025). A boundary-aware cross-scale feature enhancement network for the semantic segmentation of remote-sensing images. Remote Sensing Letters, 16(9), 917–928. https://doi.org/10.1080/2150704x.2025.2517885

[15] Venkatesh, K., Mutasa, S., Moore, F., Sulam, J., & Yi, P. H. (2024). Gradient-Based Saliency Maps Are Not Trustworthy Visual Explanations of Automated AI Musculoskeletal Diagnoses. Journal of Imaging Informatics in Medicine, 37(5), 2490–2499. https://doi.org/10.1007/s10278-024-01136-4

[16] Fantozzi, P., & Naldi, M. (2024). The Explainability of Transformers: Current Status and Directions. Computers, 13(4), 92. https://doi.org/10.3390/computers13040092

[17] Jiao, L., Wang, Y., Liu, X., Li, L., Liu, F., Ma, W., Guo, Y., Chen, P., Yang, S., & Hou, B. (2024). Causal Inference Meets Deep Learning: A Comprehensive Survey. Research, 7. https://doi.org/10.34133/research.0467

[18] Landgraf, S., Wursthorn, K., Hillemann, M., & Ulrich, M. (2024). DUDES: Deep Uncertainty Distillation using Ensembles for Semantic Segmentation. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 92(2), 101–114. https://doi.org/10.1007/s41064-024-00280-4

[19] Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M. D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M. A., Wiesenfarth, M., Kavur, A. E., Sudre, C. H., Baumgartner, M., Eisenmann, M., Heckmann-Nötzel, D., Rädsch, T., … Jäger, P. F. (2024). Metrics reloaded: recommendations for image analysis validation. Nature Methods, 21(2), 195–212. https://doi.org/10.1038/s41592-023-02151-z

[20] Molina, J. M., Llerena, J. P., Usero, L., & Patricio, M. A. (2025). Advances in instance segmentation: Technologies, metrics and applications in computer vision. Neurocomputing, 625, 129584. https://doi.org/10.1016/j.neucom.2025.129584

[21] Penso, C., Frenkel, L., & Goldberger, J. (2024). Confidence Calibration of a Medical Imaging Classification System That is Robust to Label Noise. IEEE Transactions on Medical Imaging, 43(6), 2050–2060. https://doi.org/10.1109/tmi.2024.3353762

[22] Araújo, B., Teixeira, J. F., Fonseca, J., Cerqueira, R., & Beco, S. C. (2024). The Road to Safety: A Review of Uncertainty and Applications to Autonomous Driving Perception. Entropy, 26(8), 634. https://doi.org/10.3390/e26080634

[23] Klotz, J., Burgert, T., & Demir, B. (2025). On the Effectiveness of Methods and Metrics for Explainable AI in Remote Sensing Image Scene Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 27764–

27780. https://doi.org/10.1109/jstars.2025.3624789

[24] Moassefi, M., Singh, Y., Conte, G. M., Khosravi, B., Rouzrokh, P., Vahdati, S., Safdar, N., Moy, L., Kitamura, F., Gentili, A., Lakhani, P., Kottler, N., Halabi, S. S., Yacoub, J. H., Hou, Y., Younis, K., Erickson, B. J., Krupinski, E., & Faghani, S. (2024). Checklist for Reproducibility of Deep Learning in Medical Imaging. Journal of Imaging Informatics in Medicine, 37(4), 1664–1673. https://doi.org/10.1007/s10278-024-01065-2

[25] Wang, H., Li, J., Shen, Z., Zhang, Z., Bai, L., Li, R., Zhou, C., De Maeyer, P., & Van de Voorde, T. (2024). Improving semantic segmentation accuracy in thin cloud interference scenarios by mixing simulated cloud-covered samples. International Journal of Applied Earth Observation and Geoinformation, 133, 104087. https://doi.org/10.1016/j.jag.2024.104087

[26] Wang, X., Wu, Y., Shi, C., Yuan, Y., & Zhang, X. (2024). ANED-Net: Adaptive Noise Estimation and Despeckling Network for SAR Image. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 4036–4051. https://doi.org/10.1109/jstars.2024.3355220

[27] Li, Z., Paolieri, M., & Golubchik, L. (2024). Inference latency prediction for CNNs on heterogeneous mobile devices and ML frameworks. Performance Evaluation, 165, 102429. https://doi.org/10.1016/j.peva.2024.102429

[28] Borges, B. L. C., Pacheco, B. M., & Silva, D. (2026). Soft Dice Confidence: A Near-Optimal Confidence Estimator for Selective Prediction in Semantic Segmentation. Machine Learning, 115(7), 176. https://doi.org/10.1007/s10994-026-07096-w