Reading Shallow Flood Depth from Real USGS Gauge-Camera Footage with a Learned Staff-Plate Reader

Kalp Shah
Department of Electrical Engineering
Indian Institute of Technology Gandhinagar
kalp.shah@iitgn.ac.in

Udit Bhatia
Department of Civil Engineering
Indian Institute of Technology Gandhinagar
bhatia.u@iitgn.ac.in

Abstract

How accurately can a fixed camera estimate flood water depth, and how well does that transfer across cameras? I use real USGS staff-plate cameras whose official gauge heights serve as labels. A learned reader registers each frame, rectifies the plate, classifies plate rows as submerged and decodes the water level; an image-quality gate abstains on bad frames. On the main shallow-flow camera (2259 frames, 15 months), leaving out whole months gives 1.13 cm mean absolute error on the 67% of frames passing the gate (6.97 cm for a colour-threshold baseline) and 0.79 cm on two locked months; error stays below twice native up to 5.0 cm per pixel. Applied unchanged to three cameras downloaded afterwards, with each camera held out of training in turn, it beats a no-water guess on two cameras and fails on one; a river camera is read to 12 cm over a 211 cm flood.

1 Problem and Motivation

Question. Given one frame from a fixed camera that sees a reference object of known scale, estimate the water depth h (cm) at the object. Any flood response (severity class, warning level) is a function of h, so the target is h itself, with error reported by depth range. One row is one frame; prediction uses the current frame, a one-time setup (reference views, plate geometry) and, for pose recovery only, earlier frames, never the gauge. Camera water-level work mostly targets rivers or deep urban water (Vandaele et al., 2021; Liang et al., 2023); the main camera (USGS 411229084541101, Indiana) sees shallow overland flow.

2 Data and Preparation

Source and content. Frames come from the public USGS NIMS image bucket and labels are USGS gage height (both public domain, USGS & repos, 2026); NWIS was not reachable from my environment, so the main camera's gauge values come from two open-source projects' committed result tables (USGS & repos, 2026) and were not re-checked. No camera-derived value is a label. The main camera has 2259 frames (median spacing 21 min) in 15 months; gauge -0.6 to 55 cm; 35% dry, 25% infrared night; the plate spans 0–62 cm and the gauge sits beside it. The pose changes between months: SIFT registration (Lowe, 2004) succeeds on 82% of frames (the rest reuse the last good pose); images show fog, mud, droplets, glare and over-exposure; the plate bottom is at 4.3 cm, so shallower water is unobservable.

Other cameras. Added after the method was frozen, used only for evaluation: three staff-plate cameras (≈1400 frames each, Jul–Nov 2026, gauge from the NWIS API) and a river camera (USGS 01421500, Hancock NY) whose stage reaches the plate in one flood (39 gated frames over 211 cm).

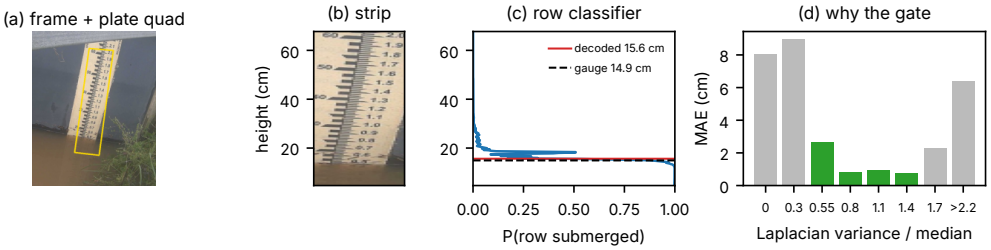


Figure 1: The reader on one held-out frame. (a) Frame and plate quad; (b) rectified strip; (c) row probability, decoded level (red) and gauge (dashed); (d) CV error by Laplacian variance / median (green: inside the gate).

Preprocessing and leakage control. Frames are resized to 1280 px and the plate quad is rectified to a 360×48 strip (row ≈ 0.16 cm). Everything learned (dry template, calibration, classifier, baseline threshold, gate thresholds) is fitted on training months only; registration references come from development months and use geometry, not labels.

3 Methods

Baselines. B0: always 0 cm. B1: a row is wet if its colour change from the dry template exceeds a tuned τ, then the decoding below.

Learned reader. For frame i and strip row k (height h_k), a gradient-boosted classifier (scikit-learn, Pedregosa et al., 2011) outputs $p_{ik} = P(\text{row submerged})$ from row colour, template change, texture, context and frame quality (Fig. 1). Labels are $y_{ik} = \mathbf{1}[h_k < g_i]$ (g_i the gauge): no manual annotation. Water level is the monotone change point $b^\star = \arg\max_b \sum_{k<b} \log(1-p_{ik}) + \sum_{k\geq b} \log p_{ik}$, mapped to height and then to centimetres by a least-squares affine fit; levels at or below the plate bottom report 0. The log-likelihood gap to the runner-up boundary is the confidence *margin*.

Cross-camera protocol. For each target camera the classifier is trained on the *other* cameras only (labels from their nominal plate geometry); the target's dry template and gate thresholds come from its earliest month. *Zero-shot* uses the nominal geometry; *calibrated* fits an affine map on that month; later months are evaluated. Hancock: one level offset from 30 development frames; *pooled* adds them to training.

Quality gate. A frame is rejected when its Laplacian variance lies outside $[0.55, 1.7]$ times the training median (low: fog, blur; high: lens dirt, glare).

4 Evaluation Design

The split unit is the calendar-month block, never the frame (consecutive frames are near duplicates; lighting and pose are shared within a month). Two main-camera blocks (2025-04, 2026-04) were locked in `make_usgs_splits.py` before any model was trained; the other 13 (1976 frames) support leave-one-block-out cross-validation (CV) and all tuning. **Disclosure:** I looked at the locked blocks twice: dev errors showed dirty-lens frames also fail, so the one-sided gate became two-sided and the run was repeated. The Hancock flood is locked and used for no template, offset or training. **Metrics.** MAE (primary), RMSE and bias (cm), overall and by gauge depth range (dry ≤ 0.5 cm, 0.5–10, >10 cm); 95% intervals bootstrap whole blocks (two locked blocks give none).

5 Results and Error Analysis

Main result. Table 1: on gated CV frames the learned reader has 1.13 cm MAE (95% CI 0.67–1.69) versus 6.97 cm for the threshold baseline and 8.80 cm for always-0. On the two locked months it reaches 0.79 cm (baseline 11.52, always-0 18.90). Without the gate the errors are 3.38 cm (CV) and 3.11 cm (locked): the gate buys reliability by abstaining on 33% (CV) and 18% (locked) of frames.

Table 1: Held-out depth error (cm). Gated = frames passing the quality gate (coverage 67% CV, 82% locked); last three columns: MAE by gauge depth range; CI: block bootstrap.

Eval.	Frames	Method	n	MAE [95% CI]	RMSE	bias	dry	0.5–10	>10
CV	gated	learned reader	1319	1.13 [0.7, 1.7]	3.1	+0.4	0.4	2.0	1.5
	gated	threshold baseline	1319	6.97 [6.0, 8.4]	10.1	+1.1	5.9	8.0	7.8
	all	learned reader	1976	3.38 [2.1, 4.5]	7.3	+0.5	1.9	5.0	4.1
Locked	gated	learned reader	233	0.79	1.2	+0.0	0.0	1.6	0.6
	gated	threshold baseline	233	11.52	14.6	-8.7	9.2	7.2	12.5
	all	learned reader	283	3.11	8.9	-2.1	1.6	2.1	3.4

Table 2: Gated MAE (cm) per camera; max = largest gauge (cm); "/ 0" = no-water guess on the same frames. † Hancock: flood range above the plate bottom; its guess is the plate bottom.

Camera	Scene	max	n	in-site / 0	held-out zero-shot / 0	held-out calib.
in_overland	cream plate	55	1319	1.13 / 8.80	6.69 / 11.16	7.14
in_north	cream plate	25	948	1.05 / 1.44	1.08 / 1.50	1.08
oh_wetland	white plate	42	1373	0.54 / 2.73	1.52 / 2.01	1.81
wi_mt1	orange plate	19	1342	0.55 / 0.27	4.96 / 0.33	4.96
hancock	river plate	211†	–	–	20 / 86	12

Per-month MAE has median 1.06 cm and worst 3.78 cm. A causal five-reading median worsens MAE (1.13 to 1.77 cm): levels move tens of cm per hour.

Other cameras (Table 2, Fig. 3 row 3). In-site (leave-one-month-out) the reader beats always-0 on in_overland, in_north (1.05 vs. 1.44 cm) and oh_wetland (0.54 vs. 2.73); wi_mt1 stays nearly dry (gauge $\leq$19 cm) and always-0 wins. With the camera held out, zero-shot reading beats always-0 on in_overland (6.69 vs. 11.16 cm, but only 47% pass the gate), in_north (1.08 vs. 1.50) and oh_wetland (1.52 vs. 2.01) but *fails* on wi_mt1 (4.96 vs. 0.33): never having seen an orange plate, the classifier reads it as fully submerged in sunny frames (69 cm at gauge 0). On frames with water (gauge $>$0.5 cm) its error is much larger (6.1 and 24.7 cm vs. 13.3 and 34.1 cm for the no-water guess), and calibration does not help. On the river plate (Fig. 2c) zero-shot MAE over the 211 cm flood is 20 cm (plate-bottom guess: 86), 12 cm calibrated; pooling helps inside the stage range it saw (4, 7 vs. 8, 11 cm at 11–16 ft) and hurts beyond it (37 vs. 22 cm at 16–19 ft).

Failure audit (Fig. 3, row 2). Infrared is *not* worse than daylight (3.30 vs. 3.41 cm); costly cases are a borrowed pose (9.82 vs. 1.91 cm) and frames outside the gate (7.90 vs. 1.13 cm). In the worst accepted day frame, glare washes out the plate and water streaks on the panel look like wet rows.

6 Graduate Depth Analysis

Resolution sweep and error model. Registration is computed once at native resolution, then frames are area-downsampled by $f \in \{1, \ldots, 64\}$ and the plate re-sampled with $360/f$ rows ($s = 0.16f$ cm/px); the reader is retrained per f. Fig. 2d: gated MAE is 1.13 cm at native scale, 1.36 cm at 2.5 cm/px, 1.84 cm at 5.0 cm/px and 3.61 cm at 10.0 cm/px (over twice native). With scale s and waterline error σ_v (px), $\mathrm{Var}(\hat{h}) = s^2\sigma_v^2 + \sigma_g^2$, σ_g being gauge–plate offset and calibration error; fitting RMSE^2 to the gated sweep gives $\sigma_v = 0.55$ px and $\sigma_g = 3.0$ cm. The σ_g floor dominates at fine scales: resolution is not the binding constraint here. Ranking frames by decode margin instead of the Laplacian gate gives 2.28 cm at the gate's coverage (gate: 1.13 cm), so a cheap image statistic abstains better than the classifier's own confidence.

7 Limitations and Conclusion

Each camera gives an effective sample size of months, not frames; the river camera contributes one flood (39 frames). Labels are a gauge beside the plate, not tape measures; water below the plate bottom (4.3 cm) is unobservable. The locked set is two months, seen twice, and the gate was designed with dev-month error in view. Stronger claims need more cameras and tape-measured depths.

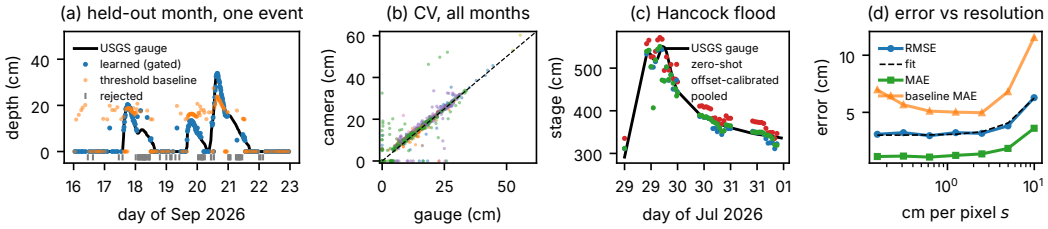


Figure 2: Held-out predictions. (a) One event (ticks: rejected frames). (b) Camera vs. gauge, coloured by month. (c) River camera, locked flood. (d) CV error vs. plate scale s; dashed: error model.

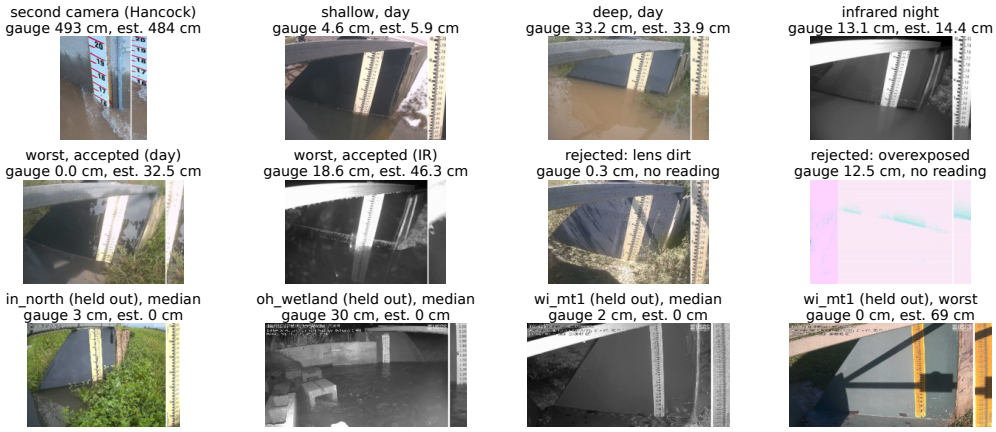


Figure 3: Real frames (crop and plate strip). Row 1: median-error examples (and Hancock). Row 2: largest accepted errors, rejected frames. Row 3: cameras held out of training, zero-shot (last: wi_mt1 failure).

Conclusion. On real USGS footage a learned plate reader with a quality gate reads depth to about 1.13 cm (CV) on accepted frames, far better than a colour-threshold baseline, and stays within twice its native error up to about 5.0 cm/px; it abstains on 33% of frames.

8 Reproducibility and AI-Use Disclosure

Code, splits, gauge tables, the trained model and all results are in `github.com/kalpshah18/FloodCam`; `scripts/make_paper_assets.py` regenerates every number, table and figure here from `results/`. Frames are public domain; no credentials or private data are used. **AI use:** Claude Code (Anthropic) helped with brainstorming, code, debugging, experiments, the dashboard and drafting; I verified the pipeline with unit tests and by inspecting strips, failures and split leakage, and I am responsible for all results.

References

R. Vandaele et al. River-level monitoring from camera images. *HESS*, 25, 2021.

Y. Liang et al. V-FloodNet: urban flood video segmentation. *Environ. Model. Softw.*, 2023.

USGS NIMS images and NWIS gage height; GitHub `waterline`, `stagesight`.

D. Lowe. SIFT keypoints. *IJCV* 60(2), 2004.

F. Pedregosa et al. Scikit-learn. *JMLR* 12, 2011.