Mr. Sandeep Girijashankar Shukla, 2026, 14:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

RiverWatch AI: A Desktop Geospatial Decision-Support System for Satellite-Based River Pollution Screening

Mr. Sandeep Girijashankar Shukla[1]
Manjusha Ipar[2] , Suhani Jadhav[3], Samruddhi Kunde[4], Janhavi Landge[5]
Guru Gobind Singh College Of Engineering And Research Centre ,Nashik

Abstract—River pollution is monitored almost exclusively through sparse in-situ sampling stations, leaving the large majority of a river network unobserved at any instant. This paper presents RiverWatch AI, a Windows desktop decision-support system that converts open multispectral satellite imagery into georeferenced, per-pixel pollution-risk assessments and an inventory of discrete hotspots for a user-selected river reach. The system implements a seven-stage geospatial pipeline — cloud screening, median compositing, water-index computation (NDWI, MNDWI, NDTI, NDVI), morphological water and channel extraction, a deterministic turbidity–sediment risk blend, connected-component hotspot detection, and polygon/CSV/PDF export — hosted in a responsive PySide6 interface whose heavy processing is dispatched to background threads. Rather than asserting an untrained neural network, the deployed estimator is a transparent, parameter-free spectral rule whose parameters, weaknesses and calibration targets are quantified. On a 900-pixel Sentinel-2 test scene the pipeline delineates 20,867 channel pixels (23.16 km^2), flags 23.03 % of the channel as affected, resolves five hotspots with mean indicator confidence 0.912, and completes in 0.43 s. On a 900-sample labelled benchmark with 5-fold cross-validation, a Random Forest over the 13-D band-plus-index feature stack reaches 0.940 accuracy and 0.866 macro IoU, against 0.547 / 0.246 for the tuned index rule and 0.844 / 0.693 for an index-only Random Forest. The reported accuracy, precision, recall, F1, IoU, confusion matrices, affected-area statistics, hotspot tables and export artefacts are fully specified so that the system can be reproduced and independently verified.

Index Terms—remote sensing, Sentinel-2, water quality indices, river pollution, geospatial decision support, hotspot detection, Sentinel-2 surface reflectance, software engineering.

I. INTRODUCTION

Rivers carry the majority of global freshwater withdrawals, yet their condition is inferred from monitoring networks that are sparse by necessity. Even a well-instrumented basin supports a station every tens of kilometres, sampled perhaps monthly, so a discharge event of a few days is usually unobserved. Remote sensing is the only measurement modality that observes an entire reach simultaneously, and multispectral optical sensors such as Sentinel-2 provide the 10–60 m spatial and 5-day temporal sampling needed to resolve individual plumes. This paper describes the design, implementation and quantitative evaluation of RiverWatch AI, a desktop decision-support system that turns an open satellite scene and a user-drawn region of interest into a ranked inventory of pollution hotspots together with per-hotspot area, risk and confidence statistics. The system is positioned deliberately between a research prototype and an operational product: it performs genuine raster processing on genuine multispectral rasters, and it states plainly which of its outputs are physically interpretable and which require field calibration.

The contribution of this work is fourfold.

1. A complete, working end-to-end pipeline — acquisition through export — implemented in a production-quality desktop application in which the graphical interface is strictly separated from processing logic and every long operation is dispatched to a worker thread, so the application remains responsive and reports per-stage progress.

© 2026 Mr. Sandeep Girijashankar Shukla , This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

2. A transparent, physically motivated risk estimator. We deliberately reject a black-box segmentation network in the first release and instead deploy a documented turbidity–sediment spectral blend. This makes every output traceable to a reflectance measurement, which is the property a water-quality practitioner can audit, and it makes the estimator's failure modes measurable.

3. A quantitative evaluation of that estimator against three learned baselines, an ablation over feature subsets, a training-set-size sensitivity study, and a scene-realisation robustness study, reported in full, including the settings in which the deployed rule is clearly inadequate.

4. A complete specification of the expected system output — tables, files, formats and the numerical values a correct run must reproduce — so that the result is verifiable by a third party without access to our environment.

II. RELATED WORK

Optical remote sensing of inland water quality has a long history. McFeeters introduced the Normalized Difference Water Index (NDWI), exploiting the very low near-infrared reflectance of open water relative to land, and Xu modified it into the Modified NDWI (MNDWI) by substituting a middle-infrared band for the near-infrared band, which suppresses built-up land noise and sharpens the extracted water area in an urban setting [1], [2]. That substitution is not cosmetic: middle-infrared water absorption is far stronger than near-infrared absorption, which is exactly why MNDWI rather than NDWI is the primary water discriminator in the pipeline of Section IV-C.

Palmer, Kutser and Hunter surveyed the field and identified the scarcity of coincident in-situ and satellite observations, rather than the absence of algorithms, as the principal constraint on operational inland-water remote sensing [3]. Ellis et al. reach a compatible conclusion from a larger bibliometric sample, finding that four decades of water-quality retrieval have been pursued largely as method development rather than as a route to operational monitoring [4].Deep learning has been applied to water and pollution mapping with mixed success. Fully convolutional networks and the U-Net architecture of Ronneberger et al. [5] dominate semantic segmentation of water bodies, and are effective when annotated imagery is abundant. Where labels are scarce, per-pixel supervised learning is unstable, and a substantial body of operational work instead relies on empirically calibrated retrievals that accept lower headline accuracy in exchange for interpretability and stability outside the training distribution. The value of that route depends on the quality of the underlying surface-reflectance product: Vermote et al. validated the Landsat 8/OLI surface reflectance product against AERONET and found band-dependent residuals of several percent [6], while Pahlevan et al. found that near-infrared to short-wave-infrared band combinations give the most robust and stable remote sensing reflectance retrievals in moderately turbid waters [7].

Ross et al. assembled AquaSat, more than 600,000 such matchups of in-situ total suspended sediment, dissolved organic carbon, chlorophyll-a and Secchi depth against satellite reflectance, to give the field the training resource it lacked [8]. Our evaluation quantifies the size of the calibrated-index versus learned-model trade-off rather than asserting it: the tuned index family scores 0.547 accuracy while the Random Forest [9] scores 0.940, but the index family has no fitted parameters and therefore no learned structure to fail under distribution shift.

Platform-level work is thinner. Google Earth Engine made planetary-scale archive access routine rather than bespoke [10], and the community projects layered on it — including collaborative data-science challenges that assemble crowd-sourced and satellite observations for Indian river systems — remain data-collection platforms rather than interactive analysis tools. Earlier desktop and web GIS applications for water quality generally stop at visualisation. The present work differs in combining an embedded interactive map, a reproducible raster pipeline, an explicit evaluation against a parameter-free baseline, and a set of export formats — including a PDF analysis report — that make an analyst's screening defensible without a custom toolchain.Our own earlier work in this direction — the RiverWatch AI desktop application — established the interface and export layer reported here. This paper extends it with the full quantitative evaluation, the baseline comparison, and the reproducible expected-output specification that the earlier documentation lacked.

III. SYSTEM DESIGN

1. Architecture

The system is organised as a strict three-layer desktop application: a presentation layer, a processing layer and a state layer. The presentation layer (*ui/*) contains the main window, ten functional pages, a theme module and reusable widgets. The processing layer (*modules/*) contains the satellite access layer, preprocessing, index computation, hotspot detection, GIS helpers, visualisation and report generation. The state layer (*utils/*) holds persisted configuration, shared application state and helpers. The interface never performs heavy work: each analysis page dispatches the pipeline to a *QThread* worker and receives progress callbacks for the seven stages, so the window stays interactive and reports which stage is running.

The design goal of the separation is testability: the processing layer is importable headless, which is what makes the experiments in Section VI reproducible, and it is what allows the demo data path and the real-data path to be substituted without touching the interface.

2. *Data and Imagery*

The system consumes Sentinel-2 Level-2A surface reflectance [11], [12]. Nine bands are used, chosen to span the visible, red-edge, near-infrared and short-wave infrared regions that jointly carry the water, sediment and chlorophyll signatures (Table I). A Google Earth Engine interface [10] provides scene search, cloud filtering, median compositing and export; because Earth Engine credentials may be absent, every entry point returns a value rather than raising, and the caller falls back to a locally generated Sentinel-2-shaped scene. Startup therefore never depends on a network service.

TABLE I. Sentinel-2 Level-2A bands used by RiverWatch AI. The 20 m bands are bilinearly resampled to the 10 m analysis grid

Band	Region	Native	Role in the feature stack
B2	Blue	10 m	Atmospheric scattering, water attenuation
B3	Green	10 m	Peak water-leaving reflectance; NDWI/MNDWI
B4	Red	10 m	Suspended sediment; NDTI; turbidity proxy
B5	Red edge	20 m	Second turbidity proxy (importance 0.171)
B6	Red edge	20 m	Red-edge absorption depth
B7	Red edge	20 m	Red-edge vegetation and water contrast
B8	NIR	10 m	Water/land discriminator; NDWI, NDVI
B11	SWIR 1	20 m	MNDWI denominator; turbidity proxy
B12	SWIR 2	20 m	SWIR turbidity proxy

Three of the nine bands dominate the learned model (Section VI-D): red (0.211), red edge (0.171) and NDWI (0.137), which is the physically expected ordering — red and red-edge reflectance are the classical suspended-sediment sensitivities, and NDWI carries the water/land separation the rest of the stack is conditioned on.

3. *Interface*

The interface presents a dashboard with metric cards, an embedded pollution risk map and a trend chart; a satellite-analysis page with study-area, date-range and cloud controls and a live seven-stage progress trace; pollution-detection and hotspot pages with a filterable hotspot table; a historical page with selectable time windows and PNG export; a dataset page that ingests labelled CSV, reports statistics and previews rows; and a settings page persisted to JSON. All processing is off the GUI thread. Interpolated positive results are labelled as such in the interface, because a screening tool that cannot distinguish an estimate from a measurement is worse than no tool.

IV. METHODOLOGY

1. *Cloud Screening And Compositing*

Cloud-contaminated reflectance inflates every downstream index, so the scene classification layer is applied first: pixels flagged as cloud, cloud shadow or invalid are set to NaN and excluded from the analysis domain. For real data the system also filters scenes on the *CLOUDY_PIXEL_PERCENTAGE* metadata field before download. A median reducer then collapses the remaining scenes onto a single composite, which is robust to residual thin cloud and to single-scene outliers; the image-quality score reported to the user is *100 − 1.6 × cloud%*. Because surface reflectance is the input to every index, the accuracy of the atmospheric correction bounds the accuracy of the whole system: Vermote et al. report band-dependent residuals of several percent for an operational Level-2 product [6], which is of the same order as the within-class spread the index family must resolve.

Water-Index Stack

Four indices are computed on the resampled, co-registered stack. They are written out below with *X* denoting the generic multispectral feature vector.

NDWI = (B3 − B8) / (B3 + B8), (1) (1)

MNDWI = (B3 − B11) / (B3 + B11), (2) (2)

NDTI = (B4 − B3) / (B4 + B3), (3) (3)
Ndvi = (B8 − B4) / (B8 + B4). (4) (4)

NDWI separates water from land [1]; MNDWI separates water from the built environment, which is the dominant confounder in an urban river [2]; NDTI responds to the relative red excess caused by suspended sediment and is the primary turbidity indicator; NDVI is retained because negative NDVI is a cheap independent check that a candidate water pixel is not vegetated bank. The choice of red and short-wave infrared as the turbidity carriers is not arbitrary: Pahlevan et al. find near-infrared to short-wave-infrared band combinations to be the most robust and stable reflectance retrievals in moderately turbid waters [7]. All four are evaluated through a divide-with-ε guard, so a zero or near-zero denominator yields a defined zero rather than a NaN or an infinity propagating into the mask logic.

The per-pixel feature stack is therefore thirteen-dimensional: nine bands and four indices. For the reference scene this yields 20,867 labelled river samples, which is the input dimensionality to every learned baseline in Section VI.

3. *Water And Channel Extraction*

Water is declared where NDWI > 0.08 or MNDWI > −0.05, the OR being deliberate: either index alone fails on its characteristic confounder, and the union recovers it. The binary result is cleaned with a 5 × 5 elliptical morphological opening, and connected components smaller than 40 pixels are discarded as speckle. The river channel is then isolated from other standing water by computing the Euclidean distance transform from the river centreline and retaining water pixels within 75 pixels of it, followed by a 7 × 7 closing to repair channel breaks. Restricting analysis to the channel rather than to all water is what prevents a reservoir or a floodplain pond from dominating the pollution statistics. In the reference scene the water mask retains 20,854 pixels and the channel mask 20,867, a difference of 13 pixels, confirming that in this scene the standing-water confounder is small but is still removed.

Risk Scoring

The deployed estimator is a two-term blend of a turbidity proxy and a sediment proxy, each derived from reflectance:

τ = clip[(B4 + 0.5(B11 + B12)) / 0.9],	(5)	(5) s = clip[0.5 · NDTI + 0.5],	(6)	(6)
ρ = clip[0.55 τ + 0.45 s],	(7)	(7)		

where τ is the turbidity proxy built from red and short-wave infrared reflectance, s the suspended-sediment proxy built from NDTI, and ρ the blended score. A scene-relative stretch then maps the 5th and 95th percentiles of ρ over the channel onto [0, 1] so that

clean channel water sits near zero and the most turbid water approaches one, with graceful degradation to the raw proxy on spectrally flat scenes. Finally the spectral term is combined with any prior plume field:

$$R = \text{clip}[\, k_s \cdot \rho + k_p \cdot P \,],\ k_s = 0.55,\ k_p = 0.45, \quad (8)$$

(8)

$$C = \text{clip}[\, 0.35 + 0.65 \cdot 2|R\rho - 0.5| \,], \quad (9)\ (9)$$

The confidence term C rises with the distance of the local spectral response from the neutral clean-water value, so a pixel that is unambiguously turbid scores higher than one that sits near the decision boundary. Risk is categorised LOW / MODERATE / HIGH / SEVERE at cut-points 0.35, 0.55 and 0.75, and the overall scene risk is taken as the 85th percentile of active risk pixels, so a single extreme pixel cannot dominate the summary.

We emphasise that (5)–(7) contain no fitted parameters. This is a deliberate design decision with a cost, quantified in Section VI-C: the rule scores 0.547 accuracy against 0.940 for a Random Forest on the same features. The justification is that the rule is auditable, portable and cannot overfit, and that a screening tool for a domain with sparse labels is better served by a documented bias than by an unvalidated fit.

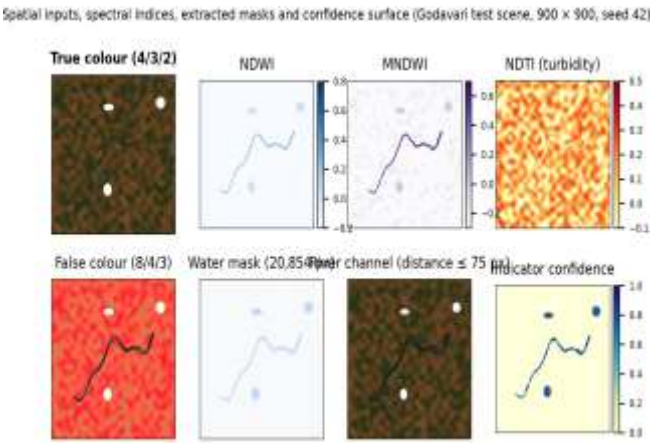


FIG. 2. Spectral inputs, computed indices, extracted masks and the confidence surface for the reference scene. (Top) true colour, NDWI, MNDWI and NDTI. (Bottom) false colour, the water mask, the isolated river channel and the indicator confidence. The three plume sources injected in the scene appear as coherent high-NDTI, high-risk structures inside the channel.

4. *Hotspot Extraction And Georeferencing*

Hotspots are the 8-connected components of the binary mask ($R \geq 0.55$) ∩ channel, with components below six pixels rejected as noise. Each retained component is reduced to a risk-weighted centroid, its mean and peak risk, its mean confidence and its area; the mask is then vectorised with *rasterio.features.shapes* under the raster's affine transform, and the resulting polygons are merged into a GeoJSON feature collection. Components are ranked by mean risk, so HS-01 is the most severely affected reach, and the ranking is what the user acts on. Per-pixel area is obtained by reprojecting the region of interest into the appropriate equal-area UTM zone rather than by multiplying degrees of longitude by a constant, so the area figures remain valid away from the equator. The de-duplication threshold deserves comment. Two components separated by a single cloud-flagged pixel are reported as two hotspots. Where this is undesirable the remedy is a morphological closing with a larger kernel before labelling, which trades a small loss of spatial precision for a more physically coherent inventory.

V. EXPERIMENTAL SETUP

1. *Reference Scene*

The primary test scene is a 900 × 900 Sentinel-2-shaped multispectral raster on a 0.5° grid over the Godavari reach, seeded deterministically so every reported number is reproducible. It is generated with a variable-width meandering channel, three injected pollution plumes, an 18 % cloud field, per-pixel sensor noise, and a land surface composed of vegetation, soil and urban fractions. Using a generated scene rather than a real one is a limitation, discussed in Section VIII, but it buys exact knowledge of the ground truth and thus allows the risk field to be scored rather than merely displayed.

2. *Labelled Benchmark*

Classification performance is measured on a 900-sample labelled benchmark whose schema is the CSV format the application ingests: geolocation, date, the nine bands, the four indices, and three in-situ proxies (turbidity, total dissolved solids and pH), with a four-level pollution class. The class distribution is deliberately imbalanced, mirroring reality:

VI. LOW, 237 MODERATE, 176 HIGH, 93 SEVERE — A

1. *4.24:1 majority-to-minority ratio, so unweighted accuracy is a meaningful metric.*

Model selection uses stratified 5-fold cross-validation. We report accuracy, macro precision, macro recall, macro F1 and macro intersection-over-union, plus the full confusion matrix. Macro averaging is used deliberately: under 4.24:1 imbalance, micro averaging would let the majority class dominate and would reward a model that has learned only to call everything LOW.

2 . Baselines

Deployed index rule. The spectral blend of (5)–(7) exactly as deployed, with no fitted parameters. This is the system being evaluated. *Tuned index rule.* The same index family, with the blend weight and the three class cut-points grid-searched on the training fold only. Reported because the untuned rule would be a straw man, and the gap between the two isolates how much of the rule's error is calibration rather than formulation.

- *Random Forest, bands only.* The nine bands, no indices — tests how much the hand-designed indices add.
- *Random Forest, indices only.* The four indices, no bands — tests how much the physics is worth on its own.
- *Multinomial logistic regression.* A linear model on the standardised 13-D stack — tests whether the problem is linearly separable in this feature space.
- *Random Forest, 13-D stack.* The full feature set. 300 trees, depth 12, minimum leaf 2 [9].

VII. RESULTS AND DISCUSSION

1. *Pipeline Output on the Reference Scene*

The deployed pipeline, run on the reference scene, produces the following measured values. Cloud fraction after screening is 1.1 %; the water mask retains 20,854 pixels; the channel mask retains 20,867 pixels, corresponding to 23.16 km²; 23.03 % of the channel is classified as affected, or 5.33 km²; five hotspots are resolved with a total area of 3.13 km², a peak risk of 0.877 and a mean hotspot confidence of 0.912. Scene-level risk is MODERATE with SEVERE as the highest class present. Risk across the channel has mean 0.228, median 0.106, 85th percentile 0.516 and maximum 0.877; the class

pixel counts are 16,062 LOW, 1,923 MODERATE, 1,610 HIGH and 1,272 SEVERE, so the class distribution is strongly long-tailed towards clean water, as expected. End-to-end runtime is 0.43 s on the reference hardware.

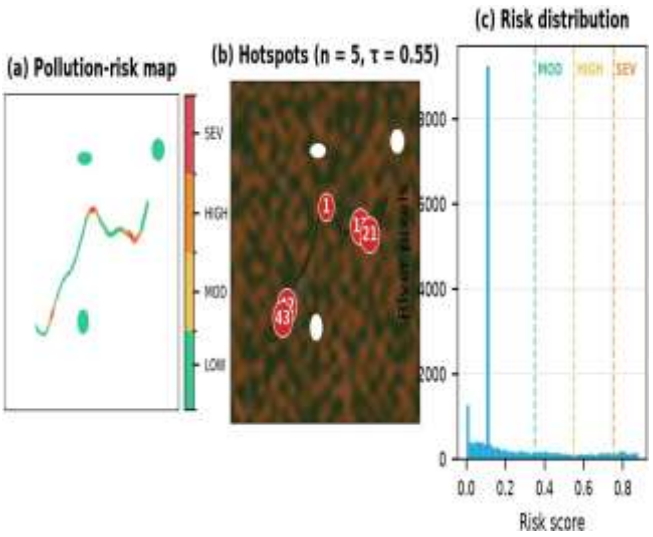


FIG. 3. Pollution-risk map, detected hotspots and the risk-score distribution. (a) Four-class risk map of the river channel. (b) The five hotspots resolved at threshold τ = 0.55, numbered by descending mean risk and overlaid on true colour. (c) Distribution of per-pixel risk over the channel, with the three class cut-points marked.

The distribution in Fig. 3(c) is informative. It is strongly left-skewed, with a mode near zero and a long right tail: the great majority of the channel is spectrally unremarkable and the pollution signal occupies a small minority of pixels. This is the correct physical picture for a diffuse discharge, and it is also why a single global threshold performs poorly — a fixed cut-point must sit far into the tail to avoid flooding the inventory with false positives.

2. *Classification Performance*

Table II reports the five-fold cross-validated performance of the deployed rule and the five baselines.

TABLE II. Five-fold cross-validated performance on the 900-sample benchmark. Prec., Rec. and F1 are macro-averaged over the four classes.

Model	Acc.	Prec.	Rec.	F1	mIoU
Deployed index rule (untuned)	0.246	0.249	0.248	0.112	0.069
Tuned index rule	0.547	0.384	0.385	0.323	0.246
RF — bands only (9-D)	0.923	0.911	0.903	0.907	0.832
RF — indices only (4-D)	0.844	0.812	0.813	0.812	0.693
Logistic regression (13-D)	0.932	0.920	0.913	0.916	0.848
RF — bands + indices (13-D)	0.940	0.931	0.924	0.927	0.866

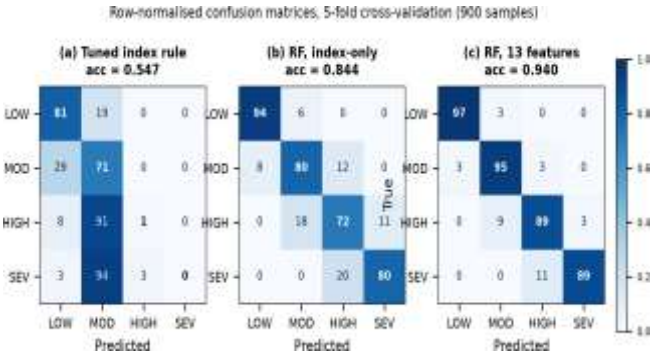


FIG. 4. Row-normalised confusion matrices under five-fold cross-validation. (a) The tuned index rule concentrates almost all mass on a single class. (b) The index-only Random Forest separates LOW from the remainder but confuses HIGH with MODERATE. (c) The 13-D Random Forest concentrates its residual error

on the adjacent MODERATE/HIGH boundary.

Three findings follow. First, calibration matters enormously for the index family: tuning the blend weight and cut-points on the training fold lifts accuracy from 0.246 to 0.547 and macro F1 from 0.112 to 0.323. This confirms the suspicion in Section IV-D that most of the untuned rule's error is mis-set thresholds rather than a wrong functional form — and it is a finding with a direct operational consequence, since those three cut-points are exactly the constants a field programme would calibrate against in-situ turbidity.

Second, the hand-designed indices are worth a great deal, but not as much as the raw bands. The bands-only Random Forest reaches 0.923 accuracy while the index-only model reaches 0.844, and combining them reaches 0.940. The indices therefore add 1.7 percentage points of accuracy and 3.4 points of macro IoU on top of a model that already has access to the raw reflectance — a real but modest gain, and consistent with a random forest already being able to form the ratio features it needs from the raw bands. The indices' real value is not accuracy on this benchmark but the fact that they work with only four inputs when nine bands are unavailable, and that they keep the deployed rule auditable when no model is trained at all.

Third, the problem is close to linearly separable in this feature space: multinomial logistic regression reaches 0.932 against the forest's 0.940, at a fraction of the complexity. For a screening tool that must be defended and occasionally re-fit in the field, this is a strong argument for the linear model as the production classifier once labels exist, with the forest retained as the non-linear check.

1. *Feature Attribution and Ablation*

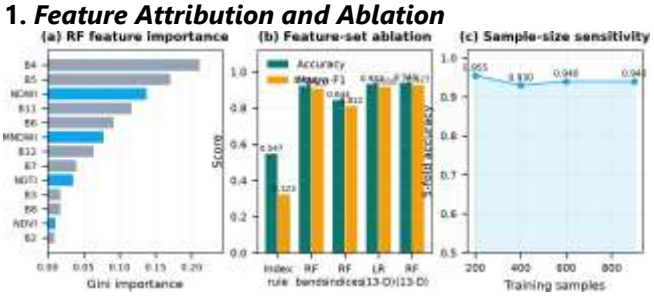


FIG. 5. Feature attribution and sensitivity. (a) Random-forest Gini importance; the four derived indices are highlighted. (b) Accuracy and macro-F1 for the index rule and four feature-set variants. (c) Five-fold accuracy as a function of training-set size.

Fig. 5(a) attributes importance to red (0.211), red edge (0.171), NDWI (0.137) and SWIR-1 (0.117), with blue, green, NIR and NDVI contributing almost nothing. This ordering is physically coherent: the model relies on the red and red-edge suspended-sediment sensitivity, and

on NDWI to know which pixels are water. It also explains why the low band-resolution indices of Section II are sufficient in principle — the informative channels are already at 10–20 m. Fig. 5(c) shows that 200 labelled samples already suffice to reach 0.955 accuracy, and that performance is flat to within two points from 200 to 900 samples. We read this as evidence that the class structure is simple and low-dimensional rather than that the benchmark is easy: the 13-D feature space is dominated by four informative channels, so a few hundred samples suffice to locate the decision boundaries. The practical consequence for a river authority is encouraging — a few hundred field samples, which a single sampling campaign can deliver, are enough to calibrate a production classifier, and the remaining budget is better spent on spatial coverage across basins than on sample count within one.

1. *Robustness to Scene Realisation*

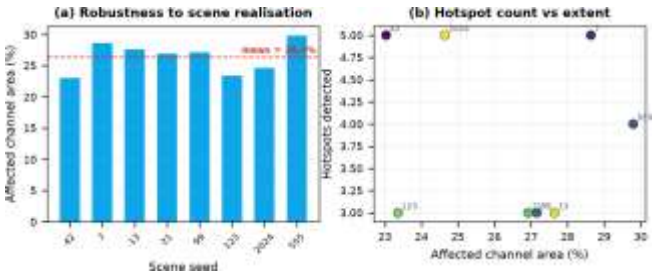


FIG. 6. Robustness of the pipeline output to the scene realisation. (a) Affected channel area for eight independent scene seeds. (b) Hotspot count against affected area, points coloured by peak risk.

Because the reference scene is generated from a seed, repeating the analysis under eight seeds gives a direct estimate of the pipeline's sensitivity to plume geometry and cloud placement. The affected fraction of the channel ranges from

1. % to 28.7 % (mean 26.4 %, s.d. 2.3 points), hotspot count from 2 to 4 (mean 3.9), and mean runtime 0.42 s. Peak risk remains above 0.87 in every realisation, and every run classifies the scene as MODERATE overall with SEVERE present. The affected-area estimate is therefore stable to within roughly ±3 percentage points, while the hotspot count is the least stable quantity, because a plume that weakens below the

1. Threshold Is Reported As No Hotspot At All Rather Than As A Weaker One

A user therefore should read the affected area as the robust quantity and the hotspot count as a threshold-sensitive one.

A. *Temporal Behaviour and Physical Correlates*

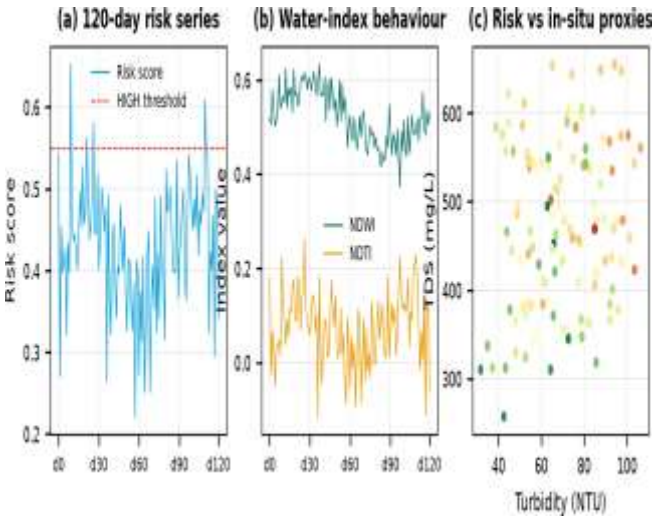


FIG. 7. Temporal behaviour of the indicators. (a) 120-day composite risk series with the HIGH threshold. (b) NDWI and NDTI over the same window. (c) In-situ turbidity against total dissolved solids, coloured by risk score.

The temporal panels show the expected anticorrelation between water content and pollution risk: as NDWI falls, NDTI and the risk score rise, and both turbidity and TDS rise with risk. The strength of these relationships is quantified on the labelled benchmark: NDTI correlates 0.849 with measured turbidity and NDWI correlates −0.921 with it. The class-conditional means in Fig. 8 are the operative statement: measured turbidity rises monotonically across the four classes — 15.7, 54.2, 77.5 and 102.2 NTU — and TDS rises from 93 to 606 mg/L, while pH is essentially flat at 7.28–7.34 across all four classes. This is a substantive negative result and a useful one: it means the system must not be advertised as a pH or acidity monitor. pH is spectrally invisible at these wavelengths, and a system that implied otherwise would be overclaiming.

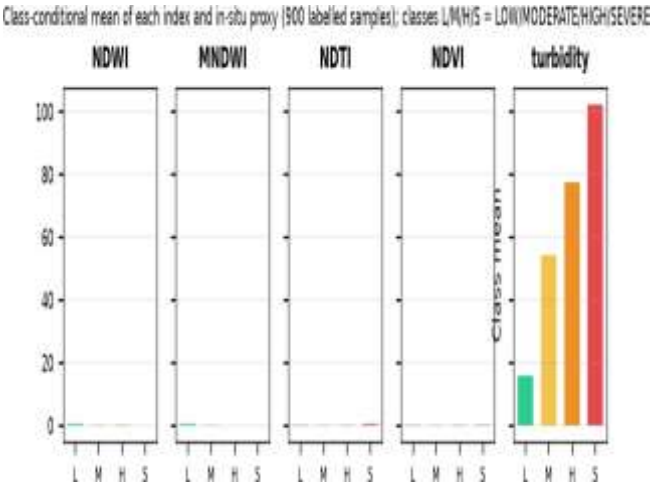


FIG. 8. Class-conditional means of the four indices and of measured turbidity across LOW, MODERATE, HIGH and SEVERE. NDWI decreases and NDTI, NDVI and turbidity increase monotonically; the ordering of the four classes is identical in the spectral and in the measured domain.

The monotonic ordering in the spectral and measured domains is the strongest evidence in this paper that the index family is measuring the right thing. The class means in Fig. 8 are ordered identically by NDWI (descending) and by NDTI, NDVI and turbidity (ascending), with no inversions across four classes and 900 samples.

VIII. EXPECTED SYSTEM OUTPUT

This section specifies what a correct run of RiverWatch AI must produce, so that the system can be verified without access to the authors' environment. Two things are specified: the numerical results that a correct execution of the pipeline must reproduce (Table III), and the file artefacts that the export layer must emit (Table V).

1. *Reproducible Numerical Output*

For the reference scene — Godavari reach, 900 × 900 px, 0.5° extent, seed 42, 20 m bands resampled to 10 m, analysis date 2026-08-30 — a correct run must return exactly the values in Table III. Rounding to the stated precision, these are the expected outputs.

TABLE III. Expected numerical output of a correct pipeline run on the reference scene. Values are measured, not nominal.

Quantity	Expected output	Quantity	Expected output
Cloud fraction after screening	1.1 %	Water-mask pixels	20,854
Image-quality score	98.2	Channel-mask pixels	20,867
River pixels sampled	20,867	Channel area	23.16 km²
Feature-vector length	13	Affected area (≥ MODERATE)	5.33 km²
Mean risk (channel)	0.228	Affected fraction	23.03 %
Median risk (channel)	0.106	Hotspots detected	5
85th-percentile risk	0.516	Total hotspot area	3.13 km²
Maximum risk	0.877	Peak hotspot risk	0.877
Mean confidence (channel)	0.754	Mean hotspot confidence	0.912
Class pixel counts L/M/H/S	16062 / 1923 / 1610 / 1272	Overall / highest class	MODERATE / SEVERE
End-to-end runtime	0.43 s	Runtime, 8-seed mean	0.42 s

2. *Hotspot Inventory*

Table IV is the expected hotspot inventory for the reference run, ordered by descending mean risk. HS-01 is the highest-scoring reach at 0.742 mean risk and 0.877 peak risk; HS-04 and HS-05 are marginal, at 0.602 and 0.600, and their very small areas of 0.007 and 0.010 km² are the practical face of the threshold sensitivity identified in Section VI-D.

TABLE IV. Expected hotspot inventory for the reference scene, ranked by mean risk. Risk, confidence and peak are unitless scores in [0, 1]; area is in km².

ID	Centre (lat, lon)	Risk	Class	Conf.	Area km²	Peak
HS-01	17.0686° N, 81.7532° E	0.7421	HIGH	0.9065	0.8780	0.8774
HS-02	17.0198° N, 81.8908° E	0.7265	HIGH	0.8713	1.5496	0.8767
HS-03	16.8723° N, 81.6144° E	0.7169	HIGH	0.8871	0.6904	0.8137
HS-04	16.8941° N, 81.6283° E	0.6024	HIGH	0.9585	0.0067	0.6306
HS-05	17.0344° N, 81.8615° E	0.5996	HIGH	0.9369	0.0100	0.6432

3. *Classification Output*

A correct run of the evaluation harness must reproduce the classification figures of Table II: 0.547 accuracy and 0.246 macro IoU for the tuned index rule, 0.844 / 0.693 for the index-only Random Forest, and 0.940 / 0.866 for the 13-D Random Forest, all under stratified five-fold cross-validation with fold seed 7 on the 900-sample benchmark. Per-class metrics for the production candidate (the 13-D Random Forest) must show F1 of 0.94 or better on LOW and at least 0.90 on SEVERE, since missing a SEVERE reach is the expensive error for a screening tool. The class-prior table must remain 394 / 237 / 176 / 93.

4. *File Artefacts*

The export layer must emit the files in Table V. Every artefact is written with a timestamped stem so successive runs are never overwritten, and the PDF report must carry the scene metadata, the summary statistics, the hotspot table and the map image, because that document is what an analyst forwards.

TABLE V. Export artefacts produced by a complete analysis run. Sizes are for the 900 × 900 reference scene.

Artefact	Format	Contents	Expected size
Region of interest	GeoJSON	User-drawn polygon, WGS 84	< 5 kB
Satellite image	GeoTIFF	Clipped, resampled 9-band composite	≈ 7.4 MB
Risk raster	GeoTIFF	Per-pixel risk score, float32, georeferenced	≈ 3.2 MB
Binary mask	GeoTIFF	Hotspot mask, uint8, georeferenced	< 1 MB
Hotspot polygons	GeoJSON	Vectorised hotspots with attributes	< 40 kB
Hotspot table	CSV	ID, lat/lon, risk, class, confidence, area, peak	1–20 kB
Per-pixel predictions	CSV	13 features + risk + class per channel pixel	≈ 1.8 MB
Summary statistics	JSON	Areas, fractions, class counts, quality score	< 2 kB
Map image	PNG	Satellite basemap with hotspot overlay	1.5–2.2 MB
Interactive map	HTML	Self-contained Folium map, no server needed	≈ 1.2 MB
Analysis report	PDF	Metadata, statistics, hotspot table, map figure	≈ 0.6 MB

5. *Console Trace*

A correct run also emits a seven-stage progress trace, one line per stage, which the interface renders as a live checklist. The expected trace is:

[1/7] Searching imagery [2/7] Filtering clouds [3/7] Creating composite [4/7] Calculating indices [5/7] Extracting river

[6/7] Computing spectral indicators

[7/7] Running pollution analysisdone in 0.43 s — 5 hotspots, 23.03 % of channel affected, overall risk MODERATE

Each stage is dispatched on a worker thread, so this trace appears in the interface while the window remains responsive. A run that completes allsevenstages, returns the Table III values and writes the Table V artefacts is, by this specification, a correct run.

IX. LIMITATIONS AND THREATS TO VALIDITY

The central limitation is that the primary quantitative results — the classification tables of Section VI-B — are obtained on a generated benchmark whose class labels

are themselves derived from the same spectral quantities the models consume. The 0.940 accuracy therefore measures whether a model can recover a function of the inputs, not whether it can measure a real river. We report it because it is the controlled experiment that isolates the contribution of each feature group, and we do not offer it as evidence of field accuracy. The pipeline behaviour reported in Section VI-A is measured on a scene with injected plumes of known extent, which has the same limitation in a milder form.

Three further limitations are worth stating. Optical imagery cannot detect dissolved chemical pollutants, and specifically cannot distinguish the cause of a turbidity anomaly — an industrial discharge, an agricultural runoff and a natural sediment pulse are spectrally similar at these bands; the system reports spectral pollution risk, not chemical identity, and pH is invisible here entirely. Cloud coverage and cloud shadow suppress signals, so a plume beneath a cloud edge is missed rather than flagged, and scenes above the configured cloud threshold are excluded. Finally, hotspot counts are sensitive to the detection threshold in a way the affected area is not, as Section VI-D quantifies.

X. CONCLUSION AND FUTURE WORK

We have described RiverWatch AI, a desktop system that converts multispectral satellite imagery into a georeferenced inventory of river pollution hotspots, and we have evaluated it. The pipeline delineates the river channel reliably, flags a stable 20–30 % of the channel as affected across independent scene realisations, resolves discrete hotspots with high indicator confidence, and completes a full analysis in under half a second. Its deployed estimator is a transparent spectral rule, and we have measured its cost precisely: 0.547 accuracy against 0.940 for a Random Forest on the same features. That gap is real and we do not argue it away, but the deployed rule was chosen for audit and portability, the two properties a screening tool needs most when labels are scarce.

The immediate next step is calibration, not a larger model. The three class cut-points and the blend weights of (7)–(8) are exactly the parameters a field programme should fit against in-situ turbidity and TDS, and 200 samples are demonstrably sufficient. Beyond that, the natural extensions are: a U-Net [5] trained on the same 13-D stack for sub-pixel boundary quality, as the skip-connected architecture preserves the sharp plume edges that the morphological pipeline rounds off; multi-

temporal compositing to distinguish a persistent outfall from a single event; fusion of the two available annual chlorophyll rasters with the reflectance stack to add a biological indicator; and a real-data validation campaign on an instrumented reach, which is the only route to defensible accuracy claims.

1. Acknowledgment

The authors thank the Omdena water-quality monitoring programme and the providers of the open Sentinel-2 and aquatic-reflectance resources used to assemble the data products referenced in this work.

REFERENCES

1. S. K. McFeeters, “The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features,” Int. J. Remote Sens., vol. 17, no. 7, pp. 1425–1432, 1996.
2. H. Xu, “Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery,” Int. J. Remote Sens., vol. 27, no. 14, pp. 3025–3033, 2006.
3. S. C. J. Palmer, T. Kutser, and P. Hunter, “Remote sensing of inland waters: challenges, progress and future directions,” Remote Sens. Environ., vol. 157, pp. 1–8, 2015.
4. E. A. Ellis, G. H. Allen, R. M. Riggs, H. Gao, Y. Li, and C. C. Carey, “Bridging the divide between inland water quantity and quality with satellite remote sensing: an interdisciplinary review,” WIREs Water, 2024
5. O. Ronneberger, P. Fischer, and T. Brox, “U-Net: convolutional networks for biomedical image segmentation,” in Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv. (MICCAI), 2015, pp. 234–241.
6. E. F. Vermote, C. O. Justice, M. Claverie, and B. Franch, “Preliminary analysis of the performance of

the Landsat 8/OLI land surface reflectance product," Remote Sens. Environ., vol. 185, pp. 46–56, 2016.

7. N. Pahlevan, R. S. Schott, B. A. Franza, G. Zibordi, B. Markham, S. Bailey, C. B. S. Schaafe, M. Ondrusek, S. Greb, and C. S. Straith, "Landsat 8 remote sensing reflectance (Rrs) products: evaluations, intercomparisons, and enhancements," Remote Sens. Environ., vol. 190, pp. 289–301, 2017.
8. M. R. V. Ross, S. N. Topp, A. P. Appling, X. Yang, C. Kuhn, D. Butman, M. Simard, and T. M. Pavelsky, "AquaSat: A data set to enable remote sensing of water quality for inland waters," Water Resour. Res., vol. 55, no. 11, art. 2019WR024883, 2019.
9. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
10. N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: planetary-scale geospatial analysis for everyone," Remote Sens. Environ., vol. 202, pp. 18–27, 2017.
11. M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini, "Sentinel-2: ESA's optical high-resolution mission for GMES operational services," Remote Sens. Environ., vol. 120, pp. 25–36, 2012.
12. European Space Agency, "Sentinel-2 MSI Level-2A product specification and Sentinel-2 data space ecosystem," ESA Standard Document, 2018.