

High-Resolution Water Segmentation for Salmon Habitat

Luana Marotti
MSc Data Science
Queen Mary University of London
London, United Kingdom
l.marotti@se25.qmul.ac.uk

Student Number: 250016521
Supervisor: Dr. Iran Roman
School of Electronic Engineering
and Computer Science
Queen Mary University of London

Abstract—Automated detection of wetted surface area in high-resolution satellite imagery is important for salmon habitat monitoring but remains challenging because shadows, sunglint, rocky surfaces, and turbid water can be spectrally confused with open water. This study compares three water segmentation approaches — an index-based method (NDWI), a pretrained deep learning model (OmniWaterMask), and a domain-specific fine-tuned U-Net — on a dataset of 64 high-resolution WorldView-2/3 scenes of Pacific Northwest salmon-rearing rivers, divided into 17,445 image chips of 512×512 pixels using scene-aware development and test splits. On 4,788 held-out test chips, NDWI achieved micro F1=0.523 with a threshold selected on development data, while OmniWaterMask (OWM) achieved micro F1=0.344 despite very high recall, reflecting severe overprediction. A U-Net with a ResNet34 encoder, Focal+Dice loss, learning rate 10^{-4} , batch size 64, and gradient clipping achieved a mean validation loss of 0.056 ± 0.016 across five folds with no training collapses. The best single-fold model achieved micro F1=0.841, while a five-fold majority-vote ensemble achieved the highest observed micro F1=0.863. These correspond to improvements of +0.318 and +0.340 over NDWI, respectively, showing that domain-specific supervised learning substantially outperforms both a spectral baseline and a domain-mismatched pretrained baseline (OWM) on this dataset.

Index Terms—water detection, semantic segmentation, U-Net, WorldView-2/3, NDWI, salmon habitat, wetted surface area, remote sensing

I. INTRODUCTION

Monitoring salmon habitat in Pacific Northwest rivers requires accurate, repeated measurement of wetted surface area — the spatial extent of water-covered ground at a given moment. Juvenile salmon depend on narrow side channels, beaver ponds, and shallow pools for feeding, shelter, and thermal refugia (Siegel and Crozier, 2018), and shifts in wetted extent driven by drought, flow variability, or restoration interventions have direct consequences for habitat capacity and population outcomes. These fine-scale features span only a few metres and are invisible at the 10–30 m resolution of freely available sensors such as Sentinel-2 and Landsat, making sub-metre imagery essential for reliable habitat mapping. Automated detection of wetted area from such imagery would directly support the National Oceanic and Atmospheric Administration (NOAA)’s streamflow-to-habitat modelling workflows.

WorldView-2 and WorldView-3, at approximately 0.50 m ground sampling distance, can resolve the narrow channels and

small water bodies relevant to salmon habitat. At this resolution, however, shadows cast by riparian vegetation, sunglint from water surfaces, rocky surfaces, and turbid or shallow water all produce spectral responses that overlap with clear open water — making automated water detection substantially harder than at coarser resolutions, where such confounders are averaged out across larger pixels.

Existing approaches fall into three broad categories. Index-based methods such as the Normalised Difference Water Index (NDWI) (McFeeters, 1996) exploit spectral contrasts without requiring labelled data, but rely on a single global threshold that cannot resolve the spectral ambiguities present in sub-metre mountain river imagery. Pretrained deep learning models such as OmniWaterMask (OWM) (Wright et al., 2025) offer broader applicability across sensors, but were developed for different domains, raising questions about transferability to WorldView imagery of Pacific Northwest rivers. Supervised deep learning, by contrast, can exploit spatial context and be fine-tuned to the specific spectral and geometric characteristics of the target domain — but requires careful design choices around class imbalance, training stability, and evaluation methodology.

U-Net (Ronneberger et al., 2015), an encoder-decoder architecture with skip connections that preserve spatial detail at multiple scales, has demonstrated strong performance across remote sensing segmentation tasks and is well-suited to detecting narrow river channels. Combined with a ResNet34 encoder (He et al., 2016) pretrained on ImageNet, it provides a practical and reproducible supervised baseline for this domain.

This dissertation addresses the following research question: *can a fine-tuned U-Net model provide a robust solution for water segmentation in high-resolution WorldView imagery of Pacific Northwest salmon rivers?* We make three contributions: (1) a systematic comparison of three water detection approaches on a challenging high-resolution dataset: an index-based method (NDWI), a pretrained deep learning model (OWM), and a domain-specific fine-tuned U-Net; (2) identification of a training configuration that eliminated the observed instability, although the individual contributions of batch size and gradient clipping were not isolated; and (3) evidence that domain-specific supervised learning substantially outperforms both a spectral baseline and a domain-mismatched pretrained

baseline (OWM) on this dataset, with implications for how automated habitat monitoring systems should be designed.

II. RELATED WORK

Water detection from remote sensing imagery has been approached through three broad paradigms: spectral indices, pretrained generalised models, and supervised deep learning. Spectral indices are transferable but rely on global thresholds that fail in spectrally complex scenes. Pretrained models offer broad coverage but suffer from domain shift when applied to sensors and environments outside their training distribution. Supervised deep learning achieves strong performance where labelled data exist, but requires careful design choices around class imbalance, training stability, and threshold selection.

A. Index-Based Water Detection

Index-based methods exploit spectral contrasts between water and surrounding land cover without requiring labelled training data. McFeeters (1996) proposed NDWI, using the ratio of green and near-infrared reflectance to highlight open water while suppressing vegetation. NDWI remains widely used due to its computational simplicity and transferability across sensors. Threshold-based indices share a fundamental limitation: they rely on a single value applied globally across a scene. As McFeeters (1996) noted, threshold selection is scene-dependent and no single value generalises across acquisition conditions. At sub-metre WorldView resolution, this problem is compounded — shadows cast by riparian trees and valley walls, wet gravel bars, and rocky surfaces all produce spectral responses that overlap with open water in the green–NIR space (Dare, 2005). In this study, NDWI achieves micro F1=0.523 when the threshold is selected on development data. Although these methods remain attractive because they require no labelled training data, their reliance on a global threshold fundamentally limits their applicability in spectrally heterogeneous, high-resolution river environments. This limitation motivates the transition towards supervised learning approaches that can exploit spatial context rather than relying solely on spectral information.

B. Deep Learning for Water Segmentation

Supervised deep learning has largely superseded index-based approaches where labelled training data are available. Long et al. (2015) established that fully convolutional networks pretrained on ImageNet could be adapted for semantic segmentation with limited labelled data. U-Net (Ronneberger et al., 2015) extended this with skip connections between encoder and decoder stages, preserving fine spatial detail at multiple scales — a critical property for narrow river channels that may span only a few pixels even at 0.5 m resolution. Wright et al. (2025) developed OmniWaterMask (OWM), a sensor-agnostic approach combining a U-Net segmentation model with NDWI and optional vector calibration data (OpenStreetMap (OSM) and National Hydrography Dataset (NHD)) to improve robustness across sensors and scales. OWM was trained on the FLAIR #1 French aerial dataset

(0.2 m) and the S1S2-Water global Sentinel dataset (10 m). Notably, OWM was originally developed and validated for farm dam monitoring in Australia — a task characterised by compact, high-contrast water bodies rather than narrow river channels in complex mountain terrain. Its training distribution differs substantially from WorldView-2/3 Pacific Northwest river imagery in sensor type, platform geometry, target water body shape, and background complexity, making it a useful but inherently limited comparison point. Unlike NDWI, OWM attempts to improve transferability by combining deep learning with external vector information. However, because its training data differ substantially from the imagery considered in this work, its performance provides an opportunity to assess the extent to which general-purpose pretrained models can replace domain-specific supervised training. Beyond U-Net-based models, modern semantic segmentation has increasingly adopted architectures such as DeepLabV3+ (Chen et al., 2018), HRNet (Wang et al., 2021), SegFormer (Xie et al., 2021), and Swin Transformer-based encoders (Liu et al., 2021), several of which have shown strong performance in remote sensing applications. These architectures offer alternative mechanisms for incorporating multi-scale context, preserving high-resolution spatial detail, or modelling long-range dependencies. A full architecture comparison is outside the scope of this study. U-Net with a ResNet34 encoder was selected because it provides a well-established, computationally efficient baseline with strong performance in remote-sensing segmentation literature. Holding the architecture fixed allows the study to isolate the effects of domain-specific training and optimisation choices rather than differences in model complexity. Consequently, the central comparison focuses on domain-specific supervised learning versus spectral and pretrained water-detection approaches.

C. Class Imbalance and Loss Functions

Water detection in high-resolution river imagery is inherently class-imbalanced: water pixels are typically rare relative to background land cover. Unweighted binary cross-entropy can be sensitive to severe class imbalance because the large number of background pixels may dominate the optimisation objective (Milletari et al., 2016). Dice loss (Milletari et al., 2016) addresses this by directly optimising foreground overlap, reducing sensitivity to severe class imbalance. Salehi et al. (2017) generalised Dice to the Tversky index, introducing asymmetric weighting that allows independent control of penalties for false positives and false negatives. Lin et al. (2017) proposed Focal loss, which down-weights easy, well-classified examples to focus learning on hard or ambiguous pixels. Combining Focal loss with Dice provides complementary benefits: Dice handles class frequency imbalance while Focal loss emphasises difficult examples.

D. High-Resolution Imagery and Confounders

The transition from moderate-resolution sensors (10–30 m) to very-high-resolution imagery (<1 m) introduces qualitatively different challenges. Dare (2005) demonstrated that

shadow detection in high-resolution satellite imagery is inherently ambiguous because shadow spectral responses depend on illumination geometry, surface composition, and sensor characteristics — none of which are captured by simple spectral indices. High-resolution WorldView imagery of Pacific Northwest rivers presents several of these confounders simultaneously: steep valley walls cast long shadows; riparian forests create heterogeneous shade patterns; wet gravel bars and turbid water are spectrally similar to shadowed land; and roads frequently run parallel to channels. Together, these challenges indicate that water segmentation in sub-metre river imagery remains an open problem. Existing index-based methods struggle with spectral ambiguity at sub-metre resolution, where shadows, wet gravel, and turbid water become indistinguishable from open water. Pretrained deep-learning models have largely been validated on different sensors and environments, leaving their transferability to this domain untested. It therefore remains an open question whether a supervised model trained specifically on high-resolution WorldView imagery of Pacific Northwest rivers can meet the accuracy requirements for operational habitat monitoring.

III. METHODOLOGY

This project follows a mixed experimental and design/build methodology. We design and implement a water-detection pipeline, establish index-based and pretrained baselines on a held-out test set, and develop a supervised segmentation model using scene-level cross-validation. All three methods are evaluated on the same held-out test set using pixel-wise metrics. Figure 1 summarises the experimental pipeline.

A. Data and Preprocessing

The dataset comprises 17,445 chips of 512×512 pixels derived from 64 WorldView-2/3 scenes covering Pacific Northwest salmon-rearing rivers, provided by NOAA. Each chip contains eight multispectral bands on a 0.50 m grid. Corresponding binary ground truth masks were manually digitised by NOAA domain experts. Chips were grouped by parent scene to prevent spatial data leakage between adjacent chips from the same scene. A scene-level split reserved 13 scenes (4,788 chips) as the held-out test set, never seen during training or validation. The remaining 51 scenes were used for training and cross-validation. Scenes span acquisitions from 2011 to 2026, covering summer (26 scenes), autumn (24 scenes), spring (11 scenes), and winter (3 scenes). The train/val and test splits are comparable in mean water fraction per scene (train/val: 0.019, test: 0.016), though the test set contains no high-water scenes (maximum scene-level water fraction 0.083 vs 0.219 in training). Two test scenes were acquired in 2012 and one in 2025 — years with no corresponding training scenes — providing a partial assessment of temporal generalisation. Four bands were selected for model input: Red (band 5), Green (band 3), Blue (band 2), and NIR1 (band 7). These approximate the RGB spectral range used in ImageNet pretraining while adding NIR1, which provides the strongest spectral contrast between water and vegetation. Since

ImageNet-pretrained encoders expect three input channels, the first convolutional layer of the ResNet34 encoder was modified to accept four channels, with the original RGB weights retained and the additional NIR1 channel weights initialised randomly. Per-band standardisation statistics were computed from the training set only and applied to all splits.

B. Baseline Methods

NDWI was computed as $(\text{Green} - \text{NIR1}) / (\text{Green} + \text{NIR1})$ on all 4,788 test chips. A threshold sweep on the 51 development scenes identified $t = 0.35$ as optimal (dev F1=0.547); applying this to the test set yields micro F1=0.523. OmniWaterMask (OWM) (Wright et al., 2025) was evaluated using scene reconstruction: chips were stitched into 13 full GeoTIFF scenes, OWM was run with OSM and NHD vector data, and predictions were cropped back to chip locations. OWM calibrates its internal threshold per scene using mapped water polygons from OSM and NHD. Narrow or unmapped channels absent from these vector datasets may not be well-represented in the calibration, potentially contributing to the high false-positive rate observed. Seven sets of dummy masks at fixed water fractions (0%, 20%, 40%, 50%, 60%, 80%, 100%) served as pipeline sanity checks.

C. U-Net Architecture and Training

We employ a U-Net (Ronneberger et al., 2015) with a ResNet34 encoder (He et al., 2016) pretrained on ImageNet, implemented via the segmentation-models-pytorch library. The encoder produces multi-scale feature maps at four resolutions; skip connections pass these to the decoder, which reconstructs the segmentation mask at the original resolution. Training used all water-containing chips plus a randomly sampled 20% of land-only chips (approximately 2,606 training chips per fold); the land sample was resampled independently at the start of each fold. On-the-fly augmentation was applied to training chips only, including horizontal/vertical flips, $90^\circ/180^\circ/270^\circ$ rotations, random crop-and-resize, per-band radiometric jitter, Gaussian noise, band dropout, shadow simulation, and sunglint simulation. Validation used unmodified original chips. A systematic instability investigation found that increasing batch size from 32 (82 batches/epoch) to 64 (41 batches/epoch) together with gradient clipping (max_norm=1.0) produced markedly more stable training. The final configuration uses AdamW optimiser (weight decay= 10^{-3}), cosine annealing LR schedule with 3-epoch warm-up, and early stopping on validation loss. Early stopping patience was set to 15 during the hyperparameter sweep to allow rapid comparison across configurations, and increased to 50 for the final five-fold cross-validation experiments to allow fuller convergence. Validation loss is used as the primary selection criterion throughout — for early stopping, hyperparameter selection, and final model checkpoint selection — because it is a smoother signal than validation F1 in the presence of class imbalance and threshold sensitivity. Architectural comparison was outside the scope of this study. The U-Net-ResNet34 architecture was held constant to isolate the effects of optimisation choices

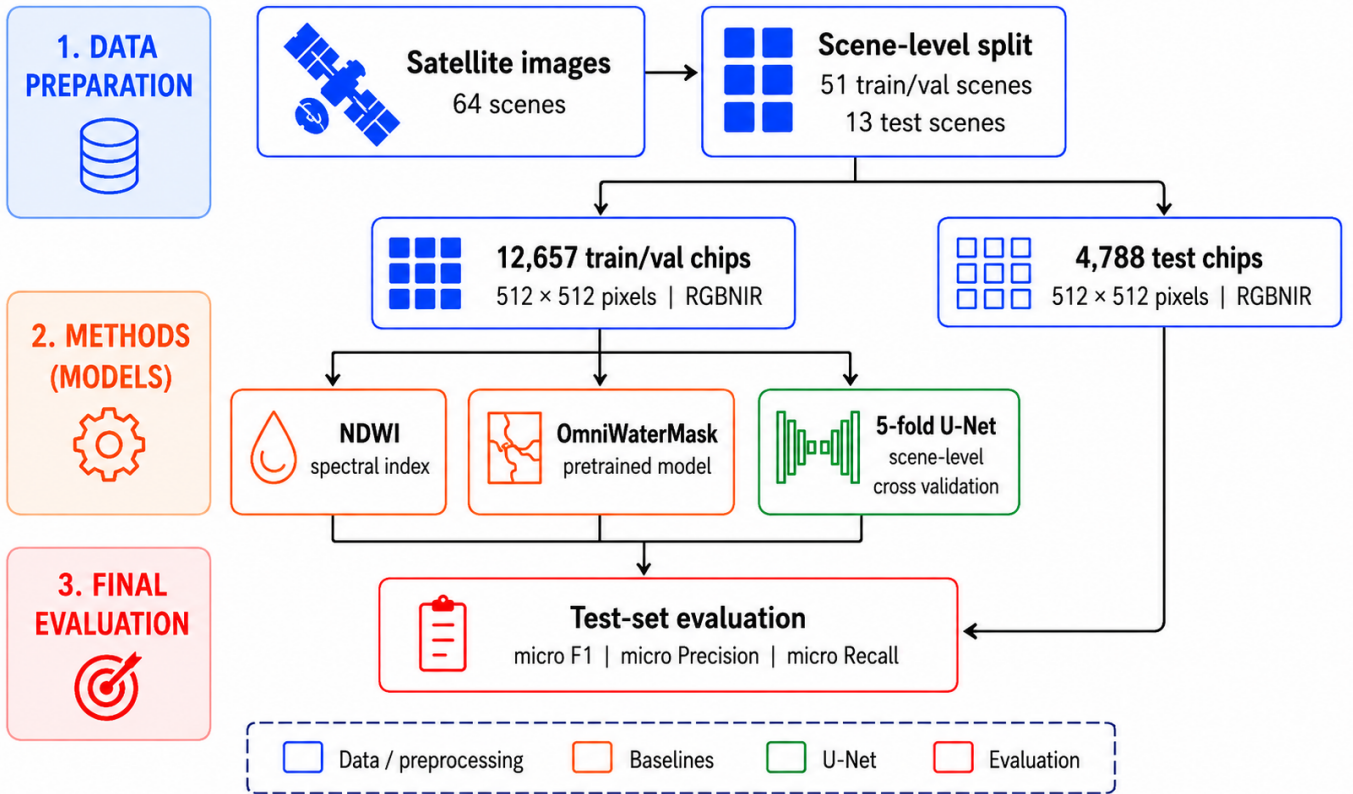


Fig. 1. Experimental pipeline overview. Data are split at the scene level into 51 train/val scenes and 13 held-out test scenes, chipped into 512×512 pixel tiles across 4 spectral bands, and standardised using training-set statistics. NDWI is computed per chip with a threshold selected on development scenes. OWM is evaluated using scene reconstruction with internal calibration. The U-Net is trained via 5-fold scene-level cross-validation; two inference strategies are evaluated: best single fold (global minimum validation loss) and majority-vote ensemble (each fold model applies its own val-selected threshold, $\geq 3/5$ predict water). All strategies are evaluated on the same 4,788 held-out test chips using micro F1, micro Precision, and micro Recall.

and to evaluate whether domain-specific supervised learning could outperform spectral and domain-mismatched pretrained baselines under a controlled architecture.

D. Hyperparameter Sweep

A two-round sweep on fold 1 identified the optimal training configuration. Round 1 fixed $LR=4 \times 10^{-4}$ and compared three loss functions: Tversky (Salehi et al., 2017), Dice (Milletari et al., 2016), and Focal+Dice — a weighted combination of Focal loss (Lin et al., 2017) and Dice loss (Milletari et al., 2016) implemented via the segmentation-models-pytorch library. Focal+Dice achieved the lowest validation loss (0.076 vs 0.102 for Tversky and 0.194 for Dice) and was selected for Round 2. Round 2 fixed Focal+Dice and compared four learning rates. Lower learning rates produced lower validation loss and more stable training; $LR=1 \times 10^{-4}$ was selected based on minimum validation loss (0.049 vs 0.069 for $LR=2 \times 10^{-4}$), consistent with the predefined selection criterion.

E. 5-Fold Cross-Validation and Test Set Evaluation

Five-fold cross-validation assesses the stability and expected generalisation of the selected configuration. Scenes are assigned to folds at the scene level to prevent spatial data leakage. Each fold uses an 80/20 scene-level train/validation split,

producing approximately 40 training scenes and 11 validation scenes. Training chips are drawn from the augmented pool described in Section III-C; validation uses unmodified original chips. Early stopping with patience=50 on validation loss determines the effective training duration. Minimum validation loss was used for checkpoint selection. Following cross-validation, two evaluation strategies are applied to the held-out test set. The *best single fold* strategy uses the global best checkpoint — the model achieving the lowest validation loss across all five folds — with a classification threshold selected via sweep on that fold’s validation scenes. The *ensemble* strategy combines all five fold models using majority voting: each model independently applies its own threshold, selected via sweep on its own validation scenes, to produce a binary prediction; a pixel is classified as water if three or more of the five models predict water. This approach ensures that every threshold is selected on validation data the corresponding model never trained on, eliminating leakage between training and threshold selection. Both strategies are evaluated once on the 4,788 held-out test chips.

F. Evaluation

All methods are evaluated on the same 4,788 held-out test chips using micro F1 as the primary metric, pooled across all

TABLE I

Summary of all results on 4,788 held-out test chips from 13 test scenes. All metrics are micro-averaged, pooled across all chips. Bold: best value per column.

Method	Micro F1	Micro P	Micro R
<i>U-Net test set evaluation</i>			
Best single fold	0.841	0.800	0.888
Ensemble (majority vote)	0.863	0.839	0.888
<i>Baselines</i>			
NDWI ($t=0.35$)	0.523	0.376	0.859
OWM	0.344	0.209	0.966
Dummy 100%	0.034	0.017	1.000

chips. Micro-averaging was selected because the operational objective is to assess total pixel-level water segmentation performance across the complete test set. A scene-macro average would assign equal weight to scenes with different numbers of evaluated pixels and substantially different water-pixel counts, and was therefore not used as the primary evaluation measure. Precision and recall are reported to characterise each method’s error profile. Classification thresholds are always selected on validation or development data, never on the held-out test set: the best single fold threshold via sweep on that fold’s validation scenes; each ensemble fold’s threshold via sweep on its own validation scenes; and the NDWI threshold via sweep on the 51 development scenes. OWM thresholds are calibrated internally using OSM and NHD vector data.

IV. RESULTS

A. Baseline Methods

NDWI achieves micro F1=0.523 when the threshold is selected on development data (Table I), already outperforming OWM (micro F1=0.344) by 18 points despite using only a spectral index. OWM’s very high recall (0.966) but low precision (0.209) indicates that it detects almost all water pixels but produces nearly four false positives for every true positive detected — a pattern consistent with severe overprediction. The dummy 100% mask, which labels every pixel as water, also achieves perfect recall but yields micro F1=0.034, confirming that high recall alone carries little diagnostic value under the severe class imbalance in the test set, where approximately 1.7% of pixels are water.

B. OWM Configuration Experiments

A systematic series of experiments across 31 development scenes tested vector source combinations (OSM only, NHD only, OSM+NHD), toggled individual pipeline components, isolated the OWM U-Net model from NDWI, and assessed vector data quality against ground truth. Across the configurations examined on 31 development scenes, none outperformed the manually thresholded NDWI baseline, and the OWM U-Net component contributed negligible value beyond the spectral index. These results suggest that simple configuration changes were insufficient to overcome the observed domain gap.

C. Hyperparameter Sweep

Round 1 compared three loss functions at $LR=4 \times 10^{-4}$. Focal+Dice achieved the lowest validation loss (0.076 vs 0.102 for Tversky and 0.194 for Dice) and was selected for Round 2. Dice reached its minimum earliest but to a considerably worse loss, suggesting that rapid convergence alone is not a reliable indicator of successful optimisation under severe class imbalance. Round 2 fixed Focal+Dice and compared four learning rates; validation loss decreased monotonically as LR decreased, and $LR=1 \times 10^{-4}$ achieved the lowest validation loss (0.049) and was selected as the final configuration. Together, these experiments show that the selected learning rate and loss configuration produced stable optimisation within the fixed U-Net–ResNet34 architecture. The absence of substantial divergence between training and validation curves suggests that no pronounced overfitting was evident for this fold, and provides indirect evidence that the instability encountered during earlier experiments was primarily an optimisation issue rather than an inherent limitation of the architecture.

D. Training Instability Investigation

Initial configurations using batch size 32 exhibited persistent oscillation: validation loss alternated between low and high values across consecutive epochs, with the model frequently collapsing to near-zero predictions. A systematic investigation across more than ten parallel experiments varied weight decay (0.0–0.1), augmentation intensity, loss function, learning rate, and batch size. Removing augmentation entirely produced identical instability, ruling it out as a cause. A separate BCE experiment produced markedly higher validation loss, consistent with threshold sensitivity under class imbalance. Increasing batch size from 32 (82 batches/epoch) to 64 (41 batches/epoch) combined with gradient clipping ($\text{max_norm}=1.0$) produced the first stably converging run. This behaviour is consistent with, but does not prove, the hypothesis that insufficient gradient averaging contributed to instability, though the individual effects of batch size and gradient clipping were not isolated because they were introduced jointly.

E. 5-Fold Cross-Validation Results

The model trains stably across all five folds with no collapses, achieving a mean validation loss of 0.056 ± 0.016 . Fold 5 achieves the lowest validation loss (0.037) and its checkpoint is used as the best single fold model. Variability in validation loss across folds reflects differences in scene composition rather than instability in the training procedure. Although all folds converged successfully, they did so at different rates and reached different validation losses. This variability is expected because each fold contains a different combination of river systems, acquisition dates, and water fractions. Folds containing a greater proportion of spectrally complex scenes or sparse-water environments are likely to present a more challenging optimisation problem, leading to higher validation losses despite identical training configurations. In practical terms, the observed variance across folds indicates that the

selected optimisation strategy behaves consistently despite differences in river morphology, acquisition conditions, and seasonal variability. Importantly, no fold exhibited unstable training or catastrophic failure, suggesting that the proposed optimisation strategy is robust across a range of scene characteristics.

F. Test Set Results

The ensemble majority vote achieved the highest observed micro F1 = 0.863, compared with micro F1 = 0.841 for the best single-fold model (Table I). Both strategies clearly exceed NDWI (micro F1 = 0.523) and OWM (micro F1 = 0.344). The ensemble’s higher precision (0.839 vs 0.800 for the best single fold) and comparable recall (0.888 vs 0.888) indicates that the corrected per-fold thresholds produced a more conservative ensemble with fewer false positives. This behaviour is expected because each fold has learned slightly different decision boundaries from different subsets of training scenes. Majority voting suppresses predictions that are specific to individual folds while preserving consistent water detections. The resulting improvement in precision therefore reflects a reduction in fold-specific false positives rather than a fundamental change in model behaviour.

The difference between the two strategies is modest (+0.022 micro F1); the largest performance gains arise from task-specific supervised learning rather than from ensembling, as both strategies clearly surpass the NDWI and OWM baselines.

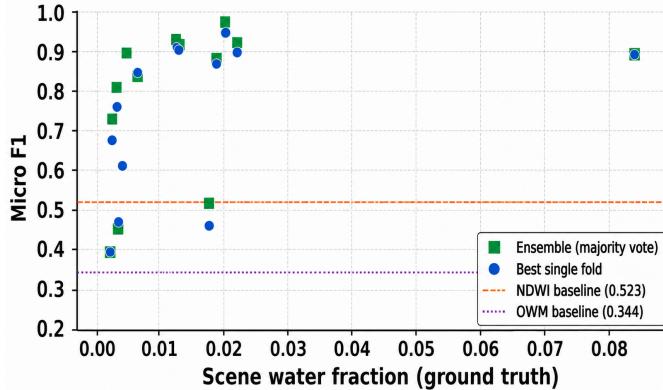


Fig. 2. Scene water fraction (ground truth) against per-scene micro F1 for both evaluation strategies across all 13 test scenes. Lines connect the same scene across strategies. The ensemble (green squares) consistently matches or outperforms the best single fold (blue circles). Water fraction is strongly associated with performance: the lowest-performing scenes have water fractions below 0.4%, while scenes above 1.0% water fraction generally achieve micro F1 above 0.85. The aggregate NDWI and OWM test-set micro F1 scores are shown as horizontal reference lines.

Per-scene analysis (Figure 2) reveals that water fraction shows the strongest observed association with test performance: the two worst-performing scenes (ensemble micro F1 = 0.389 and 0.454, water fractions below 0.4%) contrast sharply with the best-performing scene (micro F1 = 0.966, water fraction 2.0%). This pattern holds across both evaluation strategies, suggesting that extremely sparse water targets remain the primary challenge.

More broadly, these findings suggest that, within the methods evaluated in this study, domain-specific supervised training provided a larger benefit than the tested domain-mismatched pretrained model. This supports the continued value of domain-specific fine-tuning for specialised remote-sensing applications.

Figure 3 shows qualitative predictions for representative scenes.

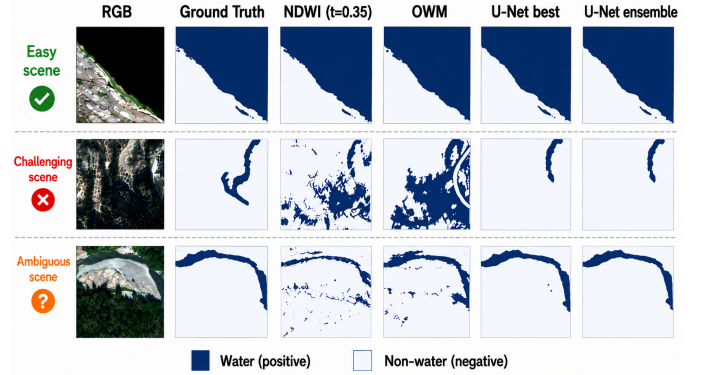


Fig. 3. Qualitative predictions for three representative test scenes. Columns: RGB chip, ground truth mask, NDWI ($t=0.35$), OWM, U-Net best fold, U-Net ensemble. Top row (easy scene, micro F1 = 0.966): the ensemble correctly delineates the main channel with minimal false positives. Middle row (challenging scene, micro F1 = 0.389): all methods struggle with ultra-sparse water targets. Bottom row (ambiguous scene, micro F1 = 0.735): NDWI and OWM produce substantial false positives from shadows and rocky surfaces; the U-Net ensemble correctly suppresses most of these ambiguous surfaces.

These results suggest that foreground sparsity is the main observed performance bottleneck within the evaluated architecture: when sufficient water pixels are present, the model performs well, whereas performance deteriorates when water is extremely sparse. Low-flow conditions — where river channels fragment into narrow pools occupying only a small fraction of the image — are precisely where performance degrades most. Since these conditions are also the most ecologically relevant for salmon habitat monitoring, improving performance under sparse-water conditions is the highest-priority direction for future work.

The qualitative examples in Figure 3 reveal that errors are concentrated in a small number of challenging scenarios rather than being uniformly distributed. Most false positives arise from deep terrain shadows and shaded riparian vegetation, while false negatives primarily occur in extremely narrow channels or isolated water patches. The proposed model therefore exhibits a broadly balanced error profile: commission errors are largely restricted to visually ambiguous land-cover types, whereas omission errors are concentrated in the most spatially fragmented water bodies. False positives may slightly overestimate wetted surface area, whereas false negatives risk omitting small habitat features from downstream ecological analyses, suggesting that future work should prioritise reducing omission errors in narrow channels while maintaining the model’s strong ability to suppress shadow-related false detections.

V. DISCUSSION AND CONCLUSIONS

The baseline results indicate that NDWI and OWM have substantial limitations for operational water segmentation in sub-metre river imagery. NDWI’s failure is primarily spectral: at 0.5 m resolution, shadows, wet gravel bars, and rocky surfaces become spectrally indistinguishable from open water, and no fixed threshold can resolve this ambiguity. OWM’s failure is primarily distributional: its training data differ substantially in sensor type, resolution, and landscape characteristics from the high-resolution WorldView imagery of Pacific Northwest river scenes used in this study, and systematic experiments across 31 development scenes suggested that simple configuration changes were insufficient to overcome the observed domain mismatch. It is worth noting that this comparison is inherently asymmetric — OWM was designed for a different sensor, environment, and water-body type, so its underperformance here is evidence of a specific domain gap rather than evidence against pretrained approaches in general. In this setting, a model trained specifically on the target data outperforms one trained on different sensors and environments. The qualitative predictions in Figure 3 make this contrast visible: NDWI and OWM produce large false-positive regions from shadows and rocky surfaces that the task-specific U-Net correctly suppresses. Taken together, these results address the research question posed in Section I by demonstrating that a fine-tuned U-Net provides a promising segmentation solution supported by a robust training configuration, with consistent performance across held-out scenes spanning diverse river systems, seasons, and acquisition years within the study area. The study therefore contributes both an empirical comparison of existing approaches and a validated training strategy for stable supervised segmentation under severe class imbalance. The instability investigation indicated that the combination of batch size 64 and gradient clipping was associated with stable optimisation in this small, class-imbalanced setting. Batch size 32 consistently produced oscillating validation loss and frequent collapses regardless of loss function, augmentation, weight decay, or learning rate. The first stable run required batch size 64 combined with gradient clipping, consistent with, but does not prove, the hypothesis that larger batches provide more stable gradient estimates when the foreground class is sparse. Because both changes were introduced jointly, their individual contributions cannot be quantified, and future work should isolate each factor.

The U-Net ensemble achieved micro F1 = 0.863, a +0.340 gain over the NDWI baseline and +0.519 over OWM. It edged the best single fold by +0.022 micro F1, with higher precision (0.839 vs 0.800) and comparable recall (0.888 vs 0.888), consistent with the corrected per-fold thresholds producing a more conservative ensemble with fewer false positives. It is worth noting that the headline micro F1 reflects a pooled aggregate across all 13 test scenes; as Figure 2 shows, a substantial share of this aggregate performance is driven by well-watered scenes where both strategies achieve F1 above 0.85, while the lowest-performing scenes are consistently

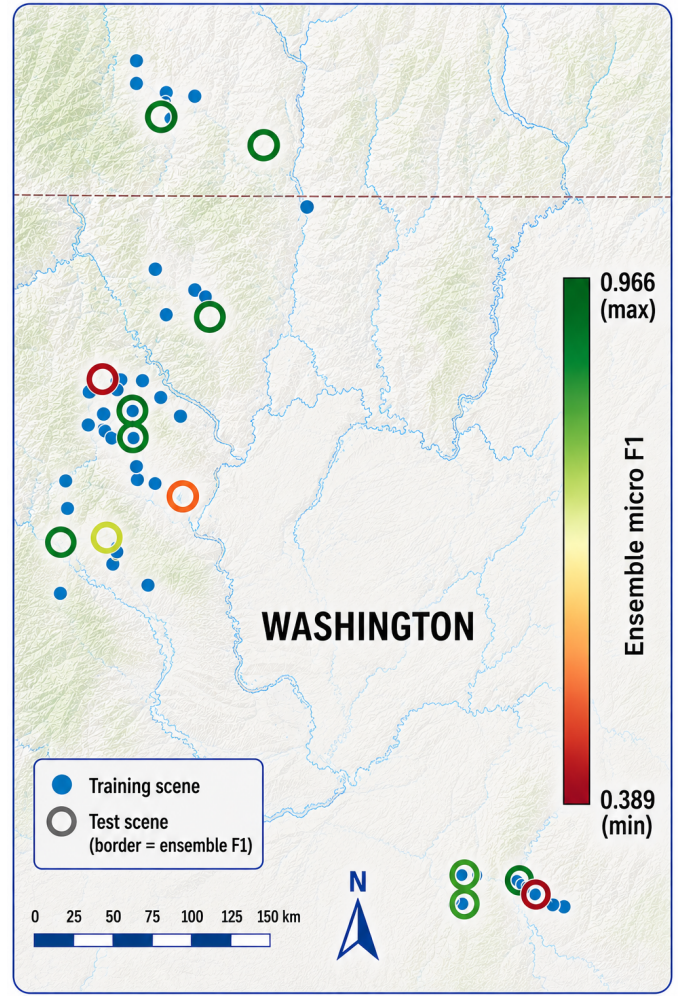


Fig. 4. Geographic distribution of training and test scenes across central Washington State and southern British Columbia. Training scenes are shown as filled blue circles; test scenes as hollow circles with borders coloured by ensemble micro F1 (colour scale spans from the minimum to maximum observed ensemble micro F1).

those with very sparse water coverage. The aggregate metric therefore somewhat understates the difficulty of the task for low-flow conditions. The per-scene pattern (Figure 2) is clear: performance is strongly driven by water fraction, with scenes below 0.4% water fraction remaining consistently difficult for both strategies. This suggests that data composition and foreground sparsity are substantial contributors to the remaining performance limitations. Consequently, future improvements may come from better data curation, targeted sampling of difficult examples, and improved annotation strategies, alongside further testing of alternative segmentation architectures. The additional context provided by Figures 4 and 5 suggests that the remaining performance variability cannot be explained by water fraction alone. Figure 4 indicates that both high- and low-performing scenes are geographically distributed across the study area rather than concentrated in a single watershed, providing some evidence that the model generalises across multiple river systems within the Pacific Northwest.

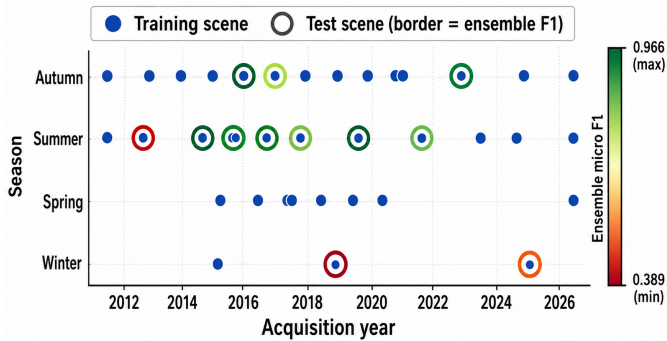


Fig. 5. Acquisition timeline showing acquisition year and season for all scenes. Training scenes are shown as filled blue circles; test scenes as hollow circles with borders coloured by ensemble micro F1 (colour scale spans from the minimum to maximum observed ensemble micro F1). Two test scenes (2012, 2025) fall outside the main training period, and the two winter test scenes cluster at lower F1 values.

Figure 5 further suggests that the lowest-performing scenes are associated with conditions that are poorly represented in the training data, including an early acquisition year (2012) and winter imagery. Together, these observations indicate that temporal and seasonal coverage may contribute to performance variability alongside foreground sparsity.

An operational NOAA deployment could incorporate confidence-based quality control, whereby scenes containing exceptionally low predicted water fractions or high model uncertainty are flagged for manual verification before downstream wetted-area estimation. Such a workflow would allow automated processing for the majority of scenes while concentrating expert effort on the relatively small subset of difficult cases. Since sparse-water conditions correspond to ecologically important low-flow periods, improving performance under these conditions is likely to have greater operational value than achieving marginal gains on already well-performing high-water scenes. Similar targeted sampling strategies may represent a more promising direction than increasing architectural complexity alone. Several limitations should be noted. The dataset is geographically constrained to Pacific Northwest rivers, and transferability to other regions or sensor families is not assessed. Ground truth masks represent single annotations without inter-annotator variability quantification. The test set contains no high-water scenes (maximum water fraction 0.083), which may limit the generalisability of these results to peak-flow conditions. Hyperparameter selection was performed on Fold 1 before evaluation across all five folds. Although the held-out test set remained fully independent, this may introduce mild optimistic bias into fold-level validation results. Winter conditions are also under-represented: of the three winter scenes in the dataset, two fall in the test set and only one appears in train/val. Consequently, predictions on winter imagery should be interpreted with caution. Overall, this work demonstrates that supervised deep learning can accurately segment surface water in challenging high-resolution WorldView-2/3 imagery where traditional spectral indices and domain-mismatched pretrained models struggle. Within the en-

vironmental and sensor conditions represented in this dataset, the proposed U-Net ensemble achieved considerably higher segmentation accuracy than both NDWI and OmniWaterMask, while scene-level evaluation reduces the risk that the result is driven by spatial leakage. The study also shows that careful optimisation, threshold selection, and data splitting are central to reliable evaluation under severe class imbalance. These findings provide a solid foundation for future NOAA-facing workflows for estimating wetted surface area in Pacific Northwest salmon habitats, while highlighting the need for further validation across additional regions, sensors, and hydrological conditions before broader operational deployment.

VI. FUTURE WORK

Several directions extend this work naturally. An Azure-based batch inference pipeline for operational deployment would deliver NOAA-facing outputs: per-pixel probability maps, thresholded masks, and wetted surface area estimates per area of interest. Hard negative mining from high-confusion areas — shadows, rocks, and sunglint — could reduce the false positives that currently limit precision. Temporal analysis across the available multi-year dataset (2011–2026) would enable change detection of wetted area in response to flow variability and restoration interventions. Transferability experiments on rivers outside the Pacific Northwest would test whether learned representations generalise across climatically and geomorphologically distinct systems. Future work could also explore transformer-based architectures such as Swin Transformer (Liu et al., 2021) and Vision Transformer (ViT) (Dosovitskiy et al., 2021) as alternative encoders to ResNet34. Finally, uncertainty quantification using Monte Carlo dropout (Gal and Ghahramani, 2016) would provide per-pixel confidence estimates, indicating where predictions are reliable and where field verification is recommended. Beyond improving model architectures, future work should place greater emphasis on the statistical characterisation of model performance. Scene-level bootstrap confidence intervals and paired scene-level bootstrap or permutation tests would provide a more rigorous and spatially appropriate assessment of whether observed improvements generalise beyond the particular test set used in this study, while respecting spatial dependence within scenes. This would be especially valuable when comparing models whose performance differs by relatively small margins, such as the best single-fold model and the ensemble (+0.022 micro F1), where the practical significance of the difference is not immediately clear from point estimates alone.

REFERENCES

- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Munich, Germany. pp. 801–818.
- Dare, P.M. (2005) Shadow analysis in high-resolution satellite imagery of urban areas. *Photogrammetric Engineering and Remote Sensing*, 71(2), pp. 169–177.

- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021) An image is worth 16x16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR 2021)*.
- Gal, Y. and Ghahramani, Z. (2016) Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. New York, NY. pp. 1050–1059.
- He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV. pp. 770–778.
- Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2017) Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy. pp. 2980–2988.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021) Swin Transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10012–10022.
- Long, J., Shelhamer, E. and Darrell, T. (2015) Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Boston, MA. pp. 3431–3440.
- McFeeters, S.K. (1996) The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7), pp. 1425–1432.
- Milletari, F., Navab, N. and Ahmadi, S.A. (2016) V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision*. Stanford, CA. pp. 565–571. IEEE.
- Ronneberger, O., Fischer, P. and Brox, T. (2015) U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2015*. Munich, Germany. pp. 234–241. Springer.
- Salehi, S.S.M., Erdogmus, D. and Gholipour, A. (2017) Tversky loss function for image segmentation using 3D fully convolutional deep networks. In: *International Workshop on Machine Learning in Medical Imaging (MLMI)*. Quebec City, Canada. pp. 379–387. Springer.
- Siegel, J. and Crozier, L. (2018) *Impacts of climate change on salmon of the Pacific Northwest: a review of the scientific literature*. NOAA Northwest Fisheries Science Center.
- Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., Liu, W. and Xiao, B. (2021) Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10), pp. 3349–3364.
- Wright, N., Duncan, J.M.A., Callow, J.N., Thompson, S.E. and George, R.J. (2025) Adaptive water body detection: integrating deep learning, normalised difference water index, and vector data for farm dam water monitoring with Omni-WaterMask. *ISPRS Journal of Photogrammetry and Remote Sensing*, 225. <https://doi.org/10.1016/j.isprsjprs.2025.07.007>
- Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P. (2021) SegFormer: simple and efficient design for semantic segmentation with transformers. In: *Advances in Neural Information Processing Systems (NeurIPS)*. pp. 12077–12090.

APPENDIX

Student Name: Luana Marotti **Student Number:** 250016521

Project Title: High-Resolution Water Segmentation for Salmon Habitat

Supervisor: Dr. Iran Roman

Have you used Generative AI tools? YES AI Tool Used:

Tool	Version	Publisher	URL
Claude	Sonnet 4.6	Anthropic	claude.ai
ChatGPT	GPT-4o	OpenAI	chatgpt.com

How AI was used:

Section	Description	Justification
Dissertation draft	Initial draft text produced with AI assistance	All content reviewed and verified by student; student takes full responsibility
Code development	Assisted with debugging and script structure	All code verified, tested, and understood by student
Concept clarification	Used AI to clarify technical concepts during the project	Used as a learning aid only, equivalent to consulting a textbook or tutor. No AI-generated explanations were included in any submitted work.
Figure improvement	ChatGPT used to improve figure aesthetics and layout	All figures reviewed and verified by student; student takes full responsibility

Declaration: I have declared all uses of AI tools. I have reviewed and verified all AI-assisted content. I take full responsibility for the originality and integrity of this submission.
Signed: Luana Marotti Date: 13 August 2026