Supplementary Material to: Bayesian evaluation of deep learning architectures and sensor modalities for remote sensing-based driftwood segmentation in the Mackenzie Delta, Arctic Canada

Carl Stadie[1,2], Ingmar Nitze[1], Begüm Demir[2], Martin Stefan Brandt[3], and Guido Grosse[1,4]

[1]Alfred Wegener Institute Helmholtz Centre for Polar and Marine Science, 14473 Potsdam, Germany
[2]Faculty of Electrical Engineering and Computer Science, Technische Universität Berlin, 10587 Berlin, Germany
[3]Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark
[4]Institute of Geosciences, University of Potsdam, 14469 Potsdam, Germany

Correspondence: Carl Stadie (carl.stadie@awi.de)

Table S1. Acquisition of high-resolution aerial imagery.

ID	Name	Date	Area (km^2)	No. images	Flightlines	y-track overlap (%)	x-track overlap (%)	GSD (cm)
A	WhitefishDriftwood	2025-07-21	65.677	5965	13	60	40	15
B	DriftwoodFields	2025-07-21	52.805	5072	10	60	40	15
C	PittIsland	2025-08-12	78.874	5979	10	60	40	15
D	OlliverIsland	2025-08-12	36.465	4348	8	60	40	15
E	DriftwoodInnerDelta	2025-07-21	57.084	5983	8	60	40	15
F	DWChannelWest	2025-08-12	86.162	9390	11	60	40	15
G	MackenzieSub	2025-08-02	64.068	5800	12	60	40	15
H	DriftwoodKitti	2025-08-08	61.307	5593	8	60	40	15
I	KittiCoast	2025-08-02	80.294	5437	14	60	40	15
J	DriftwoodTuk	2025-07-28	32.799	3353	7	60	40	15

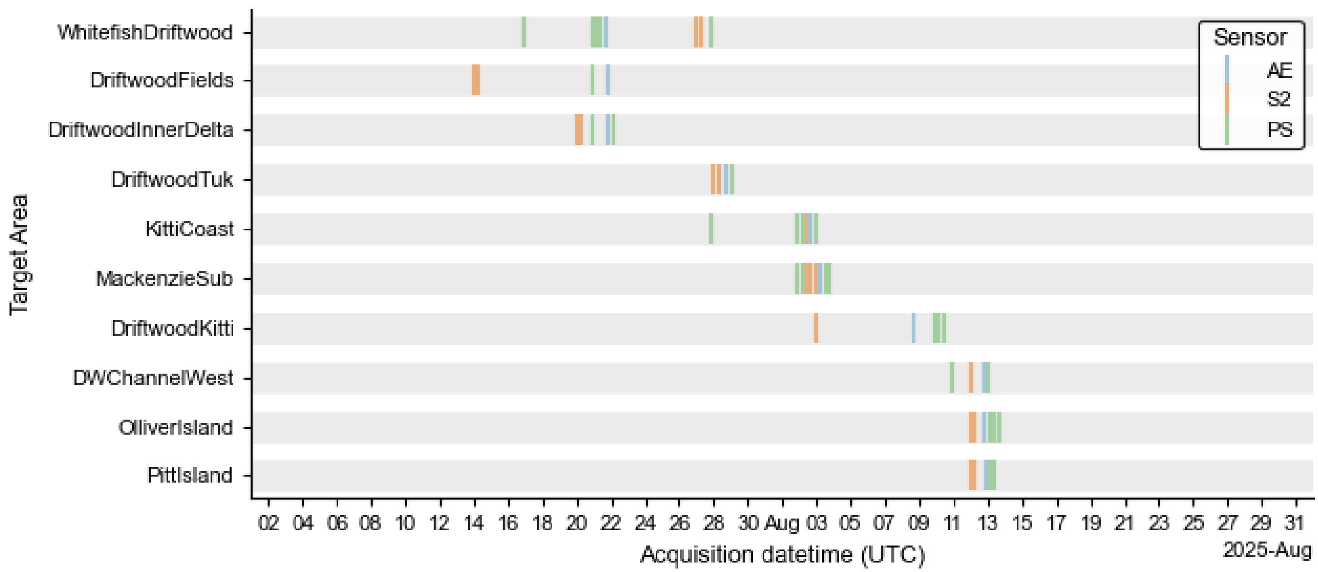

Figure S2. Temporal difference in acquisition times between aerial imagery (AE), PlanetScope (PS), and Sentinel-2 (S2) in August 2025.

Table S3: Hyperparameter search space.

Category	Hyperparameter	Models	Search Space	Explanation
Loss	Tversky α ($\beta = 1 - \alpha$)	ALL	0.3–0.9	loss function for imbalanced training data
Optimizer	optimizer	ALL	adam, adamw	
	learning rate		1×10^{-5}–5×10^{-3} (log)	
	weight_decay		1×10^{-6}–1×10^{-2} (log)	
LR Schedule	scheduler	ALL	none, cosine, onecycle	
Sampling / Imbalance	min. positive fraction	ALL	0.0, 0.01, 0.02, 0.05, 0.1	minimal fraction of positive pixels within training frames
	positive frame ratio		None, 0.1, 0.25, 0.5	ratio of positive training frames within a sampling run
U-Net architecture	dilation rate	U-Net	1, 2, 4	inserted space between kernel to increase the perceptive field
	layer count		32, 64, 96	
	l2 weight		0, 1×10^{-5}, 1×10^{-4}	
	dropout		0.0, 0.1, 0.2	
Swin-U-Net architecture	C (base channels)	Swin-U-Net	64, 96 (or 96 if pretrained)	
	drop path		0.0, 0.1, 0.2, 0.3	stochastic depth; probability to randomly drop entire layers
	patch size, window size		fixed	Swin-U-Net attention grid structure keeps patch size fixed
TerraMind architecture	decoder	TerraMind	UNetDecoder, UperNet-Decoder	
	decoder channels		128, 192, 256, 384	
	head dropout		0.0–0.2	
	freeze backbone every n epochs		0, 1, 3, 5	epochs the pretrained backbone stays frozen
	backbone learning rate		1×10^{-6}–3×10^{-5} (log)	learning rate applied to the pre-trained backbone
	learning rate multiplier		5.0, 10.0	learning rate multiplier applied to decoder/head
	weight decay		1×10^{-6}–5×10^{-4} (log)	
Validation objective	—	ALL	max Dice	
Training loss	—	ALL	Tversky(α, $\beta = 1 - \alpha$)	

Table S4. Hyperparameters for each configuration.

Hyperparameter	U-Net|AE	U-Net|PS	U-Net|S2	Swin-U-Net|AE	Swin-U-Net|PS	Swin-U-Net|S2	TerraMind|AE	TerraMind|PS	TerraMind|S2
Tversky α ($\beta = 1 - \alpha$)	0.55	0.55	0.63	0.6	0.7	0.7	0.7	0.6	0.6
optimizer	adamw	adamw	adam	adam	adamw	adamw	adamw	adam	adam
learning rate	4×10^{-4}	1.5×10^{-3}	4.2×10^{-4}	1.6×10^{-4}	1×10^{-4}	1×10^{-4}	1.00×10^{-3}	1.00×10^{-3}	1.00×10^{-3}
weight_decay	4.8×10^{-6}	1.65×10^{-5}	—	—	1.6×10^{-4}	6.3×10^{-3}	1.00×10^{-4}	—	—
scheduler	onecycle	cosine	onecycle	onecycle	onecycle	none	cosine	cosine	onecycle
min. positive fraction	1×10^{-3}	0.0	1.00×10^{-5}	5.00×10^{-5}	1×10^{-4}	1×10^{-4}	1×10^{-3}	5×10^{-4}	1×10^{-3}
positive frame ratio	None	None	0.5	0.75	None	None	0.75	0.25	0.5
dilation rate	2	4	2	—	—	—	—	—	—
layer_count	96	96	96	—	—	—	—	—	—
l2 weight	1×10^{-5}	0.0	1×10^{-5}	—	—	—	—	—	—
dropout	0.1	0.1	0.1	—	—	—	—	—	—
C (base channels)	—	—	—	96	96	96	—	—	—
drop path	—	—	—	0.0	0.2	0.2	—	—	—
patch size, window size	—	—	—	4, 7	4, 7	4, 7	—	—	—
decoder	—	—	—	—	—	—	UperNet	UNet	UperNet
decoder channels	—	—	—	—	—	—	256	384, 192, 96, 48	384
head dropout	—	—	—	—	—	—	0.05	0.18	0.007
freeze backbone every n epochs	—	—	—	—	—	—	3	5	5
backbone learning rate	—	—	—	—	—	—	2.7×10^{-5}	1.9×10^{-5}	7.5×10^{-6}
learning rate multiplier	—	—	—	—	—	—	5	10	10
weight decay	—	—	—	—	—	—	2×10^{-4}	1.01×10^{-6}	2.9×10^{-6}

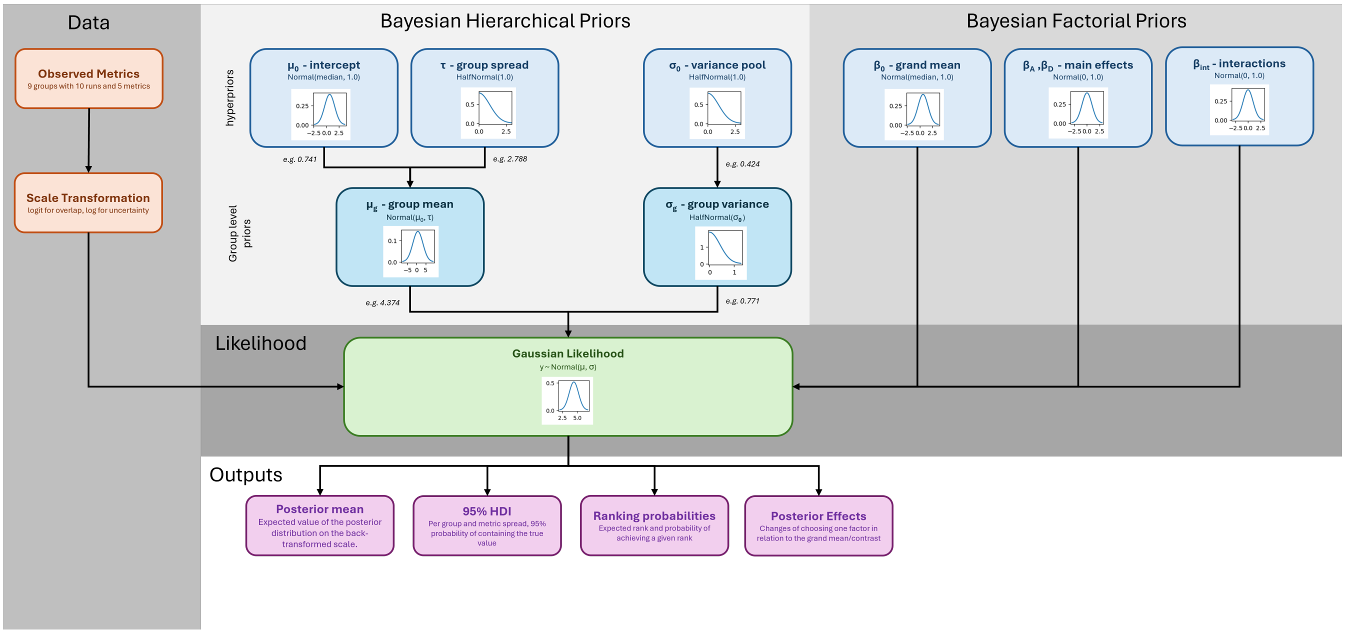

Figure S5. Graphical overview of the Bayesian hierarchical and factorial models used for performance evaluation. Observed metric values (orange) are transformed to an unconstrained scale before entering both models. In the hierarchical model (blue, centre), population hyperpriors on the intercept (μ_0), group spread (τ), and variance pool (σ_0) determine group-level priors on the partially pooled group mean (μ_g) and standard deviation (σ_g) for each sensor–architecture combination; example draws (e.g. 0.741, 2.788, 4.374) illustrate how a single prior sample propagates through the hierarchy. In the factorial model (blue, right), weakly informative priors are placed directly on the grand mean (β_0), main effects of sensor and architecture (β^d, β^a), and their interaction (β^{int}). Both models share a Gaussian likelihood (green); inset panels show the density implied by each prior and an example posterior predictive distribution. Posterior summaries (purple) include posterior means, 95 % highest density intervals (HDI), and rank probabilities for the hierarchical model, and posterior effect sizes relative to the grand mean for the factorial model. Full model specification is given in Sect. 3.3 of the main article.

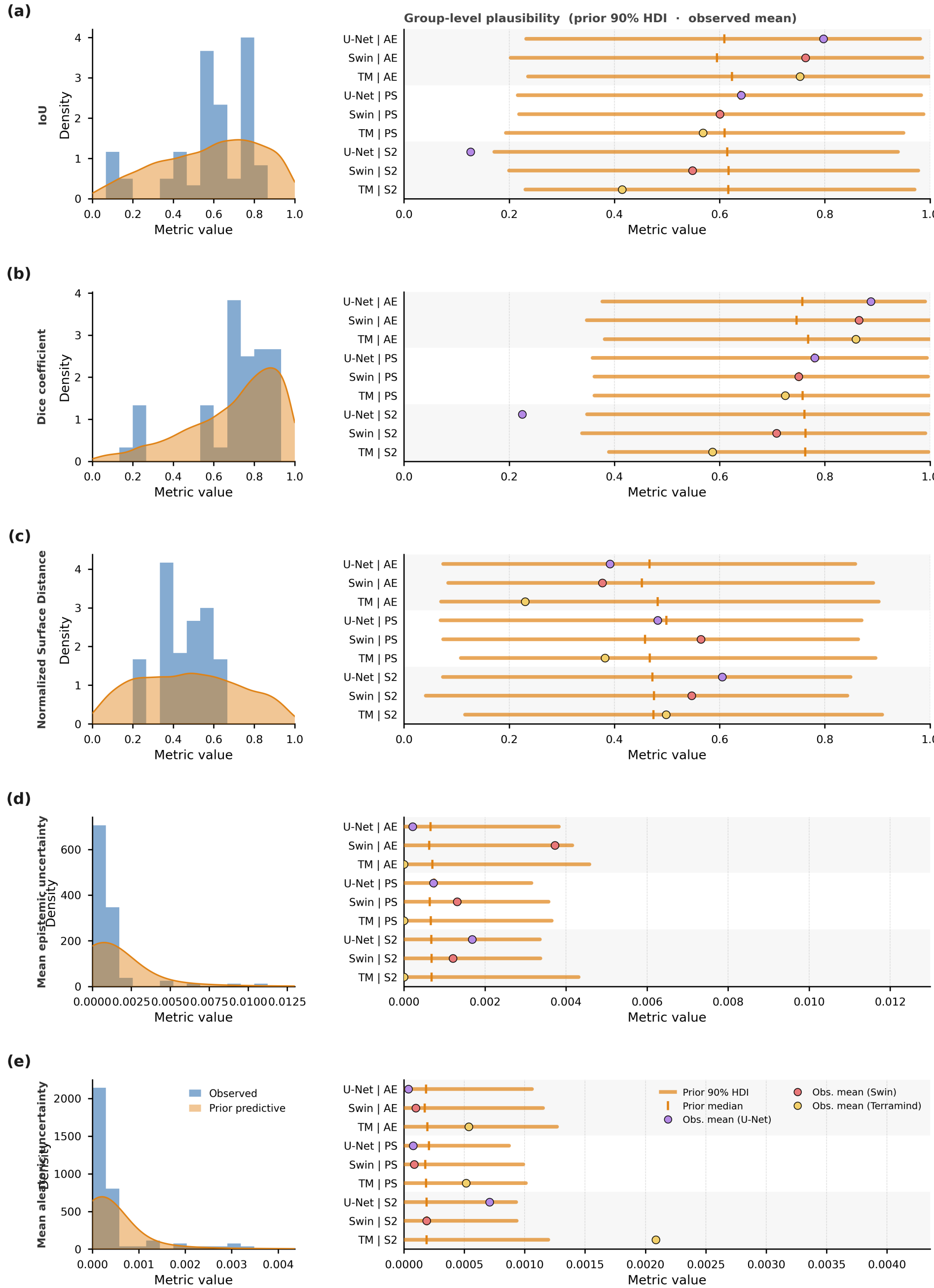

Figure S6. Prior predictive checks for the hierarchical group model across all five evaluation metrics, showing the marginal prior predictive distribution against observed data (left) and group-level plausibility of observed group means relative to the prior 90 % highest density interval (right).

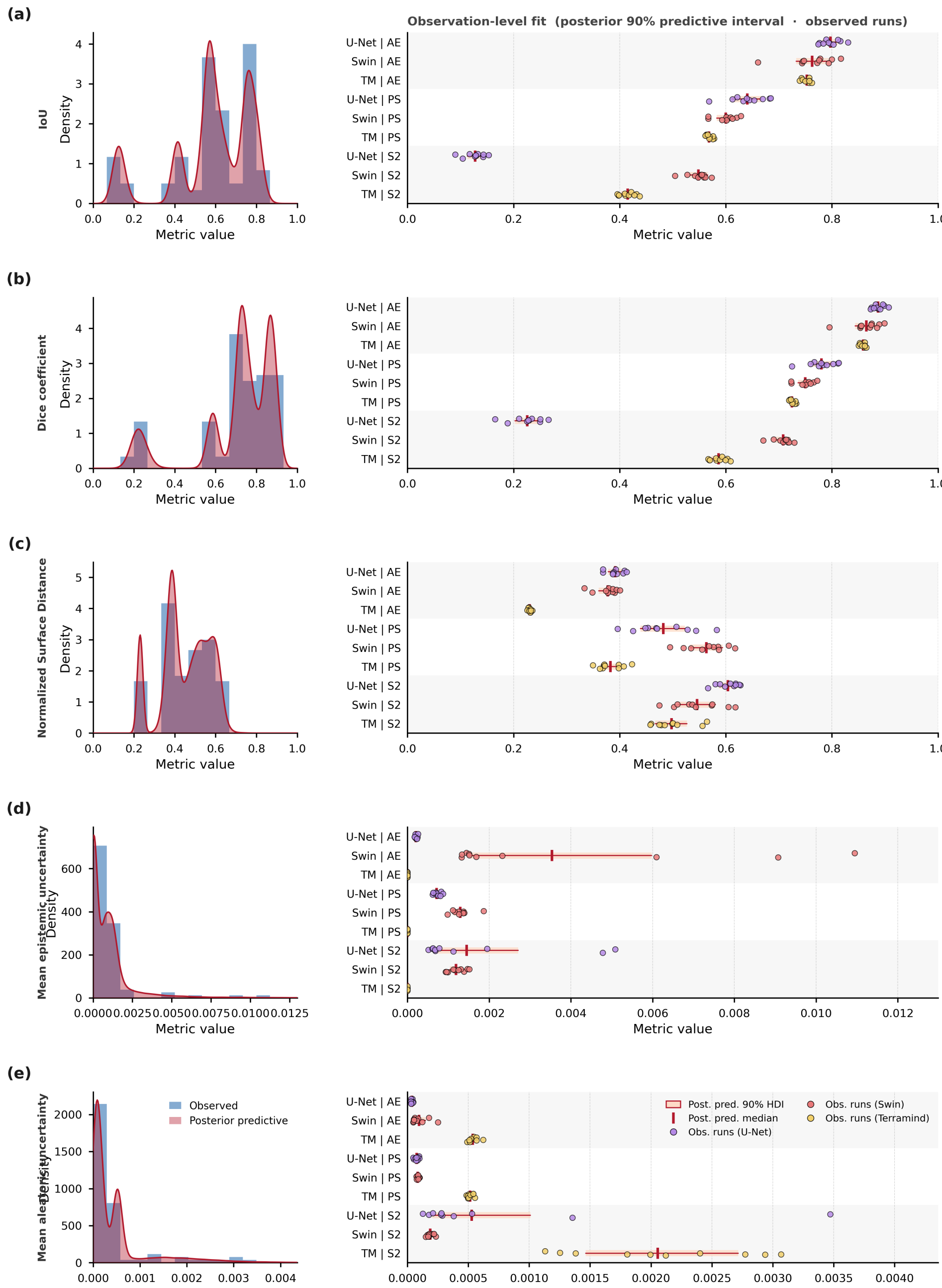

Figure S7. Posterior predictive checks for the hierarchical group model across all five evaluation metrics, showing the marginal posterior predictive distribution against observed data (left) and observation-level fit relative to the posterior 90 % predictive interval (right).

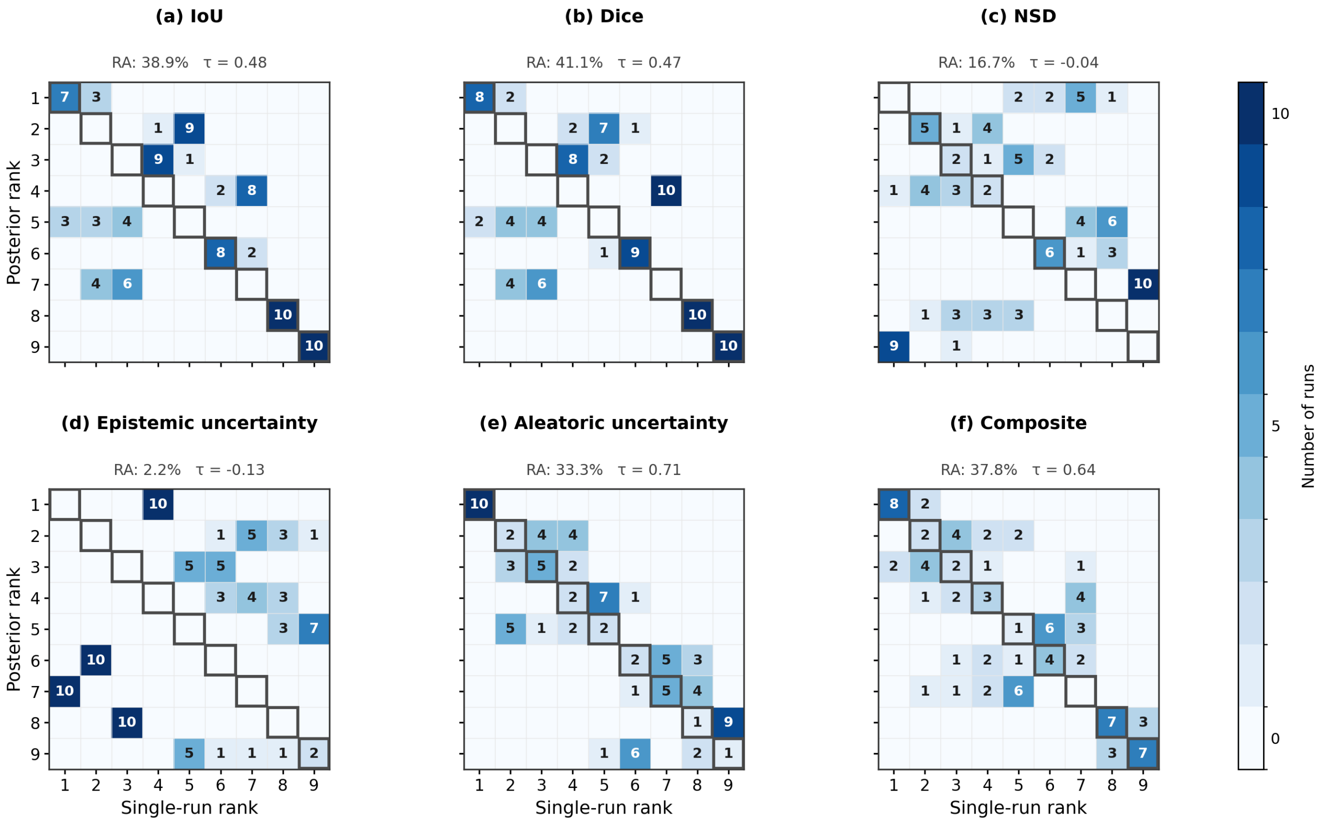

Figure S8. Rank agreement between single-run and composite posterior rankings across all five evaluation metrics and the composite score. Each panel shows a 9×9 matrix where the x-axis gives the rank assigned by a single training run and the y-axis gives the posterior expected rank per metric/composite score; cell numbers indicate how many of the 10 runs assigned that single-run rank to the configuration holding that posterior rank. Grey boxes outline the diagonal. RA: exact rank agreement averaged across runs; τ: Kendall rank correlation between single-run and posterior rank vectors, averaged across runs.

Table S9: Satellite Imagery.

ID	Study Site	Sensor	Scene Identifier	Acquisition Date
B	DriftwoodFields	PlanetScope	20250720_211307_09_2518	2025-07-20
		Sentinel-2	20250714T211519_20250714T211528_T07WFS	2025-07-14
		Sentinel-2	20250714T211519_20250714T211528_T08WMB	2025-07-14
E	DriftwoodInnerDelta	PlanetScope	20250720_211139_82_24ee	2025-07-20
		PlanetScope	20250721_210140_05_2506	2025-07-21
		Sentinel-2	20250720T203901_20250720T203902_T08WMB	2025-07-20
		Sentinel-2	20250720T203901_20250720T203902_T08WNB	2025-07-20
H	DriftwoodKitti	PlanetScope	20250809_210810_65_2532	2025-08-09
		PlanetScope	20250809_210813_01_2532	2025-08-09
		PlanetScope	20250809_210815_37_2532	2025-08-09
		Sentinel-2	20250803T211509_20250803T211512_T08WNC	2025-08-03
J	DriftwoodTuk	PlanetScope	20250728_205542_92_2514	2025-07-28
		Sentinel-2	20250728T205019_20250728T205209_T08WNB	2025-07-28
		Sentinel-2	20250728T205019_20250728T205209_T08WNC	2025-07-28
F	DWChannelWest	PlanetScope	20250810_211137_60_24ee	2025-08-10
		PlanetScope	20250812_211106_25_253c	2025-08-12
		Sentinel-2	20250812T205041_20250812T205337_T08WMB	2025-08-12
I	KittiCoast	PlanetScope	20250727_211540_28_24ed	2025-07-27
		PlanetScope	20250801_210138_59_2500	2025-08-01
		PlanetScope	20250801_210633_75_24db	2025-08-01
		PlanetScope	20250802_211323_21_24fe	2025-08-02
		Sentinel-2	20250802T205041_20250802T205041_T08WNC	2025-08-02
G	MackenzieSub	PlanetScope	20250801_210635_98_24db	2025-08-01
		PlanetScope	20250801_210638_21_24db	2025-08-01
		PlanetScope	20250802_210848_58_251c	2025-08-02
		PlanetScope	20250802_210850_82_251c	2025-08-02
		Sentinel-2	20250802T205041_20250802T205041_T08WMC	2025-08-02
		Sentinel-2	20250802T205041_20250802T205041_T08WNB	2025-08-02
		Sentinel-2	20250802T205041_20250802T205041_T08WNC	2025-08-02
D	OlliverIsland	PlanetScope	20250812_211103_91_253c	2025-08-12
		PlanetScope	20250812_211106_25_253c	2025-08-12
		PlanetScope	20250812_211130_24_24fe	2025-08-12
		Sentinel-2	20250812T205041_20250812T205337_T07WFS	2025-08-12
		Sentinel-2	20250812T205041_20250812T205337_T08WMB	2025-08-12
C	PittIsland	PlanetScope	20250812_211127_88_24fe	2025-08-12
		PlanetScope	20250812_211130_24_24fe	2025-08-12
		Sentinel-2	20250812T205041_20250812T205337_T07WFS	2025-08-12
		Sentinel-2	20250812T205041_20250812T205337_T08WMB	2025-08-12
A	WhitefishDriftwood	PlanetScope	20250716_211521_45_24ed	2025-07-16
		PlanetScope	20250720_210807_46_2516	2025-07-20
		PlanetScope	20250720_211304_89_2518	2025-07-20
		PlanetScope	20250720_211307_09_2518	2025-07-20
		PlanetScope	20250727_211147_24_24dc	2025-07-27
		Sentinel-2	20250727T212519_20250727T212521_T07WFS	2025-07-27
		Sentinel-2	20250727T212519_20250727T212521_T08WMB	2025-07-27