A. Supplementary

A.1. List of Symbols/Variables

Symbol	**Meaning**
$\Omega \subset \mathbb{R}^2$	Spatial domain (rectangular region)
H, W	Grid height/width (pixels)
Δ	Pixel spacing (≈ 150 m)
(x_i, y_j)	Grid coordinates
P	Tile/patch size (default 256)
b	Tile core border for seam-free tiling (default 96)
δ	Safety buffer to define held-out cores (default 96)
$C_{\text{tr}}, C_{\text{te}}$	Eroded train/test cores (buffered)
$s(x, y)$	Surface elevation
$v_x(x, y), v_y(x, y)$	Surface velocity components; $\mathbf{v} = (v_x, v_y)$
$\mathrm{SMB}(x, y)$	Surface mass balance
$\partial h/\partial t(x, y)$	Thickness tendency (abbrev. dhdt)
∇s	Surface gradients (aux feature)
$b_p(x, y)$	Prior bed (BedMachine)
$h_p(x, y) = s - b_p$	Prior thickness
$\hat{r}(x, y)$	*Normalized* residual thickness (network output)
μ_t, σ_t	Robust residual stats (median, $1.4826\times$MAD) for (de)norm.
$r(x, y) = \sigma_t \hat{r} + \mu_t$	*De-normalized* residual thickness
$\hat{h}(x, y) = h_p + r$	Predicted thickness
$\hat{b}(x, y) = s - \hat{h}$	Predicted bed
$P = \{(x_k, y_k, b_k^{\text{rad}})\}_{k=1}^{N}$	Radar pick set (locations and observed bed)
$m(x, y) \in \{0, 1\}$	Radar pick mask on the grid
$c(x, y) \in [0, 1]$	Radar confidence (decays with distance)
$h^{\text{rad}}(x, y)$	Thickness at picks ($= s - b^{\text{rad}}$ after splat)
$w(x, y)$	Per-pick weight (typically $w = \max(\epsilon, c)$)
$Z = \sum m\, w$	Normalizer for data loss
$d_{\text{rad}}(x, y)$	Distance to nearest radar pick (grid px)
τ	Confidence decay scale (default 12 px)
K	# neighbors for radar splat (default 9)
r	Splat radius (2.5 px $\times\Delta$)
$\mathcal{R}(x, y; \hat{h})$	Mass-conservation residual $\frac{\partial \hat{h}}{\partial t} + \nabla\cdot(\hat{h}\,\mathbf{v}) - \mathrm{SMB}$
$\nabla, \nabla\cdot(\cdot)$	Spatial gradient, divergence
$\Delta(\cdot)$	Discrete Laplacian
$\mathbf{u} = \mathbf{v}/(\lVert\mathbf{v}\rVert + \epsilon)$	Unit flow direction; $\mathbf{u}_\perp$ orthogonal unit vector
$\mathcal{S} = \{1, 2, 4\}$	Multi-scale pooling factors for $\mathcal{L}_{\text{mass}}$
$\tilde{\cdot}$	Gaussian-smoothed fields (flux smoothing)
$\mathcal{L}$	Total training loss
$\mathcal{L}_{\text{radar}}$	Masked Huber fit to thickness at picks
$\mathcal{L}_{\text{mass}}$	Multi-scale mass-conservation penalty
$\mathcal{L}_{\text{flowTV}}$	Flow-aligned total variation (cross-flow $>$ along-flow)
$\mathcal{L}_{\text{lap}}$	Laplacian/high-pass damping on residual r
$\mathcal{L}_{\geq 0}$	Non-negativity hinge on $\hat{h}$ (soft $\hat{h} \geq 0$)
$\mathcal{L}_{\text{prior}}$	Prior consistency (masked near picks; stronger where $c\downarrow$)
$\lambda_\bullet$	Loss weights for each term ($\lambda_{\text{data}}, \lambda_{\text{phys}}, \lambda_{\text{tv}}, \lambda_{\text{lap}}, \lambda_{\geq 0}, \lambda_{\text{prior}}$)
$\rho_\delta(\cdot)$	Huber penalty (with $\delta_{\text{radar}}, \delta_{\text{mass}}, \delta_{\text{prior}}$)
$\beta_\perp, \beta_{\parallel}$	Flow-TV weights (cross/along)
q	Exponent for $(1 - c)^q$ in physics weighting
EMA (θ_{ema})	Exponential moving average of weights (decay ≈ 0.999)
TTA	8-way test-time augmentation (rotations/flips)
C_{te} metrics	MAE/RMSE/R^2, SSIM, PSNR, $\lvert\Delta\mathrm{TRI}\rvert$ on test core

A.2. Implementation Details

Preprocessing and features. All rasters are reprojected to EPSG:3413 and resampled to $\Delta \approx 150\,\mathrm{m}$ on an $H \times W$ grid (two 600×600 extracts). Scalars are standardized per–channel using training-region statistics (mean/STD over valid land pixels). We append: (i) $\nabla s = (\partial_x s, \partial_y s)$ via central differences; (ii) Fourier coordinate features with $L{=}3$ bands on (x, y); (iii) prior thickness $h_p{=}s{-}b_p$. Residual normalization uses robust statistics from training radar residuals: $\mu_t{=}\,\mathrm{median}$, $\sigma_t{=}1.4826 \cdot \mathrm{MAD}$.

Radar splat and confidence. Radar picks are splatted to the grid using a cKDTree with $K{=}9$ nearest cells and Gaussian weights $\exp\big(-(d/r)^2\big)$ where $r{=}\,2.5\,\mathrm{px} \times \Delta$. The radar confidence map is $c(x,y){=}\exp(-d/\tau)$ with $\tau{=}12\,\mathrm{px}$ (grid distance to nearest pick). Weights: $w_{\text{radar}}{=}\max(\varepsilon, c)$, physics losses use $(1{-}c)$, prior loss uses $(1{-}c)^2$ and is masked at picks. To avoid pulling on steep slopes, we attenuate L_{prior} by a slope weight $w_{\nabla b_p}{=}\exp(-\|\nabla b_p\|/s_{90})$ where s_{90} is the 90th percentile of $\|\nabla b_p\|$.

Model and heads. DeepLabV3+ decoder with a ResNet-50 encoder (output stride 16). Low-level projection $1{\times}1$ (48 ch), ASPP rates $(1, 6, 12, 18)$, GroupNorm everywhere, LeakyReLU, dropout 0.1 in ASPP/decoder. Output is the normalized residual $\hat{r} \in \mathbb{R}^{H \times W}$.

Losses and weights. Total loss $L = \lambda_{\text{data}} L_{\text{radar}} + \lambda_{\text{phys}} L_{\text{mass}} + \lambda_{\text{tv}} L_{\text{flowTV}} + \lambda_{\text{lap}} L_{\text{lap}} + \lambda_{\geq 0} L_{\geq 0} + \lambda_{\text{prior}} L_{\text{prior}}$. Huber deltas: $\delta_{\text{radar}}{=}1.0$, $\delta_{\text{mass}}{=}5.0$, $\delta_{\text{prior}}{=}10.0$. Flow-aligned TV uses unit flow $\mathbf{u} = \mathbf{v}/(\|\mathbf{v}\|{+}\epsilon)$ with $L_{\text{flowTV}} = \beta_\perp \|\nabla \hat{h} \cdot \mathbf{u}_\perp\|_1 + \beta_\| \|\nabla \hat{h} \cdot \mathbf{u}\|_1$, $\beta_\perp{=}0.9$, $\beta_\|{=}0.35$. Laplacian uses the $3{\times}3$ kernel $\left(\begin{smallmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{smallmatrix}\right)$ on $\hat{r}$. Mass-conservation residual is applied multi-scale at pooling factors $\{1, 2, 4\}$, with Gaussian-smoothed fluxes; kernel schedule: first half of training uses size 11, $\sigma{=}3.5$, then size 15, $\sigma{=}5.0$.

Schedules (ramp and weights). Unless noted: $\lambda_{\text{data}}{=}2.0$, $\lambda_{\text{phys}}{=}10^{-2}$ (linear ramp 0→target over the first $\sim$ 90% of epochs), $\lambda_{\text{tv}}{=}5{\times}10^{-4}$, $\lambda_{\text{lap}}{=}2{\times}10^{-4}$, $\lambda_{\geq 0}{=}10^{-3}$, $\lambda_{\text{prior}}{=}5{\times}10^{-3}$ (ramp from 30% to 90% of training).

Optimization and training protocol. AdamW (lr $1{\times}10^{-4}$, weight decay $1{\times}10^{-4}$) with cosine warm restarts ($T_0{=}500$, $T_{\text{mult}}{=}2$); batch size 8; up to 6000 epochs with early stopping (patience 2000 epochs) on masked radar-thickness fit inside the train core. Seed fixed to 42. WeightedRandomSampler favors tiles that contain any radar in the patch core (weights 6:1).

Geo-aware augmentation and inference. During training we apply random $\pi/2$ rotations and flips with probability 0.75 (vector-aware transforms for v_x, v_y and gradient channels). At test time we use 8-way TTA (4 rotations × horizontal flip), inverse-transform and average. EMA decay 0.999; seam-free tiled inference with patch 256, stride 64, and core border $b{=}96\,\mathrm{px}$. For leakage-safe splits we erode train/test blocks by $\delta{=}96\,\mathrm{px}$ and compute metrics strictly on the held-out test core.

Evaluation specifics. We select a single global rotation/flip that minimizes RMSE against the reference once per split, then compute MAE/RMSE/R^2, SSIM, and PSNR (range = dynamic range of the reference in the core). TRI uses a $3{\times}3$ neighborhood; $|\Delta\mathrm{TRI}|$ is mean absolute difference in the core. Radar-only errors sample $\hat{b}$ at held-out picks in the core.

A.3. Distance-to-Radar Stratification

To examine how performance varies with observational support, we stratify test-core pixels by distance to the nearest radar pick into three bins: 0–2, 2–6, and $>$ 6 pixels (1 px $\approx$ 150 m), and report RMSE in each bin (Fig. 3). Across both sub-regions and both splits, our physics-guided residual model attains the lowest RMSE in *every* bin and degrades the least as distance increases. In fact, RMSE typically *decreases* farther from picks for our method (e.g., Sub-Region I—H: $9.33 \to 8.91 \to 7.48$ m; Sub-Region II—V: $5.75 \to 4.84 \to 2.76$ m), consistent with the design: near margins and complex flow (where picks cluster) the field is harder to predict, whereas in radar-sparse interiors the prior-consistency and mass terms stabilize the reconstruction. By contrast, CNN/U-Net/FPN exhibit larger near-pick errors and a shallower improvement with distance (e.g., Sub-Region I—H at $>$ 6 px: U-Net/FPN 16.87/24.29 m vs. ours 7.48 m; Sub-Region II—V at $>$ 6 px: U-Net/FPN 6.14/10.37 m vs. ours 2.76 m). These trends indicate that residual-over-prior learning with lightweight physics yields robust generalization in radar-sparse interiors while avoiding the banding and over-smoothing seen in non-physics baselines. (Note: stratification uses BedMachine as the reference).

A.4. Ablation Study Using Feature Pyramid Network (FPN) as a Backbone

Table 6 repeats the loss-component ablation with an FPN decoder in place of DeepLabV3+ while keeping the same residual-over-prior target, inputs, schedules, and leakage-safe protocol. The trends mirror the main-paper ablation: (i) **Prior-consistency** is pivotal—removing $\mathcal{L}_{\text{prior}}$ produces the largest degradation on the BedMachine comparison across splits (e.g., Sub-Region I V/H: RMSE $46.09/39.77\,\mathrm{m}$; Sub-Region II V: $70.05\,\mathrm{m}$ with $R^2{<}0$), confirming that an explicit pull toward the prior is required in radar-sparse interiors. (ii) The **non-negativity** hinge is essential for physical plausibility; without it, errors are catastrophic with strongly negative R^2 in all splits (e.g., Sub-Region I V: RMSE $381.92\,\mathrm{m}$). (iii) The **mass-conservation** term improves field structure—particularly in anisotropic splits—though its removal yields moderate increases relative to (i)/(ii) (e.g., Sub-Region I V/H: $35.90/21.50\,\mathrm{m}$ RMSE). (iv) Dropping the **radar data**

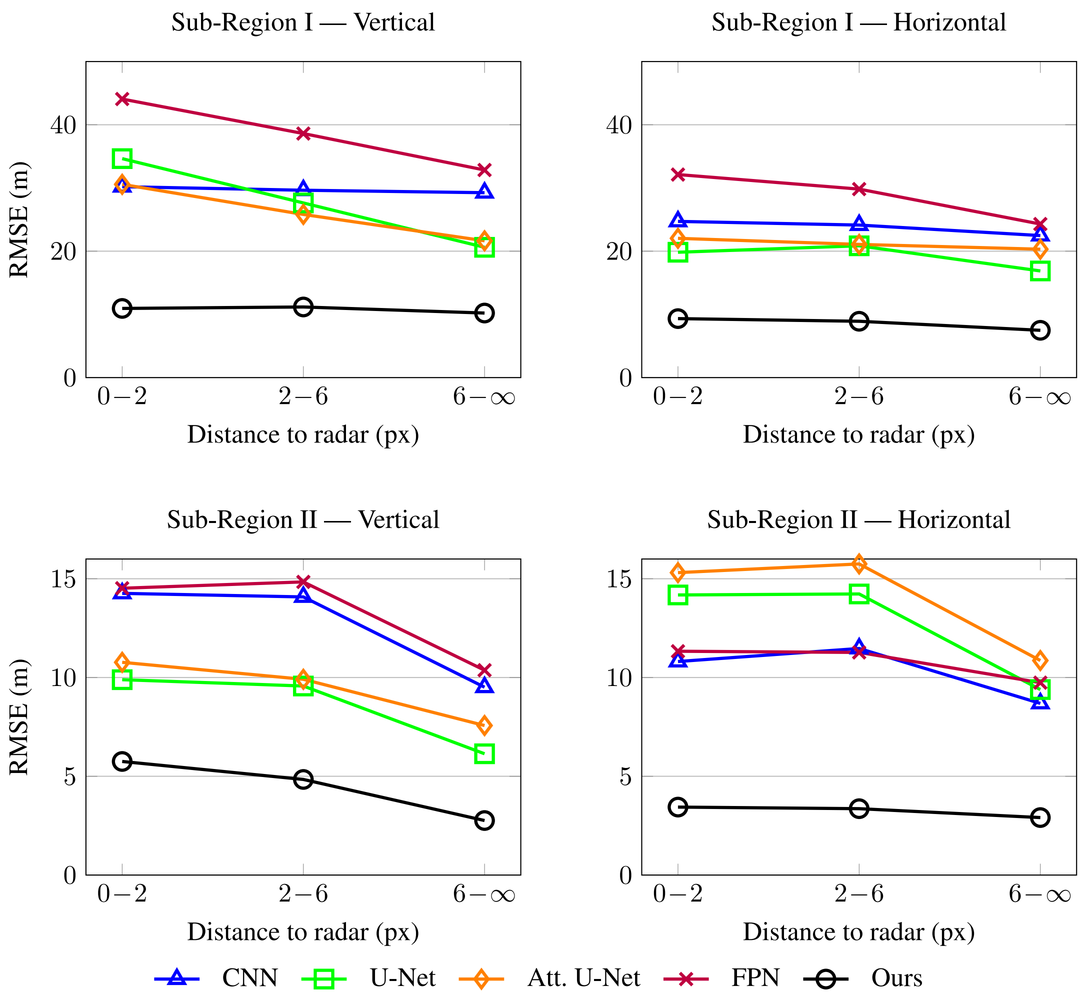


Figure 3. **RMSE vs. distance to radar on held-out test cores.** For each sub-region (rows) and split (columns), RMSE is reported in three distance bins (px) from the nearest radar pick: 0–2, 2–6, and 6–∞. Lines encode methods (markers/colors in the legend). Lower is better.

term $\mathcal{L}_{\text{radar}}$ raises pick-proximal errors (e.g., Sub-Region I V: RMSE 84.97 m vs. with-fit) and modestly worsens BedMachine agreement, indicating that $\mathcal{L}_{\text{radar}}$ anchors local corrections while the physics and prior govern global behavior.

Overall, the *ranking* of term importance (non-negativity/prior $\gg$ mass $\gtrsim$ radar) is consistent with the DeepLabV3+ results in the main paper, while absolute errors are somewhat larger with FPN—reflecting decoder capacity rather than a change in the efficacy of the loss design. These findings suggest the proposed residual+physics formulation is *backbone-agnostic*: the same components deliver the same qualitative benefits across architectures.

A.5. Quality Maps for Sub-Region II

All panels use the same elevation limits; difference maps use a zero-centered diverging colormap (white $\approx$ 0 m). A single rotation/flip alignment is fixed once per split (as in the main protocol).

Classical interpolators (IDW, Kriging) either over-smooth fjord walls and interior ridges or exhibit track-aligned streaks, resulting in structured residuals in the difference maps. CNN/U-Net/FPN reduce banding but still blur valley edges and leave halo-like artifacts near steep margins. In contrast, our physics-guided residual model preserves trough continuity and flank sharpness while producing low-amplitude, spatially diffuse differences across both vertical and horizontal splits. These visuals align with the quantitative trends reported for Sub-region II: very high structural fidelity (SSIM ≈ 0.999, PSNR up to ~ 52.94 dB) and the lowest roughness discrepancy ($|\Delta\text{TRI}| \approx 0.66$), together with low test-core RMSE (3.33 m vertical, 3.05 m horizontal). Notably, residuals for our method remain decorrelated from radar track geometry, indicating that corrections are driven by the residual-over-prior + physics design rather than memorization of pick patterns.

Table 6. **Ablations of loss components on held-out test cores using FPN as a backbone.** Effect of removing mass-conservation ($\mathcal{L}_{\text{mass}}$), prior-consistency ($\mathcal{L}_{\text{prior}}$), non-negativity, and radar data fit ($\mathcal{L}_{\text{radar}}$). Protocol matches the main-paper ablation (residual-over-prior target, identical schedules, and test-core scoring); values are in meters.

Method	Reference Data	Sub-Region I						Sub-Region II					
		Vertical			Horizontal			Vertical			Horizontal		
		MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
w/o $\mathcal{L}_{mass}$	BedMachine	25.98	35.90	0.920	14.50	21.50	0.970	6.20	9.53	0.954	6.33	10.44	0.985
	Radar	76.77	96.41	0.421	93.32	123.89	0.265	99.43	128.00	-0.020	59.11	89.63	0.447
w/o $\mathcal{L}_{prior}$	BedMachine	35.88	46.09	0.868	30.17	39.77	0.896	54.62	70.05	-1.483	25.46	30.52	0.868
	Radar	76.63	98.38	0.397	89.92	117.36	0.340	109.11	140.86	-0.235	61.98	94.40	0.387
w/o non-negativity	BedMachine	375.73	381.92	-8.093	108.13	124.96	-0.024	280.11	287.00	-40.688	87.28	97.55	-0.344
	Radar	406.45	422.31	-10.105	154.17	191.96	-0.766	330.30	361.64	-7.143	112.66	147.55	-0.499
w/o $\mathcal{L}_{radar}$	BedMachine	22.15	31.17	0.939	14.54	21.36	0.970	6.84	9.31	0.956	5.15	8.32	0.990
	Radar	65.28	84.97	0.550	92.32	121.24	0.296	97.76	126.56	0.003	58.73	88.72	0.458

A.6. Comparative Assessment with Classical/ML Baselines

Table 8 reports leakage-safe test-core performance for common machine-learning regressors trained on the same inputs (surface, velocity, SMB, dhdt, gradients, Fourier coords, prior thickness) and residual target as in the main paper. We follow the identical protocol: orthogonal block-wise splits, receptive-field buffers, single rotation/flip chosen once per split, and scoring on the held-out core. Hyperparameters were tuned on a slice of the train core to avoid test leakage. Overall, support-vector regression is the strongest ML baseline, attaining RMSE $\sim$25–33 m on Sub-region I and $\sim$15–25 m on Sub-region II against BedMachine, while linear (Ridge/ElasticNet) and tree ensembles (RF/GB) lag or overfit. KNN performs inconsistently across splits. Compared to the DeepLabV3$+$ residual+physics model in the main paper (RMSE 3–11 m), these ML baselines exhibit substantially higher errors and lower R^2, underscoring the value of: (i) learning *residual thickness* over a prior, (ii) physics-guided regularization, and (iii) leakage-safe training/evaluation.

Table 7. **Qualitative comparison on held-out test cores for Sub-region II.** For each method (rows) and split (columns: *Vertical*, *Horizontal*), we show triplets: *Prediction* | *Prior* b_p (BedMachine) | *Difference* (BedMachine $-$ $\hat{b}$). All panels use consistent color limits; difference maps use a zero-centered diverging colormap (white ≈ 0 m).

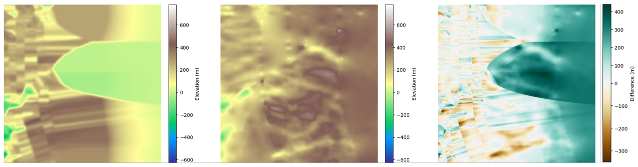
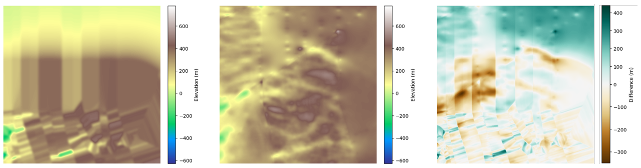
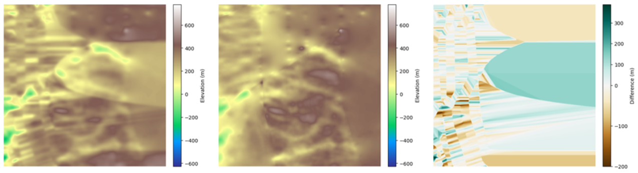
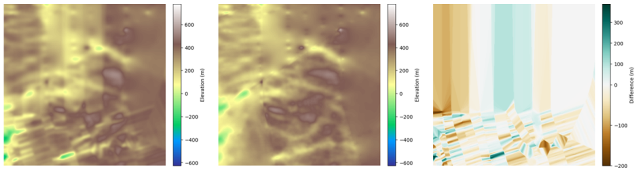
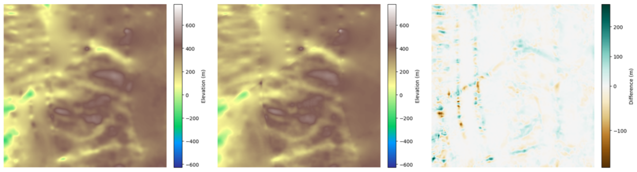
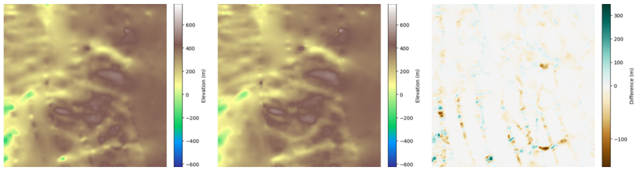
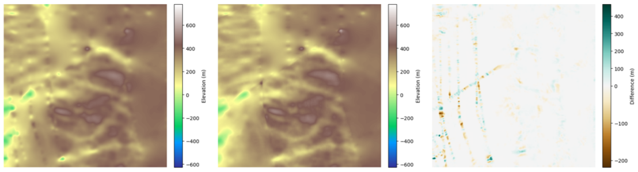
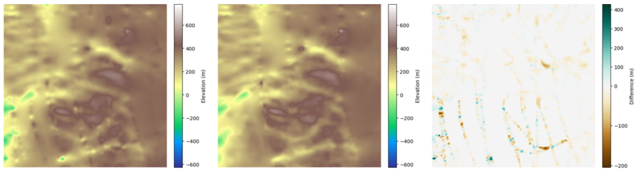
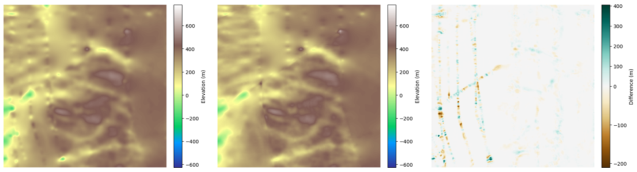
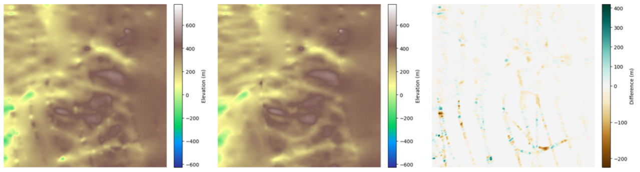
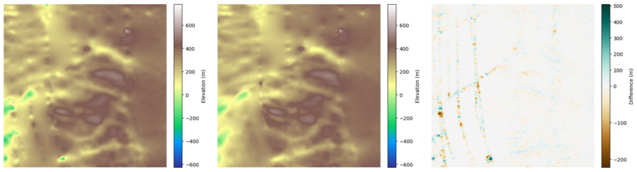
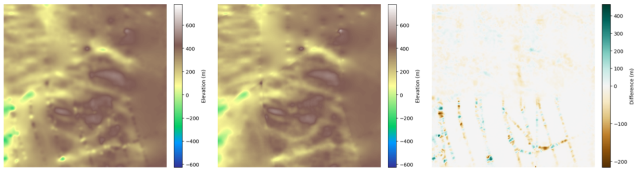
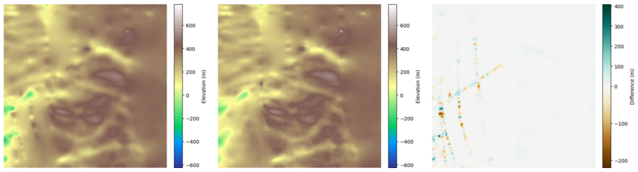
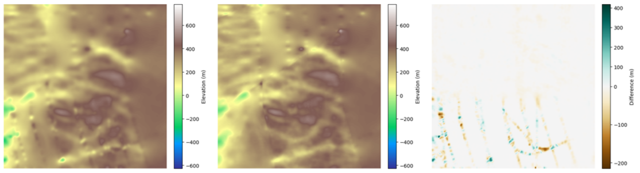
Vertical	Horizontal
Prediction | Prior b_p | Difference	
IDW [31]	
	
Kriging [9, 10, 13]	
	
CNN [12, 37, 40]	
	
U-Net [30]	
	
Att. U-Net [28, 30]	
	
FPN [21]	
	
Ours	
	

Table 8. **Machine-learning baselines on held-out test cores.** Ridge, ElasticNet, KNN, Random Forest (RF), Gradient Boosting (GB), and Support Vector Regression (SVR) trained on the same inputs and residual target as our model. "BedMachine" rows score agreement with the BedMachine prior b_p; "Radar" rows report errors at held-out radar picks. Protocol matches the main paper (block-wise splits with safety buffer, single orientation alignment per split, metrics on the test core). Values are in meters; higher R^2 is better.

Method	Reference Data	Sub-Region I						Sub-Region II					
		Vertical			Horizontal			Vertical			Horizontal		
		MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑	MAE↓	RMSE↓	R^2 ↑
Ridge	BedMachine	870.54	1027.53	-64.819	731.80	957.01	-59.064	2813.93	3164.69	-5067.668	537.27	725.08	-73.240
	Radar	664.39	893.91	-48.753	252.95	467.66	-9.478	2153.17	2705.44	-454.737	24.50	704.96	-33.208
ElasticNet	BedMachine	80.01	88.79	0.508	95.16	105.61	0.269	145.14	148.68	-10.188	36.67	48.13	0.673
	Radar	89.65	105.59	0.306	103.88	129.35	0.198	142.29	172.32	-0.849	79.75	115.22	0.086
KNN	BedMachine	94.07	128.69	-0.032	60.36	90.23	0.466	46.97	61.71	-0.927	71.68	114.78	-0.860
	Radar	101.17	134.56	-0.127	120.57	154.78	-0.148	97.78	125.95	0.012	110.02	157.71	-0.712
Random Forest	BedMachine	86.88	100.75	0.367	64.68	86.52	0.509	47.42	65.73	-1.187	42.64	58.79	0.512
	Radar	96.96	117.68	0.138	98.26	124.22	0.261	114.36	144.92	-0.308	65.23	86.66	0.483
Gradient Boosting	BedMachine	152.22	157.13	-0.539	34.49	47.68	0.851	61.93	71.18	-1.564	30.15	39.93	0.775
	Radar	114.84	135.03	-0.135	114.54	154.25	-0.140	107.59	135.90	-0.150	69.72	94.91	0.380
SVR	BedMachine	28.02	28.53	0.949	25.6	33.16	0.928	14.85	15.19	0.883	16.95	25.27	0.910
	Radar	75.82	96.11	0.425	96.16	121.86	0.289	97.77	124.61	0.033	64.72	97.38	0.347