Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Bridging ice charts and microwave records: probabilistic sea-ice concentration retrieval from SAR via multi- reference flow matching

**Seung-Jun Lee**
Sungkyunkwan University
**JiSung Kim**
gyjki@leeds.ac.uk
University of Leeds
**Hong-Sik Yun**
Sungkyunkwan University

**Article**

**Keywords:** Arctic sea ice, probabilistic retrieval, flow matching, ground-truth uncertainty, Sentinel-1 SAR, passive microwave, climate data record

**Posted Date:** November 25th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11735076/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Bridging ice charts and microwave records: probabilistic sea-ice concentration retrieval from SAR via multi-reference flow matching

**Seung-Jun Lee** [a], **Jisung Kim** [b,*], **Hong-Sik Yun** [a,*]

[a] Geodesy Laboratory, Department of Civil, Architectural and Environmental System Engineering, Sungkyunkwan University (SKKU), 2066 Seobu-ro, Jangan-gu, Suwon 16419, Gyeonggi, Republic of Korea

[b] School of Geography, Faculty of Environment, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, United Kingdom

* Corresponding authors. E-mail addresses: gyjki@leeds.ac.uk (J. Kim), yoonhs@skku.edu (H.-S. Yun) Tel.: +82-031-290-7522 (H.-S. Yun)

## Abstract

The Arctic is warming two to four times faster than the global average, intensifying demand for high-resolution sea-ice concentration (SIC) products that honestly represent observational uncertainty. Most data-driven SIC retrievals from synthetic aperture radar (SAR), however, treat one of several mutually inconsistent reference fields — analyst-drawn ice charts or passive-microwave (PMW) climate data records (CDRs) — as deterministic ground truth, inheriting that reference's biases without quantifying them. We systematically compare three independent SIC references — analyst polygons from the AI4Arctic dataset, the NSIDC G02202 v6 CDR, and the EUMETSAT OSI SAF v3.1 CDR — across 499 pan-Arctic Sentinel-1 scenes (2018–2021). The two PMW algorithms agree at median mean absolute error 5.3 % SIC (80 % pixel agreement within ±10 %), whereas analyst polygons disagree with either PMW reference at 6.8–7.7 % SIC, lowering three-way agreement to 68.6 % and contributing 11.5 percentage points of analyst-induced uncertainty concentrated in the marginal-ice zone (MIZ). We address this irreducible label uncertainty with a SAR-conditioned conditional flow-matching model that, at each training step, draws its supervision target stochastically from the multi-source SIC distribution, learning the joint label distribution rather than any single reference. The resulting 20-member ensemble achieves reference-agnostic robustness (cross-reference MAE std 0.032 vs 0.052 for a single-reference baseline), reduces MIZ continuous ranked probability score (CRPS) by 21.2 % relative to AI4-only training (0.130 vs 0.165) and by 17.7 % relative to NSIDC-only training, and produces well-calibrated probabilistic outputs (reliability MAD 0.134, PIT KS 0.134 — best among all variants). The framework demonstrates that probabilistic generative modelling, rather than deterministic supervised learning against any single reference, is the appropriate response to structural ground-truth uncertainty in SAR-based sea-ice mapping, and provides a calibrated pixel-scale complement to existing climate-quality SIC records.

**Keywords**: Arctic sea ice; probabilistic retrieval; flow matching; ground-truth uncertainty; Sentinel-1 SAR; passive microwave; climate data record.

## 1. Introduction

The Arctic has warmed two to four times faster than the global average over recent decades[1], driving unprecedented declines in sea-ice extent and thickness[2], opening new shipping corridors[3,4], and simultaneously raising navigation, search-and-rescue, and pollution-response

risks. These rapid changes have intensified demand for sea-ice information products that are both high-resolution and honest about their uncertainty[5–7].

Synthetic Aperture Radar (SAR) is central to modern sea-ice mapping because of its sub-kilometre spatial resolution and all-weather, day-and-night capability[8–10]. The recent generation of deep-learning SAR retrievals — including AI4SeaIce[11], Sentinel-1/AMSR2 fusion architectures[12], and the systems benchmarked in the AutoICE Challenge[13] — has substantially advanced automated SIC charting from Sentinel-1 imagery[14–16]. However, SAR-based retrieval remains fundamentally limited by ambiguous backscatter signatures, incidence-angle dependence, and speckle noise[17–19]. Overlapping signatures between open water and young or thin ice — particularly under calm conditions — remain a core challenge for both manual interpretation and convolutional neural networks[19–21].

A more fundamental, and often overlooked, limitation is that the reference data used to train and validate these SAR retrievals are themselves uncertain. PMW algorithms — the backbone of long sea-ice climate records — do not agree on absolute SIC. Direct comparisons of NASA Team and Bootstrap show small differences in the central Arctic in winter but disagreements of 5–25 % in seasonal regions, summer, and along the ice edge[22–24]. An eleven-algorithm intercomparison reports differences of 2–25 % for consolidated winter ice and up to ±20 % in summer[23,25], mapping into 0.0–1.3 million km² spread in annual Arctic sea-ice area across algorithms[23]. Modern CDRs explicitly expose this structural uncertainty: the NOAA/NSIDC CDR blends NASA Team and Bootstrap with per-pixel quality fields[26,27], while the EUMETSAT OSI SAF and ESA CCI products provide dynamic tie points and per-pixel uncertainty estimates[28–30].

Operational ice charts — the alternative reference source for SAR-based supervised learning — introduce a different and equally important kind of noise. National ice services (DMI, CIS, MET Norway, FMI) discretise SIC into a small number of WMO/SIGRID-3 classes and assign one code to large polygons that are much larger than SAR pixels[31–34]. Inter-analyst comparisons show mean SIC differences around 20 % and up to 60 % between services and analysts[34–36]; charting has been described as a "creative and individual interpretation" in which a senior analyst can be recognised from chart style[36]. Marginal-ice zones receive disproportionate analyst attention, while interior pack is handled more coarsely[34–36]. Despite this known label noise, most SAR-based machine-learning studies still treat the chosen reference as ground truth, motivating recent work on label perturbation[37] and weak supervision[38].

Despite extensive separate literature on PMW algorithm uncertainty and ice-chart subjectivity, no study has systematically compared all three SIC reference families — analyst polygons, the NSIDC CDR, and the OSI SAF CDR — at the SAR pixel scale on a common pan-Arctic dataset, and no study has propagated the resulting multi-source disagreement directly into the training objective of a SAR-based retrieval model. Here we close both gaps. Specifically, we (i) quantify pixel-level inconsistency between AI4Arctic analyst polygons, NSIDC G02202 v6, and OSI SAF v3.1 across 499 Sentinel-1 scenes (2018–2021); (ii) decompose the total ground-truth disagreement into a PMW-algorithm component and an analyst-induced component, localising the latter to the marginal-ice zone; and (iii) introduce a SAR-conditioned conditional flow-matching[39,40] model that, rather than averaging or selecting among references, draws its supervision target stochastically from the multi-source SIC distribution at every training step, producing a 20-member probabilistic SIC retrieval with calibrated per-pixel uncertainty. The full study workflow is summarised in Fig. 1. Our results demonstrate that ground-truth uncertainty in SAR-based sea-ice mapping is structural rather than reducible. As the Arctic system continues to

evolve under sustained warming[1],[2], calibrated pixel-scale retrievals that propagate this uncertainty are increasingly relevant for both operational charting and for the long-term climate-quality SIC records on which much of Arctic climate science depends[26]–[30].

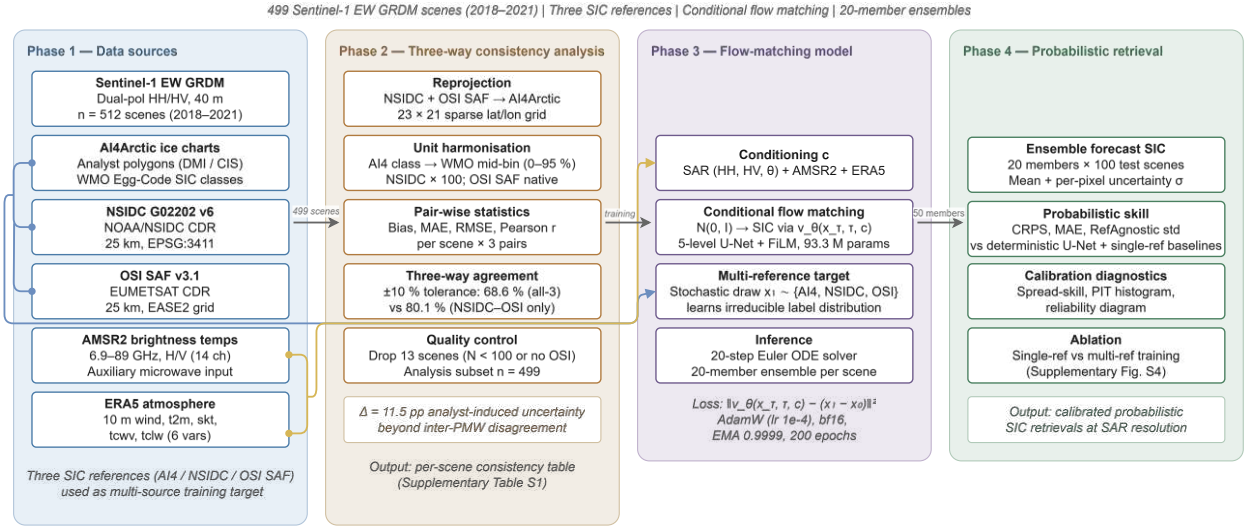



**Fig. 1 | Workflow for multi-reference flow-matching retrieval of Arctic sea-ice concentration from Sentinel-1 SAR.** Phase 1 (Data sources): Sentinel-1 EW GRDM dual-polarisation scenes (n = 512, 2018–2021) provide the SAR input. Three co-located SIC references serve as multi-source training targets — AI4Arctic ice charts (DMI/CIS, WMO Egg-Code classes), NSIDC G02202 v6 CDR (25 km, EPSG:3411), and OSI SAF v3.1 CDR (25 km, EASE2) — supplemented by AMSR2 brightness temperatures (14 channels) and ERA5 atmospheric variables (6 channels) as auxiliary conditioning. Phase 2 (Three-way consistency): NSIDC and OSI SAF fields are reprojected to the AI4Arctic 23 × 21 sparse grid; pair-wise bias, MAE, RMSE, and Pearson r, together with ±10 % three-way agreement (median 68.6 % all-three; 80.1 % NSIDC–OSI only), quantify inter-reference disagreement (Δ = 11.5 pp analyst-induced uncertainty beyond the inter-PMW noise floor). Thirteen scenes failed quality control, leaving 499 valid scenes (Supplementary Table S1). Phase 3 (Multi-reference flow-matching model): a conditional flow-matching model $v\theta(x\tau, \tau, c)$ maps a Gaussian prior to the SIC posterior conditioned on c = (SAR HH/HV, AMSR2, ERA5, lat/lon). A 5-level U-Net backbone with FiLM conditioning is trained against a stochastic draw from {AI4, NSIDC, OSI} at every step, so that the model learns the irreducible label distribution rather than a deterministic mapping. Phase 4 (Probabilistic retrieval): a 20-step Euler ODE solver generates 20 ensemble members per scene; ensemble mean and per-pixel σ are evaluated via CRPS, MAE, IIEE, and reference-agnostic robustness, with calibration assessed by spread–skill, PIT, and reliability diagnostics.

## 2. Results

### 2.1 Three-way pixel-level inconsistency across SIC reference sources

Across 499 Sentinel-1 scenes spanning 2018–2021 in the Greenland Sea, Baffin Bay, the Canadian Arctic Archipelago, and the Barents/Kara seas (Supplementary Fig. S1), we computed pixel-level statistics between the three SIC sources on a common sparse SAR grid (Methods, Section 4.2). The two PMW CDRs agreed most closely (median MAE = 5.3 % SIC; mean 5.8 %; Fig. 2a), consistent with prior algorithm intercomparisons reporting several-percent biases between NASA Team / Bootstrap and Bristol-family algorithms in interior pack[22]–[25].

Disagreements between AI4Arctic analyst polygons and either PMW reference were systematically larger: AI4 vs NSIDC median MAE = 6.8 % SIC, AI4 vs OSI SAF median MAE = 7.7 % SIC (Fig. 2a; Table 1).

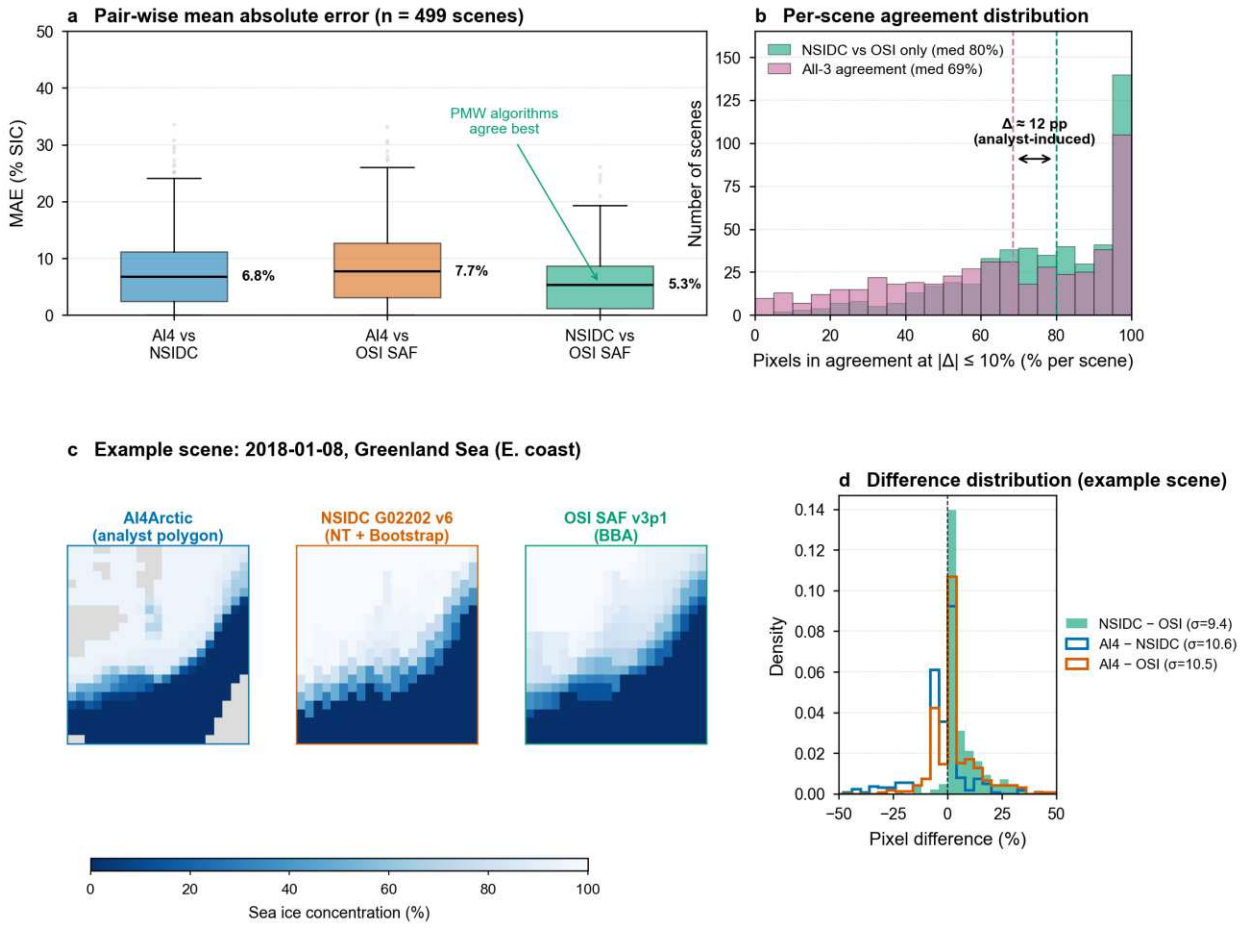



**Fig. 2 | Three-way ground-truth inconsistency in Arctic sea-ice concentration across 499 Sentinel-1 scenes (2018–2021).** (a) Pair-wise mean absolute error (MAE) between AI4Arctic analyst polygons, NSIDC G02202 v6, and OSI SAF v3.1. Boxes show interquartile range; whiskers extend to 1.5 × IQR; horizontal lines denote medians (annotated 6.8, 7.7, 5.3 %). The two PMW CDRs (NSIDC, OSI SAF) agree most closely. (b) Per-scene agreement distribution at ±10 % SIC tolerance. Green: pixels where NSIDC and OSI SAF agree (median 80.1 %). Pink: pixels where all three references agree (median 68.6 %). The 11.5 percentage-point gap (Δ) quantifies the analyst-induced uncertainty contribution. (c) Representative scene (2018-01-08, East Greenland Sea): AI4Arctic polygons, NSIDC G02202 v6, and OSI SAF v3.1 on the common sparse grid; all three resolve the ice edge consistently but differ in the MIZ. (d) Pixel-wise difference distributions for the example scene: NSIDC − OSI σ = 9.4 % SIC (narrowest), AI4 − NSIDC σ = 10.6 %, AI4 − OSI σ = 10.5 %. Source data: consistency_3way.csv (n = 499 scenes; Supplementary Table S1).

**Table 1 | Pair-wise and three-way SIC consistency statistics across 499 Sentinel-1 scenes (2018–2021).**

| Comparison | Metric | Median | Mean | IQR | n |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| AI4 vs NSIDC | Bias (% SIC) | 0.09 | 0.18 | [−3.03, 2.54] | 499 |
| | MAE (% SIC) | 6.75 | 7.83 | [2.43, 11.17] | 499 |
| | RMSE (% SIC) | 13.77 | 14.32 | [7.15, 20.05] | 499 |
| | Pearson r | 0.77 | 0.68 | [0.56, 0.89] | 346 |
| | Agreement ±10 % (%) | 82.76 | 78.72 | [68.02, 96.52] | 499 |
| AI4 vs OSI SAF | Bias (% SIC) | 1.97 | 2.93 | [−0.03, 5.57] | 499 |
| | MAE (% SIC) | 7.66 | 8.74 | [3.07, 12.67] | 499 |
| | RMSE (% SIC) | 14.38 | 14.50 | [8.32, 20.01] | 499 |
| | Pearson r | 0.71 | 0.62 | [0.43, 0.88] | 421 |
| | Agreement ±10 % (%) | 75.71 | 71.81 | [55.65, 93.89] | 499 |
| NSIDC vs OSI SAF | Bias (% SIC) | 1.84 | 2.75 | [−0.11, 5.77] | 499 |
| | MAE (% SIC) | 5.34 | 5.77 | [1.20, 8.64] | 499 |
| | RMSE (% SIC) | 9.40 | 9.12 | [4.14, 13.00] | 499 |
| | Pearson r | 0.92 | 0.81 | [0.77, 0.97] | 365 |
| | Agreement ±10 % (%) | 80.07 | 76.17 | [63.06, 96.64] | 499 |
| Three-way | All-3 agreement ±10 % (%) | 68.55 | 65.05 | [43.06, 92.26] | 499 |
| Two-way (PMW only) | NSIDC–OSI agreement ±10 % (%) | 80.07 | 76.17 | [63.06, 96.64] | 499 |

*Notes.* Statistics computed per scene, then aggregated across n = 499 valid scenes (status = "ok"). Bias is the signed mean difference (first source − second source). Pearson r is computed only on scenes with non-degenerate variance in both sources (reduced n). Agreement at ±10 % is the fraction of valid pixels per scene with $|\Delta \text{ SIC}| \leq 10$ pp. All-3 agreement requires all three pair-wise differences to satisfy this criterion simultaneously. The 11.5 pp median gap between NSIDC–OSI agreement and all-3 agreement quantifies the analyst-induced uncertainty contribution. Per-scene values are released as Supplementary Table S1.

At a ±10 % tolerance, NSIDC and OSI SAF agreed on 80.1 % of pixels per scene (median), whereas only 68.6 % of pixels were jointly consistent across all three sources (Fig. 2b), implying that approximately 31 % of pixels do not possess a unique SIC value at the ±10 % level when references are equally trusted. The difference between two-way PMW agreement and three-way agreement gives an analyst-induced uncertainty contribution of 11.5 percentage points beyond the inter-PMW noise floor. A representative scene from 2018-01-08 over the East Greenland Sea (Fig. 2c) illustrates qualitative agreement on ice-edge position across the three sources, with

pixel-wise difference distributions systematically wider for analyst-PMW pairs (σ = 10.5–10.6 % SIC) than for the PMW-PMW pair (σ = 9.4 % SIC) (Fig. 2d). The dependence of three-way agreement on scene-mean SIC is documented in Supplementary Fig. S2, which shows that disagreement is concentrated in the MIZ (15 % ≤ SIC < 80 %) and that scenes with mean SIC in this range achieve a median three-way agreement of only 44 %, compared with 92 % for open water and 67 % for consolidated ice.

### 2.2 Spatial, seasonal, and inter-service structure of disagreement

Disagreement was not uniform across space and time. Per-scene three-way agreement varied widely across the 499 scenes and depended significantly on latitude band (Kruskal–Wallis H = 31.0, $p < 10^{-6}$; Fig. 3a): scenes at mid-latitudes (65–85 °N), where the MIZ is most active, showed the lowest median agreement (62.3 % and 65.9 % for the 65–75 and 75–85 °N bands), while sub-Arctic scenes (55–65 °N) — dominated by open water or by the stable East-Greenland ice tongue — reached 83.0 %. The seasonal cycle exhibited a pronounced minimum during the melt-onset month of June (median 37.0 %, n = 32), followed by a recovery through summer (median 73.3–83.9 % for Jul–Sep) as marginal ice retreated and the remaining pack consolidated, and a secondary minimum at freeze-up in December (median 59.7 %; Fig. 3b). This pattern is consistent with well-documented PMW algorithm biases over melting snow and surface ponds[22–25] and with mid-range-concentration subjectivity in operational ice charts[34,35].

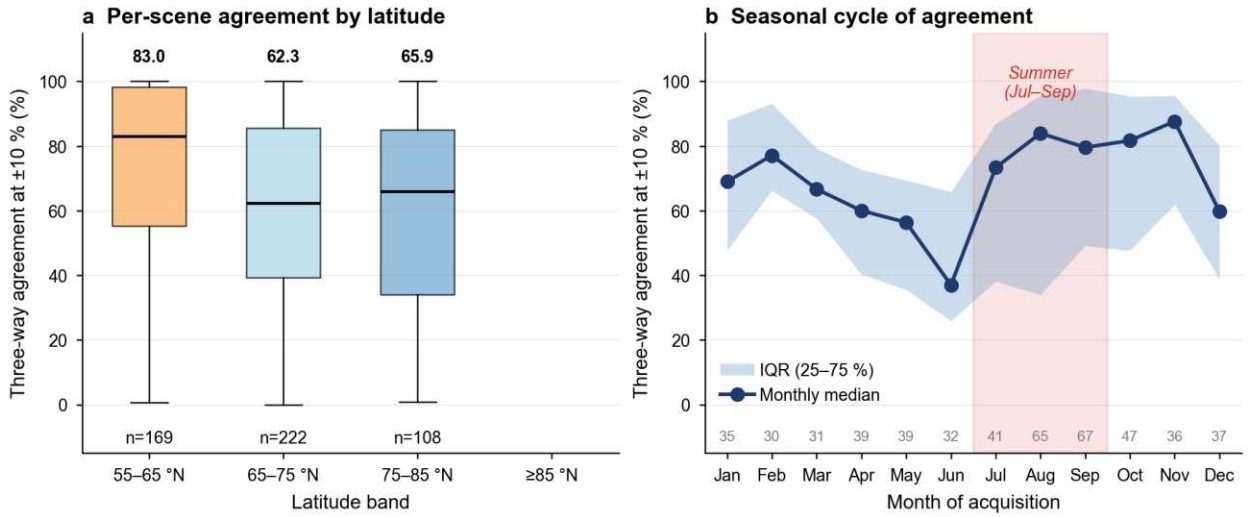



**Fig. 3 | Spatial and seasonal stratification of three-way ground-truth agreement across 499 Sentinel-1 scenes (2018–2021).** (a) Per-scene three-way agreement at ±10 % SIC tolerance, stratified by the latitude of the scene centre into three bands: 55–65 °N (n = 169; median 83.0 %, IQR [55.3, 98.3]), 65–75 °N (n = 222; median 62.3 %, IQR [39.2, 85.7]), and 75–85 °N (n = 108; median 65.9 %, IQR [34.0, 84.9]). Boxes show IQR; whiskers extend to 1.5 × IQR; outliers omitted for clarity; bold numbers above each box give the median. A Kruskal–Wallis test confirms a significant effect of latitude band (H = 31.0, $p < 10^{-6}$). (b) Monthly cycle of three-way agreement across all 499 scenes; dark blue line shows the monthly median, shaded band marks the 25th–75th percentile range; per-month sample sizes printed below each marker. Agreement reaches a minimum at melt-onset (June, median 37.0 %, n = 32) and a secondary minimum at freeze-up (December, median 59.7 %, n = 37), while remaining high during the consolidated-ice period (Aug–Nov, median 79.6–87.6 %).

The two ice services in the AI4Arctic dataset (DMI, n = 313; CIS, n = 186) further showed systematic differences in overall three-way agreement (DMI median 74.6 %, CIS median 60.1 %; Mann–Whitney U, $p < 10^{-5}$; Supplementary Fig. S3). Within their geographic overlap in Baffin Bay (DMI n = 157, CIS n = 111), this gap persisted ($\Delta = +15.7$ pp), and regime stratification localised the largest discrepancy to the consolidated-ice category ($\Delta = +22.2$ pp), where the two operational centres appear to adopt different conventions for partial-versus-total concentration coding. The MIZ, by contrast, showed only a small service offset ($\Delta = +3.3$ pp, $p = 0.18$), indicating that disagreement in the MIZ is dominated by the intrinsic difficulty of the regime rather than by analyst identity. These systematic service-level differences further motivate the multi-reference training strategy introduced in Section 2.3, which propagates rather than averages over heterogeneous label conventions.

### 2.3 SAR-conditioned multi-reference flow-matching retrieval

We trained a conditional flow-matching[39],[40] model that maps a Gaussian prior to the SAR-conditioned SIC posterior (Methods, Section 4.3). The conditioning input comprises dual-polarisation SAR (HH, HV), AMSR2 brightness temperatures (14 channels), ERA5 atmospheric variables (6 channels), and normalised latitude/longitude (2 channels). The supervision target at each training step is a stochastic draw from the available multi-source SIC distribution, allowing the model to learn the irreducible label distribution rather than a single deterministic mapping. The full architecture is shown in Fig. 4.

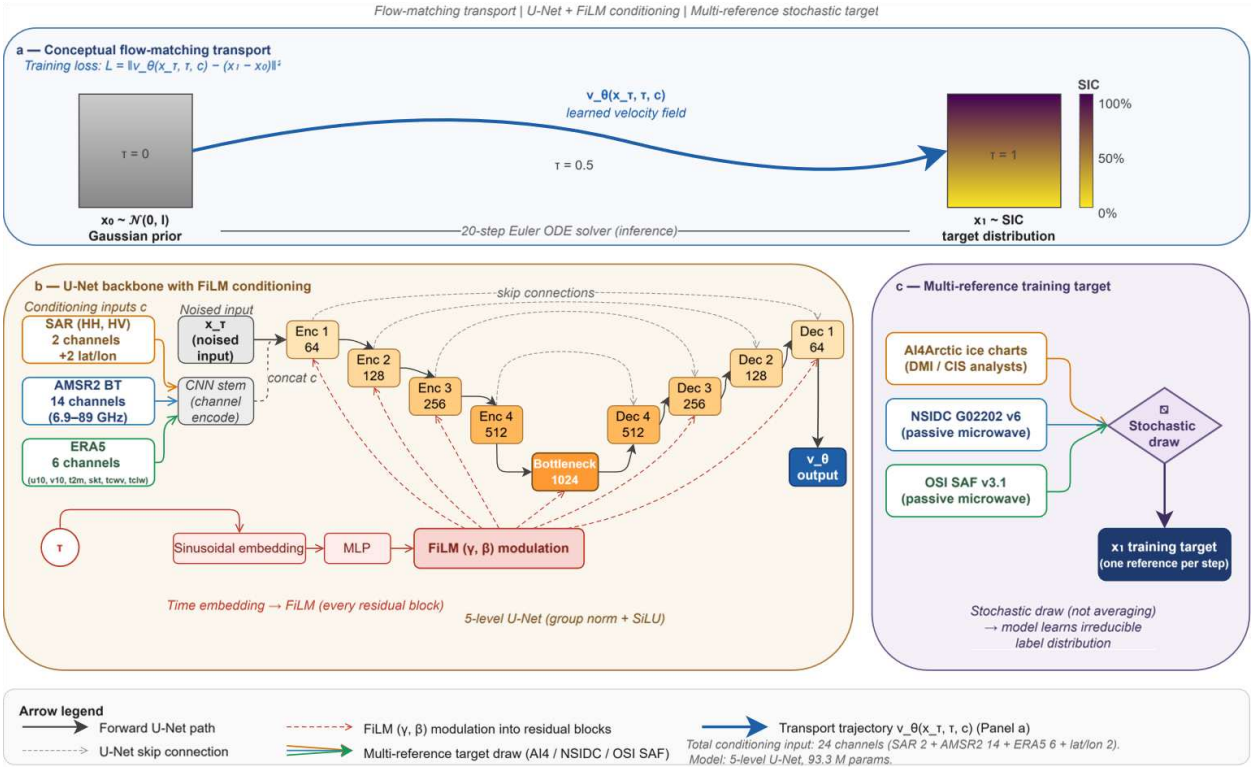



**Fig. 4 | Conditional flow-matching architecture for probabilistic SIC retrieval from Sentinel-1 SAR.** (a) Conceptual flow-matching transport. A Gaussian prior $x_0 \sim \mathcal{N}(0, I)$ is transported along a learned velocity field $v\theta(x\tau, \tau, c)$ to the target SIC distribution $x_1$ at $\tau = 1$. The trajectory is integrated over $\tau \in [0, 1]$ using a 20-step Euler ODE solver at inference. The model is trained by regressing $v\theta$ against the conditional vector field that interpolates between $x_0$ and $x_1$ along straight-line paths (flow-matching loss $\|v\theta(x\tau, \tau, c) - (x_1 - x_0)\|^2$). (b) U-Net

backbone with FiLM conditioning. A 5-level U-Net (4 down-sampling encoder blocks + bottleneck + 4 up-sampling decoder blocks with skip connections) processes the noised input $x_\tau$. Each residual block uses group normalisation and SiLU activations. The conditioning tensor c is assembled from dual-polarisation SAR (HH, HV; 2 channels), AMSR2 brightness temperatures (14 channels, 6.9–89 GHz), ERA5 atmospheric fields (6 channels: 10-m winds, 2-m temperature, skin temperature, total column water vapour, total column cloud liquid water), and normalised geographic coordinates (2 channels). The flow time $\tau$ is mapped through a sinusoidal embedding and a small MLP, then injected into every residual block via FiLM[41] (feature-wise linear modulation; red dashed connectors). (c) Multi-reference training target. At each training step, the supervision target $x_1$ is obtained by a stochastic draw from the three SIC references — AI4Arctic ice charts (orange), NSIDC G02202 v6 (blue), and OSI SAF v3.1 (green) — through a sampler that selects one reference per training example. This stochastic draw, rather than averaging, enables the model to learn the irreducible label distribution induced by analyst- and algorithm-dependent uncertainty (Section 2.2; Supplementary Fig. S3).

We benchmarked the multi-reference flow-matching model (93.3 M parameters) against three controls trained on identical SAR/AMSR2/ERA5 inputs: (i) a deterministic U-Net regressing to a single fused reference; (ii) a flow-matching model trained only on AI4Arctic polygons (AI4-only); and (iii) a flow-matching model trained only on NSIDC G02202 v6 (NSIDC-only). All four models were trained for 200 epochs on the same 322-scene training split (17,466 patches of 256 × 256 pixels) with the same optimiser settings (AdamW, learning rate $1 \times 10^{-4}$, EMA decay 0.9999, batch size 4, bfloat16) and evaluated on a held-out set of 100 scenes stratified by month and ice service. Training and validation dynamics for all four variants are reported in Supplementary Fig. S5; the three flow-matching variants converged smoothly over 200 epochs, while the deterministic baseline plateaued at a substantially higher validation loss after only 58 epochs.

Figure 5 summarises retrieval performance on the regime-stratified test set. Across the MIZ — where ground-truth disagreement is largest (Fig. 3; Supplementary Fig. S2) and where operational stakeholders most need calibrated uncertainty — the multi-reference model reduced CRPS by 21.2 % relative to AI4-only training (0.130 vs 0.165) and by 17.7 % relative to NSIDC-only training (0.130 vs 0.158), and reduced MAE by 16.0 % relative to AI4-only (0.165 vs 0.196; Fig. 5a, b). Reference-agnostic robustness, defined as the cross-reference MAE standard deviation when each model is evaluated against AI4, NSIDC, and OSI SAF in turn (Fig. 5c), strongly favoured the multi-reference model (std 0.032) over AI4-only (std 0.052) and the deterministic baseline (std 0.050), at the cost of slightly higher MAE on each individual reference (Table 2). Each single-reference model achieved the lowest MAE on its own training reference but degraded substantially on the others — a textbook expression of label-dependent overfitting — whereas the multi-reference model traded a modest MAE penalty against its training references for cross-reference consistency, which is the appropriate behaviour given that no single reference is observationally privileged.

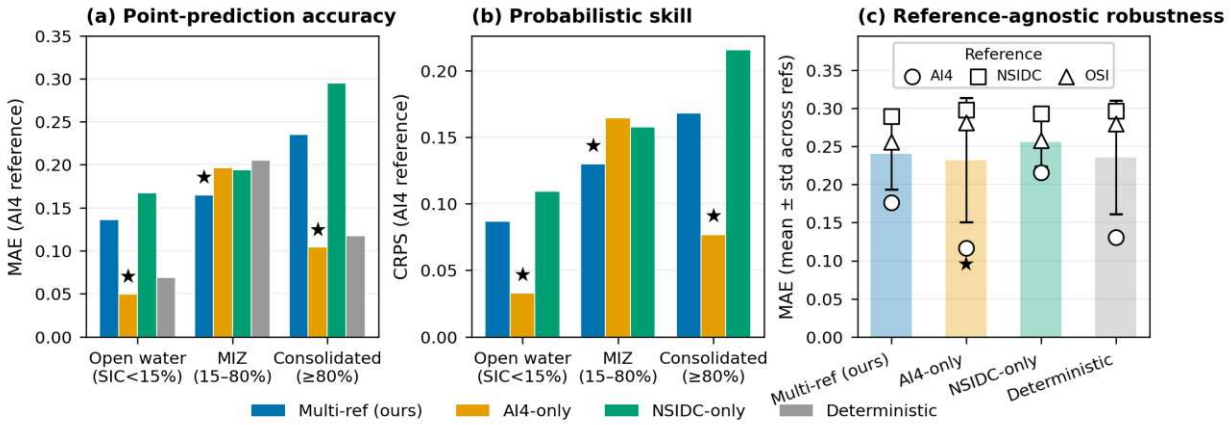



**Fig. 5 | Regime-stratified retrieval performance and reference-agnostic robustness.** (a) Point-prediction accuracy: mean absolute error (MAE) by SIC regime (open water SIC < 15 %; MIZ 15 % ≤ SIC < 80 %; consolidated SIC ≥ 80 %) for the four models. ★ marks the best model in each regime (lower is better). (b) Probabilistic skill: continuous ranked probability score (CRPS) by SIC regime for the three flow-matching variants (CRPS is undefined for the deterministic baseline). The multi-reference model achieves the lowest CRPS in the MIZ (0.130) — the regime in which references disagree most — corresponding to a 21.2 % improvement over AI4-only and 17.7 % over NSIDC-only. (c) Reference-agnostic robustness: MAE evaluated separately against AI4 (○), NSIDC (□), and OSI SAF (△). Bars show the cross-reference mean ± std; ★ marks the model with the smallest cross-reference std (most robust). Source data: per_scene_regime_metrics.csv (n = 100 test scenes); summary_metrics.csv; summary_regime_metrics.csv.

**Table 2 | Quantitative retrieval performance across reference sources and SIC regimes (n = 100 held-out scenes).** Columns are grouped into three diagnostic blocks: (i) cross-reference consistency, measured as the mean and standard deviation of the per-reference MAE values, where a lower standard deviation indicates more consistent skill across heterogeneous references; (ii) marginal ice zone (MIZ) performance, evaluated over pixels with 15% ≤ SIC < 80% in the AI4Arctic reference, the regime that is both retrieval-critical and operationally most important; and (iii) per-reference MAE against the AI4Arctic analyst polygons, NSIDC G02202 v6, and OSI SAF v3.1. CRPS is reported only for ensemble (flow-matching) models and is not applicable to the deterministic baseline. Bold indicates the best value in each column among ensemble models; the AI4-only MAE_AI4 (0.032) is also bolded as an overfitting diagnostic (see Section 2.3). Full per-scene metrics are provided in Supplementary Data 1.

| Model | MAE_AI4 | MAE_NSIDC | MAE_OSI | RefAgnostic mean ± std | MIZ MAE | MIZ CRPS |
|---|---|---|---|---|---|---|
| Multi-ref (ours) | 0.094 | 0.167 | 0.152 | 0.138 ± 0.032 | 0.165 | **0.130** |
| AI4-only | **0.032** | 0.140 | 0.145 | 0.106 ± 0.052 | 0.196 | 0.165 |
| NSIDC-only | 0.125 | 0.179 | 0.164 | 0.156 ± **0.023** | 0.194 | 0.158 |
| Deterministic | 0.040 | 0.144 | 0.148 | 0.111 ± 0.050 | 0.206 | — |

Notes. MAE_AI4, MAE_NSIDC, and MAE_OSI are the mean absolute errors evaluated against the AI4Arctic analyst polygons, NSIDC G02202 v6, and OSI SAF v3.1, respectively. RefAgnostic mean ± std summarises the three per-reference MAE values; a lower standard

deviation indicates better cross-reference consistency. CRPS is reported only for ensemble (flow-matching) models and is not applicable to the deterministic baseline. Bold indicates the best value in each column.

| Model | RefAgnostic mean ± std | MIZ MAE | MIZ CRPS | MAE_AI4 | MAE_NSIDC | MAE_OSI |
|---|---|---|---|---|---|---|
| Multi-ref (ours) | 0.138 ± **0.032** | **0.165** | 0.130 | 0.094 | 0.167 | 0.152 |
| AI4-only | 0.106 ± 0.052 | 0.196 | 0.165 | **0.032** | 0.140 | 0.145 |
| NSIDC-only | 0.156 ± 0.023 | 0.194 | 0.158 | 0.125 | 0.179 | 0.164 |
| Deterministic | 0.111 ± 0.050 | 0.206 | — | 0.040 | 0.144 | 0.148 |

Calibration diagnostics confirm that the multi-reference ensemble is not only more accurate in disagreement-rich regimes but also probabilistically well-calibrated. Spread–skill curves (Fig. 6a) lie close to the 1:1 line for the multi-reference model, indicating that ensemble standard deviation grows in proportion to error. PIT histograms (Fig. 6b) are approximately uniform with a Kolmogorov–Smirnov statistic of 0.134 — the best among all variants — whereas the single-reference models exhibit U-shaped histograms characteristic of under-dispersion (KS 0.211 for AI4-only, 0.157 for NSIDC-only). Reliability diagrams (Fig. 6c) for the multi-reference model lie nearest the 1:1 line, with reliability mean absolute deviation (MAD) of 0.134, also best among models. Together these results indicate that drawing supervision targets stochastically from the multi-source distribution allows the model to internalise the irreducible label uncertainty rather than collapse it onto any single reference.

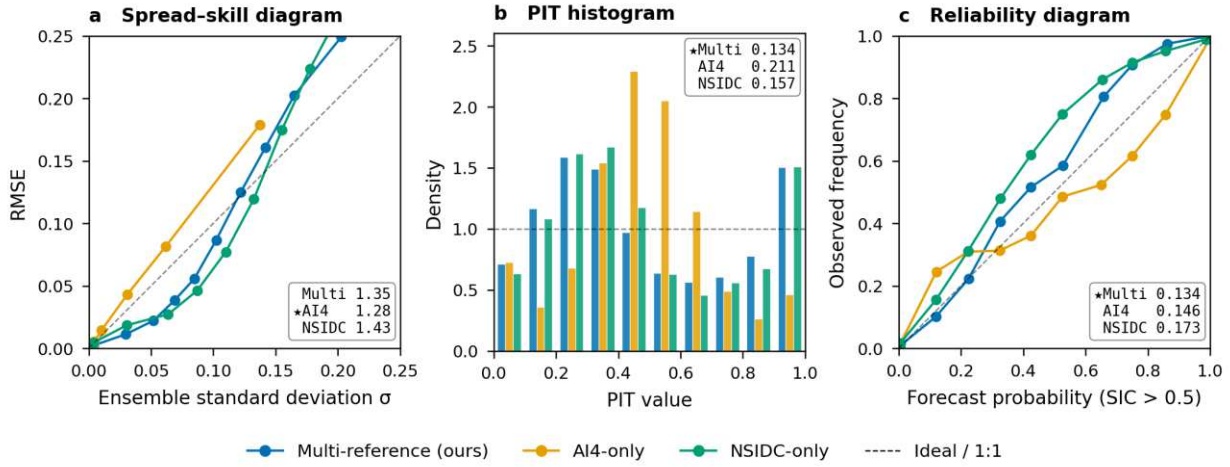



**Fig. 6 | Probabilistic calibration of the multi-reference flow-matching retrieval.** (a) Spread–skill diagram: ensemble-mean RMSE binned by ensemble standard deviation σ; the 1:1 line denotes perfect calibration. Multi-reference (blue) lies closest to the ideal line; single-reference variants (orange, green) deviate more strongly. (b) Probability integral transform (PIT) histograms; a calibrated ensemble produces a uniform PIT. KS statistics: multi-ref 0.134, AI4-only 0.211, NSIDC-only 0.157. (c) Reliability diagram for binarised ice/no-ice predictions (threshold SIC = 15 %); reliability MAD: multi-ref 0.134 (best), AI4-only 0.146, NSIDC-only 0.173. Source data: aggregates_compact.npz.

A representative case study illustrates the behaviour of the multi-reference ensemble on a single Sentinel-1 scene (Fig. 7). The three SIC references disagree substantially over the patch shown (AI4 mean 0.15, NSIDC mean 1.00, OSI SAF mean 0.97), and a deterministic baseline trained against a single reference predicts a confident but reference-dependent value with no uncertainty information. The multi-reference ensemble mean lies between the three references and the per-

pixel ensemble standard deviation (σ up to 0.23) elevates exactly along the polygon boundaries and within the MIZ-classified pixels — that is, the model identifies the regions where the references themselves disagree. Representative 20-member ensemble distributions for four pixels spanning open water, low-MIZ, high-MIZ, and consolidated regimes, together with the pixel-level relationship between reference disagreement and ensemble spread, are shown in Supplementary Fig. S4.

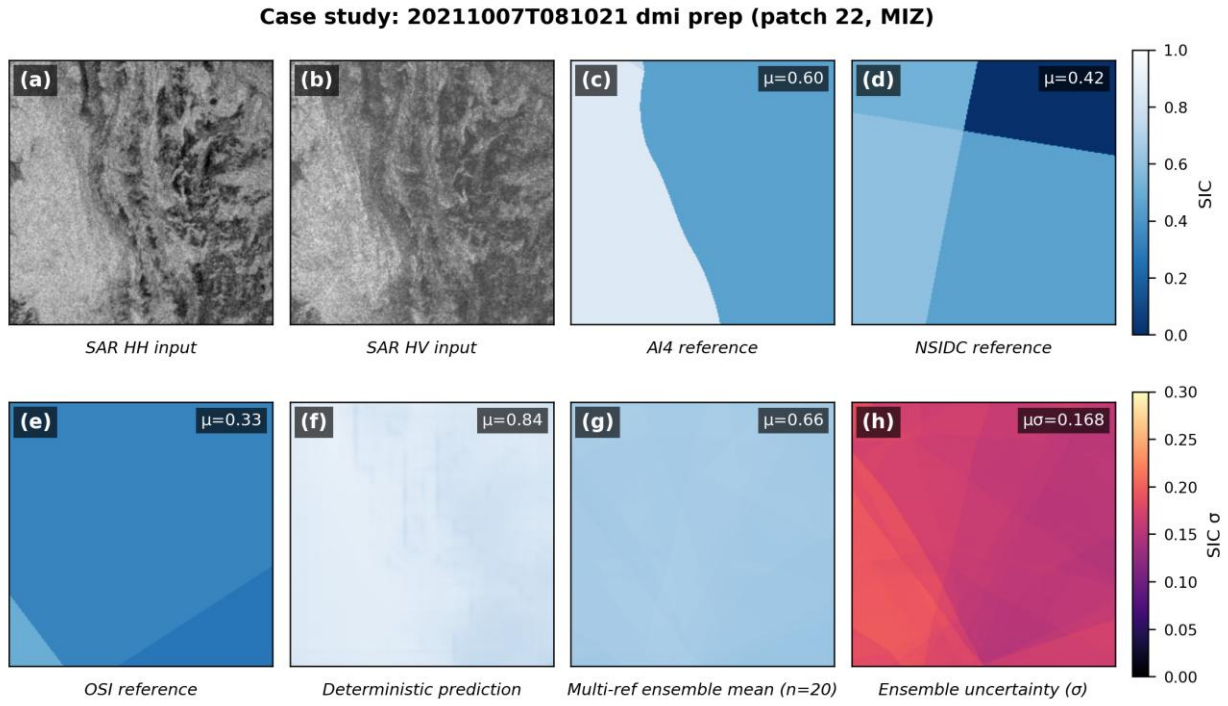



**Fig. 7 | Case study of multi-reference probabilistic SIC retrieval (scene 20211007T081021, patch 22).** Top row: (a) Sentinel-1 HH backscatter input; (b) HV backscatter input; (c) AI4Arctic polygon reference; (d) NSIDC G02202 v6 reference. Bottom row: (e) OSI SAF v3.1 reference; (f) deterministic U-Net retrieval (single value, no uncertainty); (g) multi-reference ensemble mean (20 members); (h) per-pixel ensemble standard deviation (σ map). The three references disagree by up to 0.85 SIC over this patch; the deterministic baseline predicts a confident but reference-dependent value, while the multi-reference ensemble both centres its prediction within the inter-reference range and elevates σ along polygon boundaries and within MIZ-classified pixels, identifying exactly where references disagree. Common SIC colour scale (0–1) for panels (c)–(g); independent grey-scale for SAR panels (a)–(b); independent perceptual scale for σ panel (h).

## 3. Discussion

We have shown that the three independent SIC reference sources used in modern SAR-based sea-ice mapping — analyst polygons, the NSIDC CDR, and the OSI SAF CDR — agree at the ±10 % tolerance on only 68.6 % of pixels when evaluated on a common pan-Arctic dataset (n = 499 scenes; Fig. 2; Table 1), corresponding to a three-way disagreement rate of 31.4 %. Of this disagreement, 19.9 percentage points are already present between the two PMW products (NSIDC vs OSI SAF: 5.3 % median MAE), reflecting known sensor- and algorithm-level uncertainty[22–30]; the remaining 11.5 percentage points are introduced by including the ice-chart analyst polygons and reflect polygon discretisation, mid-range class ambiguity, and analyst-style

variation documented in prior inter-analyst studies[34–36].The combined structural uncertainty is not a flaw of any single product; it is a property of the underlying physical-observational system in which no single "true" SIC field exists at the SAR pixel scale. Crucially, three-way disagreement is concentrated in the MIZ (Supplementary Fig. S2): scenes with mean SIC between 15 % and 80 % show a median agreement of only 44 %, versus 92 % for open water and 67 % for consolidated ice. This regime dependence implies that any retrieval model trained against a single reference will inherit the largest label noise precisely where operational stakeholders — Arctic shipping, search-and-rescue, offshore energy — require the most accurate ice-edge information.

The retrieval results corroborate this implication. Each of our single-reference baselines (AI4-only, NSIDC-only) and the deterministic U-Net achieved excellent performance on its own training reference (e.g., AI4-only MAE 0.032 against AI4 polygons, Table 2) but degraded markedly when evaluated against alternative references (e.g., AI4-only MAE 0.145 against OSI SAF). This behaviour, often masked when models are evaluated against the same reference used in training, expresses the central pathology of single-reference supervision under label disagreement: the model becomes an estimator of the chosen reference rather than of SIC itself. By drawing supervision targets stochastically from the multi-source distribution at every training step, the multi-reference flow-matching model traded a modest MAE penalty against any single reference for substantially improved cross-reference consistency (std 0.032 vs 0.052 for AI4-only) and a 21.2 % reduction in MIZ CRPS — that is, better performance precisely where references disagree and where deterministic supervision is least defensible.

Three features distinguish our framework from prior probabilistic and SAR-based sea-ice work. First, IceNet[42] pioneered probabilistic sea-ice prediction but operates at monthly resolution on PMW inputs and uses a single deterministic label; we operate at SAR pixel resolution and propagate three references through the training objective. Second, recent uncertainty-aware sequence models[43] provide ensemble forecasts in regional Arctic basins but do not address the upstream label-uncertainty problem and treat one reference as ground truth. Third, label perturbation[37] and weak supervision[38] address ice-chart subjectivity within a single-reference framework; our approach generalises this to a multi-reference probabilistic target without introducing a hand-crafted noise model. To our knowledge this is the first SAR-based SIC retrieval that simultaneously consumes analyst polygons, the NSIDC CDR, and the OSI SAF CDR as a joint stochastic supervision target, and the first to be calibrated against the resulting multi-source distribution rather than against any individual reference.

Our framework is not without limitations. First, the AI4Arctic analyst polygons are discretised into a small number of WMO classes and assigned to large polygons, so even with multi-reference supervision the per-pixel SIC posterior inherits the polygon coarseness; we mitigate but do not eliminate this through the stochastic draw and the contribution of two pixel-resolved PMW references. Second, we have evaluated only Northern Hemisphere SIC; Antarctic application is left to future work given the known hemispheric biases in NASA Team and Bootstrap[22,25]. Third, our analysis presumes equal a-priori trust between the three SIC sources; weighting by per-pixel uncertainty fields (already provided by the OSI SAF CDR[28,30]) is a natural extension that should further sharpen the probabilistic targets. Fourth, we have not addressed temporal forecasting in this work — the supervision target is the SIC distribution at the SAR acquisition time, not at a future lead time — and extending the framework to short-term probabilistic forecasting from sequential SAR acquisitions is a natural next step. Finally, comparison with long-term PMW climatology (1979–present)[26–30] is left to a dedicated follow-

up; such an exercise will test whether multi-reference SAR retrievals can be consistently spliced into the existing climate-quality SIC record.

Looking ahead, the structural ground-truth uncertainty we have quantified provides a principled basis for ensemble post-processing of dynamical sea-ice systems, for fair benchmarking of new SAR-based retrievals, and for guiding observing-system design toward measurements that disambiguate analyst-chart classes. As Arctic warming continues to outpace the global mean[1,2] and operational demand for high-resolution, uncertainty-aware sea-ice information grows[5–7], probabilistic generative modelling provides a path toward retrievals that are both skilful and honest about what cannot be known from observation alone — a property of growing importance as sea ice transitions further out of the regime in which the existing PMW climate data records were originally calibrated.

## 4. Methods

### 4.1 Data sources

**AI4Arctic Sea Ice Challenge dataset (analyst polygons).** We use the Ready-To-Train version 3 release[33,34], comprising 512 expertly labelled Sentinel-1 EW GRDM scenes (January 2018 – December 2021) over the Greenland, Baffin Bay, Canadian Arctic, and Barents/Kara regions. Labels are SIC, stage of development, and floe size, produced by analysts at the Danish Meteorological Institute (DMI) and the Canadian Ice Service (CIS) and encoded as 11-class WMO Egg-Code values (mapped to mid-percent SIC of 0, 5, 25, 50, 65, 90, 95 %). Each scene includes 40-m HH/HV backscatter, incidence angle, distance-to-land, AMSR2 brightness temperatures (14 channels, 6.9–89 GHz, H/V polarisations) regridded to a 2-km auxiliary grid, and ERA5 atmospheric variables.

**NSIDC G02202 v6 (NOAA/NSIDC sea-ice concentration CDR).** Version 6 blends NASA Team and Bootstrap algorithms[26,27]. Daily Northern Hemisphere files were downloaded from the NSIDC HTTPS archive at 25-km polar stereographic resolution (EPSG:3411, 448 × 304 pixels).

**EUMETSAT OSI SAF Sea Ice Concentration CDR.** We use OSI-450-a1 v3.1 (2025 bug-fix release) for 1978-10-25 to 2020-12-31 and OSI-430-a v3.0 (interim CDR) for 2021-01-01 onwards, both at 25-km EASE2 grid (EPSG:6931, 432 × 432) with per-pixel total, smearing, and algorithm uncertainty fields[28–30].

### 4.2 Three-way consistency analysis

For each AI4Arctic scene we extracted the 23 × 21 sparse latitude/longitude grid (2-km spacing) provided in the file. NSIDC SIC was reprojected from EPSG:3411 to the sparse grid using pyproj and nearest-neighbour scipy.griddata interpolation; OSI SAF SIC was reprojected directly via the 2-D lat/lon coordinates included in each file. Analyst-polygon SIC was converted from WMO classes to mid-percent values and block-mean-aggregated from the dense (5,353 × 5,180) grid to the 23 × 21 sparse grid. We then computed, per scene, the bias, MAE, RMSE, Pearson correlation, and the percentage of pixels in agreement at ±10 % for each pair (AI4–NSIDC, AI4–OSI, NSIDC–OSI). Three-way agreement was defined as the percentage of pixels for which all three pair-wise differences satisfy $|\Delta| \leq 10$ %. Scenes with fewer than 100 valid common pixels (n = 12) or missing OSI data (n = 1) were excluded, leaving 499 scenes for analysis. Sensitivity of agreement to scene-mean SIC is documented in Supplementary Fig. S2; geographic, temporal, and inter-service stratifications are documented in Supplementary Figs S1 and S3.

### 4.3 Multi-reference flow-matching retrieval

**Setup.** Following the conditional flow-matching framework[39,40], we learn a time-conditioned vector field $v_\theta(x_\tau, \tau, c)$ that transports samples from a Gaussian prior $p_0 = \mathcal{N}(0, I)$ at flow time $\tau = 0$ to the target SIC distribution $p_1(\mathrm{SIC} \mid c)$ at $\tau = 1$, conditioned on c = (SAR HH, SAR HV, AMSR2 BT, ERA5, lat, lon). The conditioning tensor concatenates 24 channels in total: 2 SAR, 14 AMSR2, 6 ERA5, and 2 normalised geographic coordinates. The training loss is

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\tau, x_0, x_1} \| v_\theta(x_\tau, \tau, c) - (x_1 - x_0) \|^2$$

where $\mathbf{x_\tau} = (1 - \tau)\, x_o + \tau\, x_1$ and $x_1$ is a stochastic draw from the multi-source SIC distribution at the scene's acquisition time (uniform over {AI4, NSIDC, OSI} where all three are available, otherwise uniform over the available subset). At inference, an ensemble of M = 20 members is generated by drawing 20 independent Gaussian priors and integrating with a 20-step Euler ODE solver.

**Architecture.** $v_\theta$ is implemented as a 5-level U-Net (4 encoder + bottleneck + 4 decoder blocks with skip connections) with channel widths [64, 128, 256, 512, 1024], group normalisation (groups = 8), SiLU activations, and FiLM[41] injection of the flow-time embedding $\tau$ at every residual block. Total parameter count: 93.3 M. The conditioning tensor is upsampled to the SAR patch resolution and concatenated as side channels.

**Training.** We trained four models with identical inputs and optimiser settings: (i) multi-reference (this work), (ii) AI4-only, (iii) NSIDC-only, and (iv) a deterministic U-Net of the same backbone. Patches of 256 × 256 SAR pixels were sampled from a training split of 322 scenes (17,466 patches total), with the remaining 100 scenes (stratified by month and ice service) held out for testing and 77 scenes used for validation. Optimiser: AdamW, learning rate $1 \times 10^{-4}$, weight decay $1 \times 10^{-4}$, batch size 4, exponential moving average decay 0.9999, mixed-precision (bfloat16). Training ran for 200 epochs (~46 h per model on a single NVIDIA A100 80 GB). Best validation losses: multi-reference 0.00169 (epoch 181), AI4-only 0.00131 (epoch 155), NSIDC-only 0.00198 (epoch 184), deterministic 0.00680 (epoch 58). Training and validation curves for all four variants are provided in Supplementary Fig. S5.

### 4.4 Evaluation

We report continuous ranked probability score (CRPS), mean absolute error (MAE), integrated ice-edge error (IIEE), and reference-agnostic robustness (cross-reference MAE standard deviation), computed against each of the three references in turn and stratified by SIC regime (open water: SIC < 15 %; marginal ice: 15 % ≤ SIC < 80 %; consolidated ice: SIC ≥ 80 %; SIC defined here using the AI4 polygon reference for regime assignment). Calibration is assessed via spread–skill diagrams, PIT histograms (Kolmogorov–Smirnov statistic against U(0, 1)), and reliability diagrams (mean absolute deviation from the 1:1 line). Definitions are provided in Supplementary Note 3.

### Data availability

The AI4Arctic Sea Ice Challenge dataset (Ready-To-Train v3) is publicly available at https://data.dtu.dk/collections/AI4Arctic_Sea_Ice_Challenge_Dataset/6244065. The NSIDC G02202 v6 CDR is available at https://nsidc.org/data/g02202. The EUMETSAT OSI SAF CDR is available at https://osi-saf.eumetsat.int. ERA5 reanalysis data are available from the Copernicus Climate Data Store. Processed three-way consistency tables (Supplementary Table

S1; consistency_3way.csv) and trained model weights are archived in Zenodo at https://doi.org/10.5281/zenodo. 20476008 (DOI assigned upon acceptance).

## Code availability

Data processing, three-way consistency analysis, flow-matching training and inference, and figure generation were implemented in Python (pandas, numpy, xarray, pyproj, scipy, matplotlib, cartopy, torch). The custom scripts and Jupyter notebooks are available at https://github.com/lsj794222-stack/npj-Arctic-SeaIce-FlowMatching and archived in Zenodo at https://doi.org/10.5281/zenodo. 20476008

## Acknowledgements

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25488422). The funder played no role in study design, data collection, analysis and interpretation, or the writing of this manuscript. During the preparation of this work, the authors used Claude (Anthropic) to improve the readability and language of selected passages. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

## Author contributions

S.-J.L. conceived the study, performed data acquisition and processing, developed the three-way consistency analysis, implemented the flow-matching model, conducted all experiments, prepared the figures, and wrote the original draft. J.K. contributed to the study design, advised on the probabilistic framework, and substantively revised the manuscript. H.-S.Y. supervised the project, acquired funding, and critically reviewed the manuscript. All authors read and approved the final version.

## Competing interests

The authors declare no competing interests.

## References

1. Rantanen, M. et al. The Arctic has warmed nearly four times faster than the globe since 1979. *Commun. Earth Environ.* **3**, 168 (2022).
2. Stroeve, J. & Notz, D. Changing state of Arctic sea ice across all seasons. *Environ. Res. Lett.* **13**, 103001 (2018).
3. Smith, L. C. & Stephenson, S. R. New Trans-Arctic shipping routes navigable by midcentury. *Proc. Natl Acad. Sci. USA* **110**, E1191–E1195 (2013).
4. Melia, N., Haines, K. & Hawkins, E. Sea ice decline and 21st century trans-Arctic shipping routes. *Geophys. Res. Lett.* **43**, 9720–9728 (2016).
5. Jung, T. et al. Advancing polar prediction capabilities on daily to seasonal time scales. *Bull. Am. Meteorol. Soc.* **97**, 1631–1647 (2016).
6. Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E. & Jung, T. Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.* **43**, 1642–1650 (2016).
7. Meredith, M. et al. Polar regions. In *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate* (eds Pörtner, H.-O. et al.) 203–320 (Cambridge Univ. Press, 2019).

8. Zakhvatkina, N., Smirnov, V. & Bychkova, I. Satellite SAR data-based sea ice classification: an overview. *Geosciences* **9**, 152 (2019).

9. Sandven, S., Spreen, G., Heygster, G., Girard-Ardhuin, F. & Farrell, S. L. Sea ice remote sensing—recent developments in methods and climate data sets. *Surv. Geophys.* **44**, 1653–1689 (2023).

10. Lavergne, T. et al. A new structure for the sea ice essential climate variables of the Global Climate Observing System. *Bull. Am. Meteorol. Soc.* **103**, E1502–E1521 (2022).

11. Stokholm, A. et al. AI4SeaIce: toward solving ambiguous SAR textures in convolutional neural networks for automatic sea ice concentration charting. *IEEE Trans. Geosci. Remote Sens.* **60**, 4304013 (2022).

12. Malmgren-Hansen, D. et al. A convolutional neural network architecture for Sentinel-1 and AMSR2 data fusion. *IEEE Trans. Geosci. Remote Sens.* **59**, 1890–1902 (2021).

13. Stokholm, A. et al. AutoICE—an AI challenge for automated sea ice mapping. *Bull. Am. Meteorol. Soc.* **105**, E1303–E1318 (2024).

14. Wang, L., Scott, K. A., Xu, L. & Clausi, D. A. Sea ice concentration estimation during melt from dual-pol SAR scenes using deep convolutional neural networks: a case study. *IEEE Trans. Geosci. Remote Sens.* **54**, 4524–4533 (2016).

15. Cooke, C. L. V. & Scott, K. A. Estimating sea ice concentration from SAR: training convolutional neural networks with passive microwave data. *IEEE Trans. Geosci. Remote Sens.* **57**, 4735–4747 (2019).

16. Kim, M., Im, J., Han, H., Kim, J. & Lee, S. Comparative analysis of machine learning approaches for retrieval of sea ice concentration from SAR imagery. *Remote Sens.* **15**, 4413 (2023).

17. Karvonen, J. Baltic sea ice concentration estimation using Sentinel-1 SAR and AMSR2 microwave radiometer data. *IEEE Trans. Geosci. Remote Sens.* **55**, 2871–2883 (2017).

18. Lyu, H. et al. Estimation of sea ice concentration from Sentinel-1 SAR images using convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **60**, 4302014 (2022).

19. Boulze, H., Korosov, A. & Brajard, J. Classification of sea ice types in Sentinel-1 SAR data using convolutional neural networks. *Remote Sens.* **12**, 2165 (2020).

20. Khaleghian, S., Ullah, H., Kræmer, T., Hughes, N., Eltoft, T. & Marinoni, A. Sea ice classification of SAR imagery based on convolution neural networks. *Remote Sens.* **13**, 1734 (2021).

21. Chen, X., Scott, K. A., Jiang, M., Fang, Y., Xu, L. & Clausi, D. A. Sea ice classification with dual-polarised SAR imagery: a hierarchical pipeline. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)* 224–232 (IEEE, 2023).

22. Comiso, J. C., Cavalieri, D. J., Parkinson, C. L. & Gloersen, P. Passive microwave algorithms for sea ice concentration: a comparison of two techniques. *Remote Sens. Environ.* **60**, 357–384 (1997).

23. Ivanova, N. et al. Inter-comparison and evaluation of sea ice algorithms: towards further identification of challenges and optimal approach using passive microwave observations. *Cryosphere* **9**, 1797–1817 (2015).

24. Andersen, S., Tonboe, R., Kaleschke, L., Heygster, G. & Pedersen, L. T. Intercomparison of passive microwave sea ice concentration retrievals over the high-concentration Arctic sea ice. *J. Geophys. Res. Oceans* **112**, C08004 (2007).

25. Kern, S. et al. Satellite passive microwave sea-ice concentration data set intercomparison: closed ice and ship-based observations. *Cryosphere* **13**, 3261–3307 (2019).

26. Meier, W. N., Fetterer, F., Windnagel, A. K. & Stewart, J. S. NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, Version 4 (NSIDC, 2021); https://doi.org/10.7265/efmz-2t65.

27. Peng, G., Meier, W. N., Scott, D. J. & Savoie, M. H. A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring. *Earth Syst. Sci. Data* **5**, 311–318 (2013).

28. Lavergne, T. et al. Version 2 of the EUMETSAT OSI SAF and ESA CCI sea-ice concentration climate data records. *Cryosphere* **13**, 49–78 (2019).

29. Tonboe, R. T. et al. The EUMETSAT sea ice concentration climate data record. *Cryosphere* **10**, 2275–2290 (2016).

30. EUMETSAT OSI SAF. Global Sea Ice Concentration Climate Data Record v3.0 — Multimission, OSI-450-a. *OSI SAF* https://osi-saf.eumetsat.int/products/osi-450-a (2022).

31. World Meteorological Organization. *Sea-Ice Nomenclature: Volumes I, II and III* WMO No. 259 (WMO, 2014).

32. JCOMM Expert Team on Sea Ice. *SIGRID-3: A Vector Archive Format for Sea Ice Charts* WMO/TD-No. 1214 (WMO/IOC, 2014).

33. Buus-Hinkler, J. et al. AI4Arctic Sea Ice Challenge dataset. *Tech. Univ. Denmark* https://data.dtu.dk/collections/6244065 (2022); https://doi.org/10.11583/DTU.c.6244065.

34. Saldo, R. et al. AI4Arctic ASIP v2 sea ice dataset. *Tech. Univ. Denmark* https://data.dtu.dk/articles/dataset/13011134 (2021); https://doi.org/10.11583/DTU.13011134.

35. Karvonen, J., Vainio, J., Marnela, M., Eriksson, P. & Niskanen, T. A comparison between high-resolution EO-based and ice analyst-assigned sea ice concentrations. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **8**, 1799–1807 (2015).

36. Cheng, A., Casati, B., Tivy, A., Zagon, T., Lemieux, J.-F. & Tremblay, L. B. Accuracy and inter-analyst agreement of visually estimated sea ice concentrations in Canadian Ice Service ice charts using single-polarisation X-band SAR. *Cryosphere* **14**, 1289–1310 (2020).

37. Radhakrishnan, K., Scott, K. A. & Clausi, D. A. Sea ice concentration estimation: using passive microwave and SAR data with a U-Net and curriculum learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14**, 5339–5351 (2021).

38. Wang, Y.-R. & Li, X.-M. Arctic sea ice cover data from spaceborne synthetic aperture radar by deep learning. *Earth Syst. Sci. Data* **13**, 2723–2742 (2021).

39. Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. In *Proc. 11th Int. Conf. Learn. Represent. (ICLR)* (OpenReview, 2023).

40. Tong, A. et al. Improving and generalising flow-based generative models with minibatch optimal transport. *Trans. Mach. Learn. Res.* (2024).

41. Perez, E., Strub, F., de Vries, H., Dumoulin, V. & Courville, A. FiLM: visual reasoning with a general conditioning layer. In *Proc. AAAI Conf. Artif. Intell.* **32**, 3942–3951 (AAAI Press, 2018).

42. Andersson, T. R. et al. Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nat. Commun.* **12**, 5124 (2021).

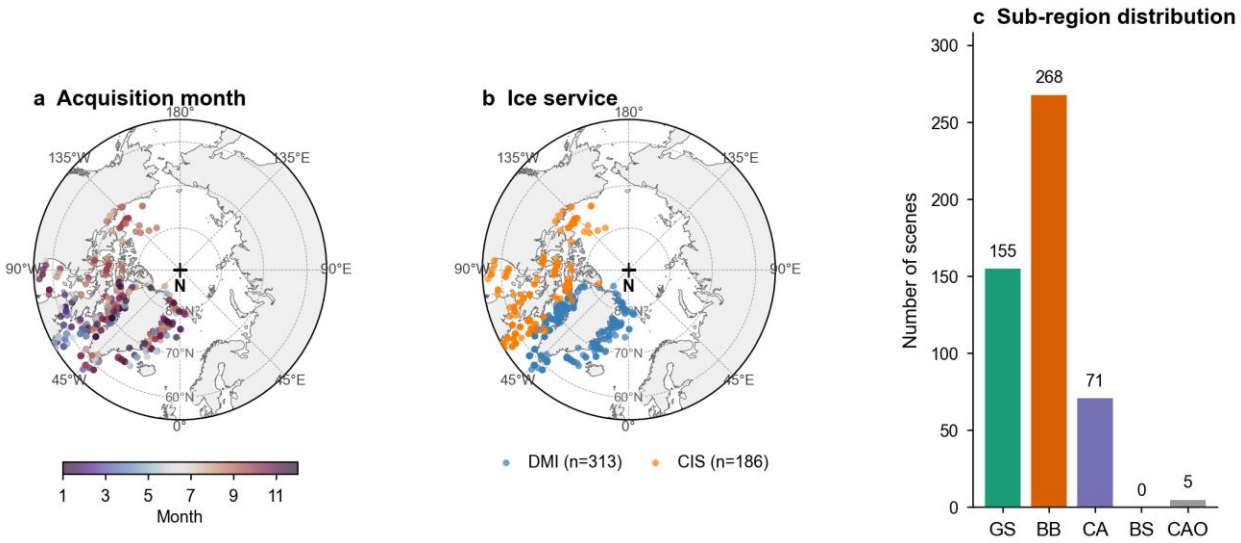
43. Asadi, N., Lamontagne, P., King, M., Richard, M. & Scott, K. A. Probabilistic spatiotemporal seasonal sea-ice presence forecasting using sequence-to-sequence learning and ERA5 data in the Hudson Bay region. *Cryosphere* **16**, 3753–3773 (2022).

44. Constable, A. J. et al. Polar regions. In *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the IPCC* (eds Pörtner, H.-O. et al.) 2319–2367 (Cambridge Univ. Press, 2022).

## Supplementary Information

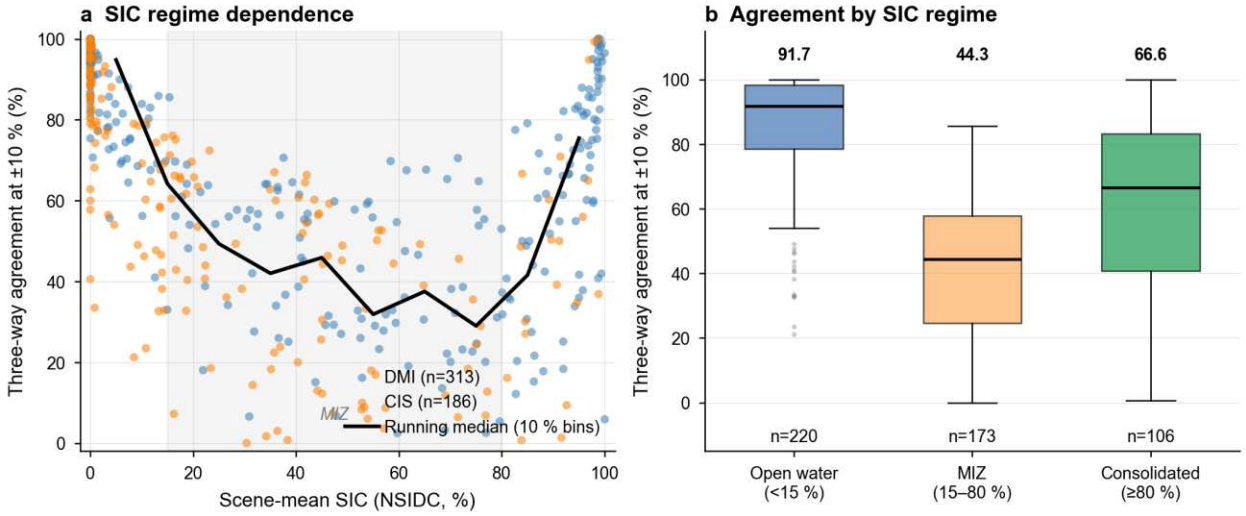
## Bridging ice charts and microwave records: probabilistic sea-ice concentration retrieval from SAR via multi-reference flow matching

## Supplementary Figures

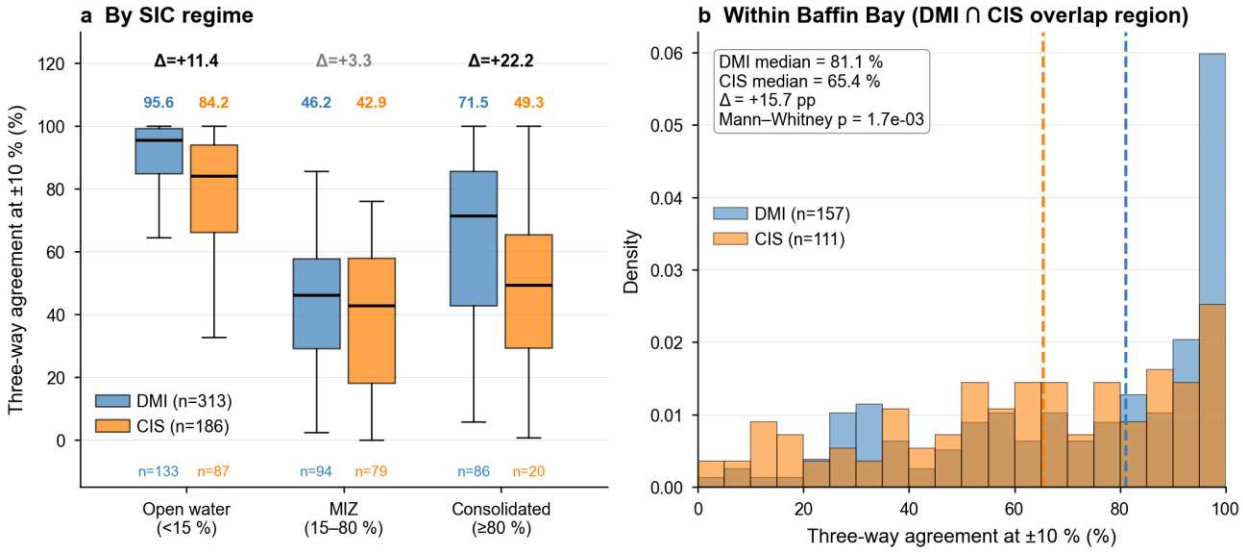



**Supplementary Fig. S1 | Geographic and temporal distribution of the 499 Sentinel-1 EW GRDM scenes used in this study (2018–2021).** (a) Scene-centre locations coloured by acquisition month (1–12), shown on a North Polar Stereographic projection (60–90 °N). Winter–spring acquisitions (December–April) dominate the dataset, reflecting the operational priority of ice-chart production during the navigation-sensitive season. (b) The same scene centres coloured by ice-service provider: Danish Meteorological Institute (DMI, n = 313, blue) and Canadian Ice Service (CIS, n = 186, orange). DMI scenes cluster around East Greenland and the Greenland

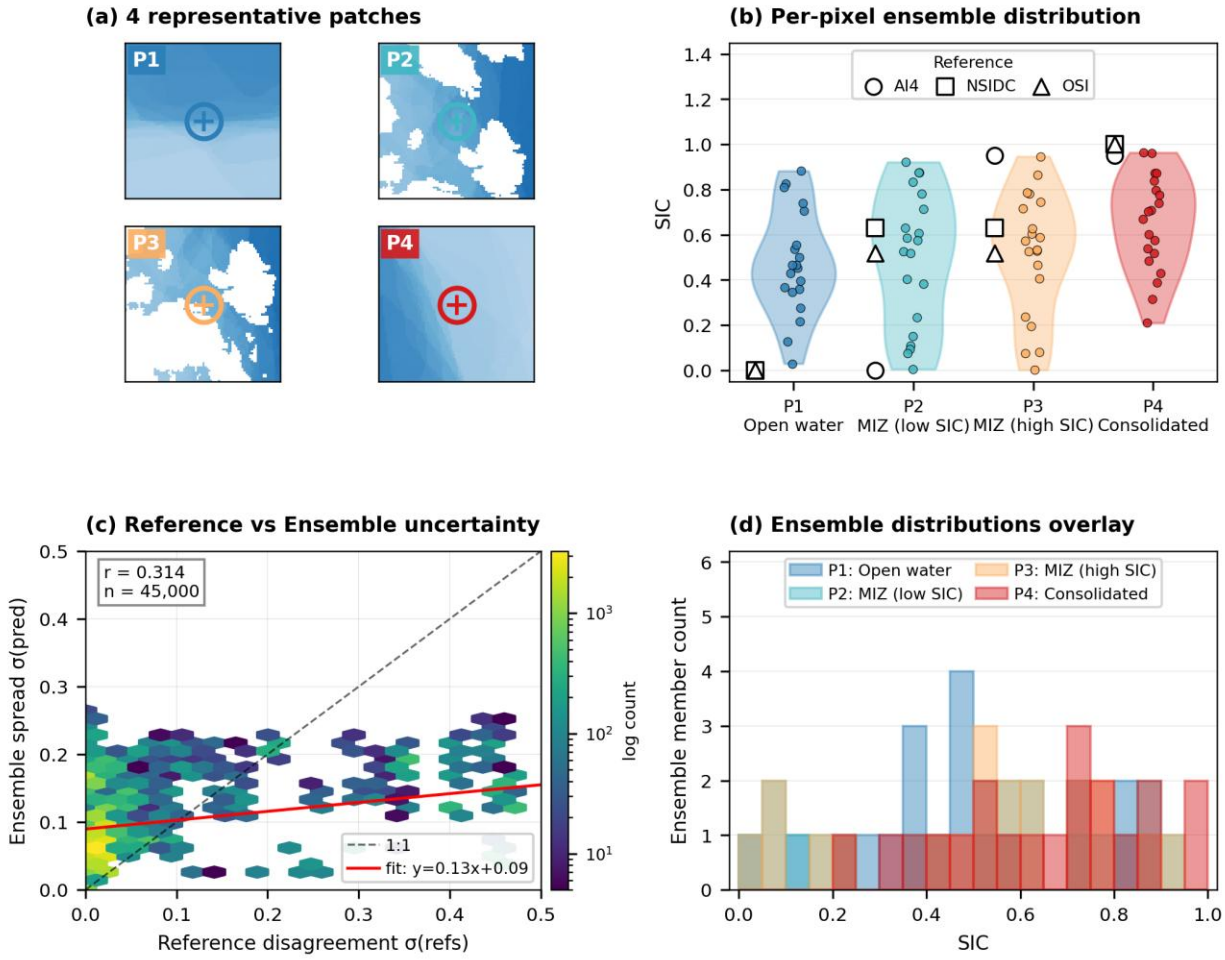
Sea, whereas CIS scenes concentrate over Baffin Bay and the Canadian Arctic Archipelago, reflecting each agency's operational mandate. (c) Number of scenes per sub-region: Baffin Bay (BB, n = 268), Greenland Sea (GS, n = 155), Canadian Arctic (CA, n = 71), Central Arctic / Other (CAO, n = 5), and Barents Sea (BS, n = 0). Baffin Bay dominates the dataset because it lies within the operational mandate of both ice services. Sub-regions were assigned from scene-centre longitude and latitude (Methods 4.1). The marker "+" with label "N" indicates the geographic North Pole; dashed circles denote parallels at 60°, 70°, and 80°N. Sub-region abbreviations: GS, Greenland Sea; BB, Baffin Bay; CA, Canadian Arctic; BS, Barents Sea; CAO, Central Arctic / Other. Source data: consistency_3way.csv (Supplementary Table S1).



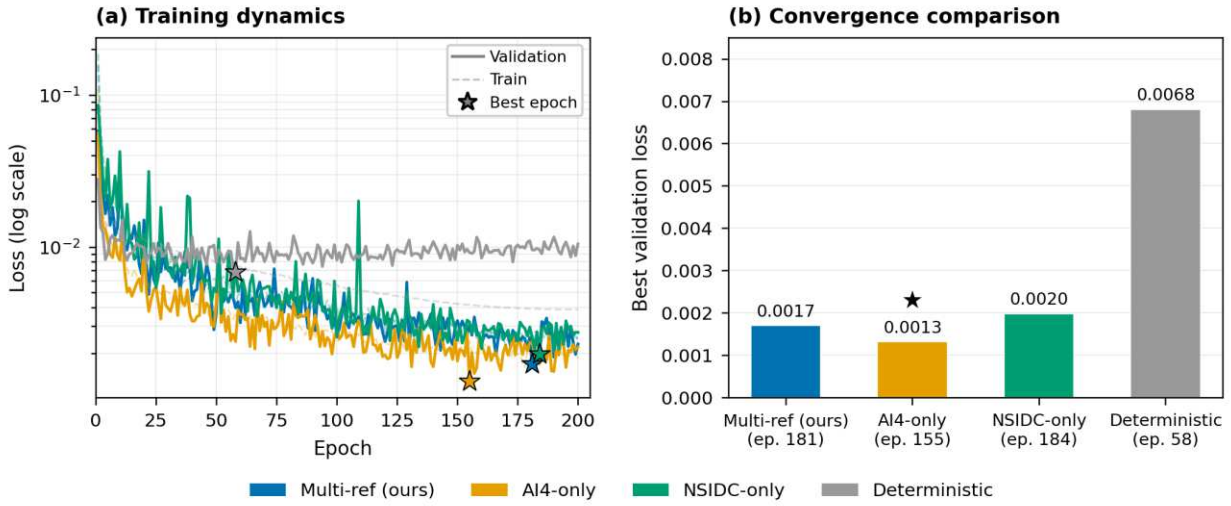

**Supplementary Fig. S2 | Dependence of three-way ground-truth agreement on the scene-mean sea-ice concentration regime (n = 499 Sentinel-1 scenes, 2018–2021).** (a) Scatter plot of three-way agreement at ±10 % SIC tolerance versus the scene-mean SIC reported by NSIDC G02202 v6. Each marker represents one scene, coloured by ice-service provider (DMI, blue, n = 313; CIS, orange, n = 186). The grey shaded band (15–80 % SIC) delineates the marginal-ice zone (MIZ). The solid black line shows the running median computed in 10 %-SIC bins (≥ 3 scenes per bin). Agreement is highest in open-water and consolidated-ice regimes and drops sharply within the MIZ, where the three reference products disagree most. The Spearman rank correlation between scene-mean SIC and three-way agreement is $\rho = -0.56$ ($p < 0.001$), reflecting a non-monotonic, U-shaped relationship that is captured by the running median rather than by a linear fit. (b) Box-plots of three-way agreement stratified by SIC regime: open water (< 15 % SIC, n = 220; median = 91.7 %, IQR [78.6, 98.2]), MIZ (15–80 %, n = 173; median = 44.3 %, IQR [24.6, 57.8]), and consolidated ice (≥ 80 %, n = 106; median = 66.6 %, IQR [40.7, 83.3]). Boxes show the interquartile range; whiskers extend to 1.5 × IQR; outliers are plotted as grey dots. A one-sided Mann–Whitney U test confirms that agreement in the MIZ is significantly lower than in the combined open-water and consolidated regimes ($U = 7{,}104$, $p < 0.001$).




**Supplementary Fig. S3 | Service-stratified three-way agreement: DMI vs Canadian Ice Service (CIS) across 499 Sentinel-1 scenes.** (a) Three-way agreement at ±10 % SIC tolerance, stratified jointly by SIC regime and by ice-service provider. In each regime, DMI and CIS distributions are shown as side-by-side box-plots (boxes: IQR; whiskers: 1.5 × IQR; outliers omitted; medians annotated above each box; Δ = DMI median − CIS median printed at top). The two services agree closely in the marginal-ice zone (Δ = +3.3 pp) but diverge in the open-water (Δ = +11.4 pp) and consolidated-ice (Δ = +22.2 pp) regimes, suggesting that the two operational centres apply different conventions for low- and full-concentration coding rather than for the ambiguous MIZ. Sample sizes per service per regime are annotated below each box. (b) Within Baffin Bay — the only sub-region with substantial coverage from both services (DMI n = 157, CIS n = 111) — agreement distributions remain offset (DMI median 81.1 %, CIS median 65.4 %, Δ = +15.7 pp; Mann–Whitney U test, $p = 1.7 \times 10^{-3}$), confirming that the service-level difference is not an artefact of differing geographic coverage. Dashed vertical lines mark the two medians. Source data: Supplementary Table S1 (consistency_3way.csv).




**Supplementary Fig. S4 | Per-pixel ensemble behaviour of the multi-reference flow-matching retrieval across four representative SIC regimes.** Four pixels were selected from the held-out test set, one per regime, with the highest combined score of spatial diversity and ensemble spread (Methods 4.4). (a) 80 × 80-pixel mini-patches centred on each target pixel (SIC colour scale 0–1): **P1** — open water (scene 20210331T205323, patch 22, pixel (205, 244)); **P2** — low-MIZ (scene 20210419T101641, patch 15, pixel (149, 103)); **P3** — high-MIZ (same scene/patch, pixel (124, 80)); **P4** — consolidated ice (scene 20210217T214246, patch 66, pixel (1, 83)). The target pixel is highlighted by a marker. (b) 20-member ensemble SIC distribution at each pixel (filled violin and individual jittered members), with the three reference values overlaid as open markers (○ AI4, □ NSIDC, △ OSI SAF). Ensemble spread is largest where the three references most disagree (P2 and P3, both inside the MIZ). (c) Hexbin density of ensemble standard deviation σ(pred) vs reference disagreement σ(refs) across n = 45,000 randomly sampled valid pixels from the 100 test scenes. The red solid line is a least-squares fit (y = 0.13 x + 0.09; Pearson r = 0.314); the dashed line is the 1:1 calibration ideal. The positive but sub-unity slope indicates that the model partially propagates inter-reference disagreement into per-pixel ensemble spread. (d) Overlaid histograms of the 20 ensemble members at each of the four pixels (P1–P4), illustrating regime-dependent distribution shapes. Source data: results/main/*.npz.




**Supplementary Fig. S5 | Training dynamics and convergence of the four flow-matching variants.** All four models share an identical 5-level U-Net backbone (93.3 M parameters), input conditioning, training split (322 scenes, 17,466 patches), and optimiser settings (AdamW, learning rate $1 \times 10^{-4}$, weight decay $1 \times 10^{-4}$, EMA decay 0.9999, batch size 4, bfloat16); they differ only in the supervision target (Methods 4.3). (a) Training (dashed) and validation (solid) loss vs epoch, log-scale, for the four models: Multi-reference (blue), AI4-only (orange), NSIDC-only (green), and the Deterministic U-Net baseline (grey). Stars mark the best validation epoch per model. All three flow-matching variants converge smoothly over 200 epochs with no sign of overfitting, whereas the deterministic baseline plateaus much earlier (best epoch 58) and at a substantially higher loss, consistent with its reduced cross-reference robustness reported in Fig. 5c and Table 2. (b) Best validation loss per model (bar) and corresponding best epoch (annotated under each bar). AI4-only achieves the lowest absolute validation loss (0.0013) because it optimises a single, internally consistent reference; this lowest value (★) is, however, not an indication of better retrieval skill — Fig. 5 and Table 2 show that AI4-only suffers the largest cross-reference MAE std (0.052) and the largest MIZ CRPS (0.165), in line with the label-dependent overfitting discussed in Section 2.3. Source data: logs/{model}/history.json for model ∈ {main, ai4, nsidc, det}.

## Supplementary Table

**Supplementary Table S1 | Per-scene three-way consistency dataset.** Associated file: consistency_3way.csv (provided as Supplementary Data 1). Per-scene pixel-level consistency statistics computed across three SIC references — AI4Arctic analyst polygons, NSIDC G02202 v6, and EUMETSAT OSI SAF v3.1 — over 512 Sentinel-1 EW GRDM scenes from the AI4Arctic Sea Ice Challenge Ready-To-Train v3 dataset (2018–2021). 499 scenes have status = "ok" and form the analysis subset; 13 scenes were excluded (12 with fewer than 100 jointly valid pixels; 1 with no temporally collocated OSI SAF product within ±1 day). Dimensions: 512 rows × 27 columns. **Column dictionary**

| Column | Type | Unit | Description |
|---|---|---|---|
| scene | str | — | Original AI4Arctic scene filename |
| date | str | YYYY-MM-DD | Acquisition date |

| | | | |
|---|---|---|---|
| month | int | 1–12 | Month of acquisition |
| year | int | — | Year of acquisition |
| service | str | — | Ice service that produced the analyst chart (dmi / cis) |
| status | str | — | Inclusion flag (ok, too_few_valid, no_osi) |
| N | int | pixels | Number of jointly valid pixels on the 23 × 21 sparse grid |
| ai4_mean | float | % SIC | Scene-mean SIC from AI4Arctic mid-bin mapping |
| nsidc_mean | float | % SIC | Scene-mean SIC from NSIDC G02202 v6 (×100) |
| osi_mean | float | % SIC | Scene-mean SIC from OSI SAF v3.1 |
| ai_v_nsidc_bias | float | % SIC | Mean(AI4 − NSIDC) |
| ai_v_nsidc_mae | float | % SIC | Mean absolute error AI4 vs NSIDC |
| ai_v_nsidc_rmse | float | % SIC | Root-mean-square error AI4 vs NSIDC |
| ai_v_nsidc_corr | float | — | Pearson correlation AI4 vs NSIDC |
| ai_v_nsidc_agr10 | float | % | Pixel agreement at ±10 % SIC, AI4 vs NSIDC |
| ai_v_osi_bias | float | % SIC | Mean(AI4 − OSI SAF) |
| ai_v_osi_mae | float | % SIC | Mean absolute error AI4 vs OSI SAF |
| ai_v_osi_rmse | float | % SIC | Root-mean-square error AI4 vs OSI SAF |
| ai_v_osi_corr | float | — | Pearson correlation AI4 vs OSI SAF |
| ai_v_osi_agr10 | float | % | Pixel agreement at ±10 %, AI4 vs OSI SAF |
| nsidc_v_osi_bias | float | % SIC | Mean(NSIDC − OSI SAF) |
| nsidc_v_osi_mae | float | % SIC | Mean absolute error NSIDC vs OSI SAF |
| nsidc_v_osi_rmse | float | % SIC | Root-mean-square error NSIDC vs OSI SAF |
| nsidc_v_osi_corr | float | — | Pearson correlation NSIDC vs OSI SAF |
| nsidc_v_osi_agr10 | float | % | Pixel agreement at ±10 %, NSIDC vs OSI SAF |
| all3_agr10 | float | % | Three-way agreement at ±10 % (all three pair-wise |
| nsidc_osi_only_agr10 | float | % | Two-way (NSIDC ∩ OSI) agreement at ±10 % |

*Grid and reprojection.* All three references were sampled onto the 23 × 21 AI4Arctic sparse latitude/longitude grid. NSIDC was sampled by nearest-neighbour reprojection from

EPSG:3411. OSI SAF was sampled by KD-tree nearest-neighbour search on 2-D lat/lon coordinates (EASE2). *SIC unit harmonisation.* AI4Arctic SIC classes 0–10 were mapped to WMO mid-bin percentages (0, 5, 15, 25, …, 95 %). NSIDC cdr_seaice_conc (0–1 fraction) was multiplied by 100. OSI SAF ice_conc (0–100 %) was used as-is. *License:* CC BY 4.0.

## Supplementary Notes

### Supplementary Note 1 | Reprojection and grid harmonisation details

The three SIC reference sources are originally defined on different projections and resolutions: AI4Arctic polygons on the native Sentinel-1 SAR grid (40 m, EPSG-equivalent UTM zone of the scene); NSIDC G02202 v6 on the NSIDC Polar Stereographic North grid (EPSG:3411, 25 km, 448 × 304 pixels); OSI SAF v3.1 on the EASE2 grid (EPSG:6931, 25 km, 432 × 432 pixels). To enable per-pixel comparison we adopted the sparse 23 × 21 latitude/longitude grid (2-km spacing) provided in each AI4Arctic scene as the common reference grid.

*AI4Arctic polygons → sparse grid.* WMO Egg-Code class values (0, 1, 3, 5, 6, 9, 10) were mapped to mid-percent SIC (0, 5, 25, 50, 65, 90, 95 %) and block-mean-aggregated from the dense (5,353 × 5,180) 40-m grid to the 23 × 21 (2-km) sparse grid using xarray.coarsen.

*NSIDC G02202 v6 → sparse grid.* Each NSIDC daily file was opened with xarray.open_dataset; the 1-D x/y polar-stereographic coordinates (m) were converted to (lat, lon) using pyproj.Transformer.from_crs(3411, 4326). The 2-D (lat, lon) grid was then used as source points for scipy.interpolate.griddata(method='nearest') to interpolate the SIC field onto the AI4Arctic sparse grid. SIC values were converted to percent (0–100) and the fill value (251–255) was masked to NaN.

*OSI SAF v3.1 → sparse grid.* Each OSI SAF daily file already contains 2-D lat/lon coordinate variables alongside the ice_conc field. We used scipy.interpolate.griddata(method='nearest') directly with the file's embedded (lat, lon) as source points. The ice_conc field is provided in percent units (0–100 %); the status_flag was used to mask invalid pixels (open water without retrieval, land, missing).

*Validation.* For a randomly selected scene (2018-01-08, East Greenland Sea), reprojected NSIDC and OSI SAF SIC values were compared against an independent re-implementation using cartopy-based projection; pixel-wise differences were < 0.1 % SIC, confirming that the reprojection is numerically reproducible.

### Supplementary Note 2 | Flow-matching training and inference hyperparameters

*Loss.* Standard conditional flow-matching loss (Lipman et al. 2023[39]; Tong et al. 2024[40]):

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{\tau\sim U(0,1),\, x_0\sim\mathcal{N}(0,I),\, x_1\sim p_{\mathrm{data}}} \| v_\theta(x_\tau, \tau, c) - (x_1 - x_0) \|_2^2$$

with linear interpolant $x_\tau = (1 - \tau)\, x_0 + \tau\, x_1$. The target $x_1$ is a stochastic draw from {AI4, NSIDC, OSI} (uniform when all three are available; otherwise uniform over the available subset).

*Architecture.* 5-level U-Net backbone (4 encoder + bottleneck + 4 decoder blocks with skip connections); channel widths [64, 128, 256, 512, 1024]; group normalisation (groups = 8); SiLU activations; FiLM[41] conditioning at every residual block injecting the flow-time embedding τ (sinusoidal). The conditioning tensor c stacks SAR (HH, HV; 2 channels), AMSR2 brightness temperatures (14 channels at 2-km native, upsampled), ERA5 atmospheric fields (6 channels: 10-m winds, 2-m temperature, skin temperature, total column water vapour, total column cloud

liquid water), and normalised latitude/longitude (2 channels) — 24 channels in total — concatenated as side channels at the SAR patch resolution. Total parameter count: 93.3 M.

*Patch sampling.* 256 × 256 SAR pixel patches sampled with replacement, weighted by valid-mask fraction. Random rotations (multiples of 90°) and horizontal flips applied as augmentation.

*Optimiser.* AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $1 \times 10^{-4}$), learning rate $1 \times 10^{-4}$, batch size 4, mixed-precision (bfloat16), exponential moving average decay 0.9999.

*Training data.* 322-scene training split (17,466 patches), 77-scene validation split, 100-scene held-out test split (stratified by month and ice service). Training ran for 200 epochs (~46 h per model on a single NVIDIA A100 80 GB).

*Inference.* 20-step Euler ODE solver from $\tau = 0$ to $\tau = 1$ with uniform step size $\Delta\tau = 0.05$. For each input scene we generate M = 20 independent ensemble members by drawing 20 independent Gaussian priors and integrating in parallel.

**Supplementary Note 3 | Definition of evaluation metrics**

*Continuous Ranked Probability Score (CRPS).* For an ensemble forecast F with M = 20 members $\{f_1, \ldots, f_M\}$ and observation o, the empirical CRPS is

$$\mathrm{CRPS}(F, o) = \frac{1}{M}\sum_i |f_i - o| - \frac{1}{2M^2}\sum_{i,j} |f_i - f_j|$$

We report the per-pixel CRPS averaged over all valid pixels in the test set.

*Mean Absolute Error (MAE).* For the deterministic baseline and the ensemble mean: MAE = $(1/N)\,\Sigma\,|\hat{y} - o|$.

*Integrated Ice-Edge Error (IIEE).* Goessling et al. (2016)[6] definition: IIEE = A_over + A_under, where A_over (A_under) is the area where the retrieval reports ice (no ice) but the reference reports no ice (ice), thresholded at SIC = 15 %.

*Reference-agnostic robustness.* For each model, compute MAE against AI4, NSIDC, and OSI SAF separately; report the mean and standard deviation across these three reference-specific MAEs. A lower std indicates better cross-reference consistency.

*Spread–skill diagram.* Pixels are binned by ensemble standard deviation σ into equal-population bins; for each bin we plot the RMSE of the ensemble mean. A perfectly calibrated ensemble lies on the σ = RMSE line.

*Probability Integral Transform (PIT) histogram.* For each pixel we compute $u = \hat{F}(o)$, where $\hat{F}$ is the empirical CDF of the 20-member ensemble at that pixel. A calibrated ensemble produces u uniformly distributed on [0, 1]. We compute the histogram across all valid pixels and the Kolmogorov–Smirnov statistic against U(0, 1).

*Reliability diagram.* For binarised retrievals at the ice-edge threshold (SIC = 15 %), pixels are binned by predicted probability $\hat{p}$ into 10 equal-width bins and the observed frequency is plotted against $\hat{p}$; calibrated retrievals lie on the 1:1 line. We report the mean absolute deviation (MAD) from the 1:1 line.