

Methods note · Sea ice and ocean

Sea-ice forecast skill starts with a baseline

Eirik Solheim · 15 October 2026

Fictional author and institute · Editorial demonstration, not peer reviewed

ABSTRACT

Forecast accuracy does not establish added information unless it is compared with an explicit alternative. This methods note draws on primary ice-edge verification and subseasonal prediction studies to define a fair comparison among a model, climatology, and persistence-based forecasts. A hypothetical Brier-score calculation shows how a model can be skilful against one baseline and unskilful against another. A spatial example explains why correct total extent can coexist with a misplaced ice edge. The note examines issue-time information boundaries, changing climatology, calibration, dependence, and observation uncertainty without evaluating an operational system. Its practical conclusion is a conditional skill statement naming the target, lead, dates, region, reference observations, and baseline, rather than treating a high percentage of correct outcomes as sufficient evidence of usefulness.

Keywords: sea-ice forecasts · forecast skill · baseline · Brier score · ice-edge error

A source-based editorial text, not a report of original Still North research. Examples and imagery are illustrative.

A sea-ice forecast can look accurate because much of the ocean remains in its usual state. Predicting no ice in a region that is almost always ice-free may succeed frequently without offering much new information. Skill asks a different question from raw accuracy: does the forecast improve on an explicit alternative that could have been issued with the same information at the same time?

This methods note develops that comparison for forecasts of ice presence and ice-edge location. It draws on primary studies of integrated ice-edge error, probabilistic contour verification, and operational subseasonal prediction. Its worked scores are hypothetical and do not evaluate a real forecast system. The purpose is to make the target, baseline, score, and verification sample visible before a forecast is described as skilful.

Fix the forecast event

Define ice presence using a stated concentration threshold, grid, observation product, and valid time. A common ice-edge definition uses the 15% concentration contour, but a threshold is a convention for the forecast target, not a safe-navigation boundary. Daily presence, monthly mean concentration, regional extent, and the date of retreat are different outcomes. A forecast verified against one cannot inherit a claim of skill for the others.

Lead time and issue date also matter. A forecast issued ten days before a daily target is not comparable with one issued a month before a monthly target merely because both mention the same calendar date. Specify what observations were available at issue time, the ensemble size, post-processing, and the region over which the score is calculated. These details determine the task being tested rather than decorating an otherwise complete result.

Choose a baseline that can actually be issued

Climatology predicts the event frequency for the relevant date and region, estimated from a declared training

record. Persistence carries forward the latest observed state. A damped-anomaly persistence forecast allows an observed anomaly to weaken towards an estimated seasonal background. They exploit different information and may perform differently with season and lead. A serious evaluation often needs more than one, particularly where a weak climatological baseline makes a complex forecast look better than it is.

None of the baselines should use the verification outcome during fitting. In a changing climate, full-record climatology can both leak information from later years and misrepresent the present background state. Use a defensible fixed training period or a documented rolling procedure that only uses information available before each forecast. Any trend adjustment must obey the same rule. The objective is an honest competitor, not the easiest denominator for a flattering skill score.

A worked probabilistic score

For a binary event, the Brier score is the mean of $(p - o)^2$, where p is the forecast probability of ice presence and o is one when ice occurs and zero otherwise. Lower scores are better. Consider four hypothetical verification cases whose observed outcomes are 1, 0, 1, and 0. A model issues probabilities 0.8, 0.4, 0.6, and 0.2. A baseline trained before these cases issues 0.5 every time.

The model's mean score is 0.10 and the baseline's is 0.25. Defining skill as $1 - \text{model score}/\text{baseline score}$ gives 0.60, or a 60% reduction in mean squared probability error against this particular baseline on these four invented cases. It does not mean 60% of forecasts were correct, probabilities are calibrated, or the model improves every case. Four teaching cases are far too few to establish performance.

Change the comparison and the conclusion can change. If an alternative issued baseline scores 0.08 on the same cases, the model's skill against it is $1 - 0.10/0.08 = -0.25$. The model has not changed; the comparator has. If the

Case	Observed event	Model probability	Model squared error	Baseline squared error
1	1	0.8	0.04	0.25
2	0	0.4	0.16	0.25
3	1	0.6	0.16	0.25
4	0	0.2	0.04	0.25

baseline score is zero, this ratio is undefined. Report the raw scores and baseline definition alongside the skill value so a reader can distinguish genuine improvement from a convenient reference.

Regional extent can conceal a misplaced edge

Imagine a hypothetical forecast that predicts the correct total ice-covered area but places part of it in the wrong sector. Its extent error could be zero while its local presence forecast is wrong across substantial areas. Goessling and colleagues introduced integrated ice-edge error, defined as the area where forecast and reference disagree about presence above the chosen concentration threshold. Its decomposition separates absolute extent error from misplacement.

For an invented map, suppose the forecast incorrectly includes 100 km² of ice in one sector and misses 100 km² elsewhere. The net extent error is zero, but the disagreement area is 200 km². A user concerned with the edge cannot regard those misplaced areas as cancelling benefits. Area weighting is necessary when grid cells differ in size; simply counting latitude-longitude pixels can distort a spatial score.

Goessling and Jung's probabilistic contour score extends this logic by integrating local probability scores over space. The method responds to ensemble properties and allows comparison with persistence and climatological forecasts. The local Brier example above teaches the underlying probability-error idea; it is not an implementation of every normalisation and convention in their Spatial Probability Score. Reproducing a published metric requires following its exact definition.

Verify the deployment, not the training archive

A hindcast evaluation should reproduce the issue-time information boundary. Fit calibration and bias correction without using the held-out verification year. Keep tuning separate from final evaluation, and compare identical dates, regions, and valid observations across systems. Otherwise a model tested only on easy conditions may appear superior to a baseline evaluated everywhere. Missing forecasts should be reported rather than silently excluded from a service-level assessment.

Uncertainty must respect dependence. Adjacent grid cells and successive daily forecasts share information; treating each as independent can produce an implausibly narrow confidence interval. Use paired comparisons over meaningful blocks, such as years or appropriately defined forecast episodes, and explain the sample's limitations. A basin-wide mean should be accompanied by performance across season, lead, and decision-relevant regions.

An average improvement can coexist with a systematic weakness in a particular coastal sector.

Probability quality and user value are additional questions

A skilful aggregate score does not guarantee calibrated probabilities. When a model repeatedly forecasts 70%, the event frequency in comparable cases should be examined, with uncertainty and enough examples to make that comparison meaningful. Ensemble spread, bias, and resolution can contribute differently to performance. Nor does a good average score guarantee operational usefulness for a specific action whose costs depend disproportionately on missed ice.

Zampieri and colleagues found substantial differences among operational systems in subseasonal ice-edge prediction. Their result supports evaluating the actual system and target, not transferring a general claim that 'sea ice is predictable' to every product. Potential predictability in an idealised model experiment is also distinct from realised skill in an operational forecast with imperfect initialisation, model bias, and observational uncertainty.

Limits and the useful headline

The examples do not establish a sample size, confidence interval, or navigation recommendation. Reference concentration has uncertainty, especially near the edge and during melt. Verifying against several justified observation products can reveal sensitivity, but cannot create independent truth when those products share inputs. A well-scoring large-scale presence forecast does not resolve ridges, thin ice, floe motion, or other conditions required for local safety decisions.

A meaningful headline therefore names the comparison: a specified forecast reduced a specified error relative to a stated baseline, over named dates, leads, regions, and observations. That sentence is longer than 'high accuracy'. It is also the shortest claim that tells a reader what useful information the forecasting system actually added.

References

- Goessling, H. F., Tietsche, S., Day, J. J., Hawkins, E., and Jung, T. (2016). Predictability of the Arctic sea ice edge. *Geophysical Research Letters*, 43, 1642–1650.
- Goessling, H. F., and Jung, T. (2018). A probabilistic verification score for contours: methodology and application to Arctic ice-edge forecasts. *Quarterly Journal of the Royal Meteorological Society*, 144, 735–743.
- Zampieri, L., Goessling, H. F., and Jung, T. (2018). Bright prospects for Arctic sea ice prediction on sub-seasonal time scales. *Geophysical Research Letters*, 45, 9731–9739.