NEURIPS 2026

Njord

A Probabilistic Graph Neural Network for Ensemble Ocean Forecasting

Daniel Holmberg[1] Joel Oskarsson[2] Erik Wikingsson[3]
Fredrik Lindsten[3] Teemu Roos[1]

1 UNIVERSITY OF HELSINKI
2 ETH AI CENTER
3 LINKÖPINGS UNIVERSITET



The Norse sea god Njord, *Edda Oblongata*, c. 1680

01
Motivation

MOTIVATION

Ocean forecasting needs uncertainty

- Ocean forecasts drive navigation, fisheries, coastal hazards and environmental monitoring.
- Numerical models (NEMO) take *hours on CPU clusters* at operational resolution.
- ML emulators (GLONET, WenHai, XiHe, ...) match or even outperform physics-based systems in some cases at a fraction of the cost, but are all **deterministic**:
  - trained with MSE $\Rightarrow$ predict the conditional mean
  - smooth out fine scales
  - lack estimate of forecast confidence

> Ocean dynamics are **chaotic**. Decisions benefit from the *full distribution* of future states, as opposed to a single trajectory.

In weather, conditional generative models (GenCast, Graph-EFM, AIFS-CRPS) already give skilful ensembles.

No comparable probabilistic data-driven model exists for the ocean.

MOTIVATION

Njord in a nutshell

Njord is the first **probabilistic** data-driven model for short-range, high-resolution **global and regional** ocean forecasting.

1. **Generative ensembles**, with every member and step sampled in a *single forward pass* (3 s globally on one MI250X).
2. **K-means cluster meshes** that adapt to irregular coastlines.
3. **Sea ice** with physical constraints via a density channel.
4. **Global Njord at 0.25°** has the lowest error on average across upper-ocean variables *against real-world observations*.
5. **Njord-Baltic at 2 km** matches SeaCast while also giving ensembles.

02
Method

METHOD

Problem formulation

- Ocean states $X^{1:T}$, with $X^t \in \mathbb{R}^{N \times d_x}$:
  d_x variables at N ocean grid points
  (T, S, U, V at several depths; SSH, sea ice)
- Initial conditions $X^{-1:0}$
- Forcing $F^{-1:T}$: atmosphere from a weather model, static fields, time of year

GOAL: MODEL THE FULL DISTRIBUTION

$$p\left(X^{1:T} \mid X^{-1:0}, F^{-1:T}\right)$$

ENSEMBLE MEMBER

Each sample $X^{1:T} \sim p(X^{1:T} \mid X^{-1:0}, F^{-1:T})$ is one possible trajectory of the ocean.

Markov assumption: sample X^t from $p(X^t \mid X^{t-2:t-1}, F^{t-2:t})$, then condition on it for the next step *autoregressively*.

A deep latent variable model

Following Graph-EFM [1], previously used for weather forecasting.

$$\hat{p}_\theta\left(X^{1:T} \mid X^{-1:0}, F^{-1:T}\right) = \prod_{t=1}^{T} \int \underbrace{p_\theta\left(X^t \mid Z^t, X^{t-2:t-1}, F^{t-2:t}\right)}_{\text{predictor}} \underbrace{p_\theta\left(Z^t \mid X^{t-2:t-1}, F^{t-2:t}\right)}_{\text{latent map}} \mathrm{d}Z^t$$

Latent map (the prior)

- Latent random variable $Z^t \in \mathbb{R}^{|\mathcal{V}_L| \times d_z}$: one row Z^t_α per node α in the top graph level $\mathcal{V}_L$
- *Mean* from a GNN g_θ, *fixed unit variance*:

$$p_\theta\left(Z^t \mid X^{t-2:t-1}, F^{t-2:t}\right) = \prod_{\alpha \in \mathcal{V}_L} \mathcal{N}\left(Z^t_\alpha \,\middle|\, g_\theta\left(X^{t-2:t-1}, F^{t-2:t}\right)_\alpha, I\right)$$

Predictor

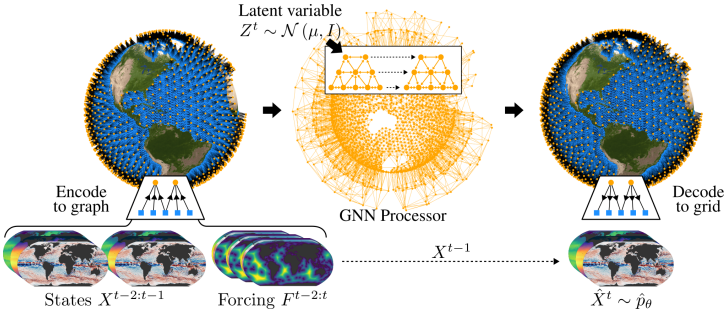
- Hierarchical GNN with a skip connection:

$$\hat{X}^t = X^{t-1} + \hat{r}^t\left(Z^t, X^{t-2:t-1}, F^{t-2:t}\right)$$

- Z^t enters at the coarsest level and can influence *every* output
- One sample = *one forward pass*, no iterative denoising

METHOD

The Njord architecture



One-step prediction: the predicted residual is added to X^{t-1} to give the sample $\hat{X}^t$.

- **Encode** $X^{t-2:t-1}, F^{t-2:t}$ onto the ocean-only graph
- **Process** up $\mathcal{G}_1 \rightarrow \cdots \rightarrow \mathcal{G}_L$; add $Z^t \in \mathbb{R}^{|\mathcal{V}_L| \times d_z}$ at the top level
- **Decode** to the grid: $\hat{X}^t = X^{t-1} + \hat{r}^t$
- Independent $Z_1^t, \ldots, Z_M^t \Rightarrow M$ members in parallel; roll out over $t = 1, \ldots, T$

METHOD

Training: ELBO + almost-fair CRPS

$$\mathcal{L} = -\mathbb{E}_{q_\phi}\left[\log p_\theta\left(X^t \mid Z^t, X^{t-2:t-1}, F^{t-2:t}\right)\right] + \lambda_{\mathrm{KL}}\, D_{\mathrm{KL}}\left(q_\phi\left(Z^t \mid X^{t-2:t}, F^{t-2:t}\right) \,\middle\|\, p_\theta\left(Z^t \mid X^{t-2:t-1}, F^{t-2:t}\right)\right) + \lambda_{\mathrm{CRPS}}\, \mathcal{L}_{\mathrm{CRPS}}$$

- Encoder q_ϕ, a separate GNN, sees the target X^t and learns *mean and std*; training only. The KL pulls the prior p_θ towards it.
- Likelihood $p_\theta(X^t \mid Z^t, \ldots) = \mathcal{N}(\hat{X}^t, \sigma^2)$ with a fixed σ per variable, so term 1 is a weighted MSE.

Almost-fair CRPS [2] on $M = 2$ samples from the prior:

$$\mathcal{L}_{\mathrm{CRPS}} = \frac{1}{M}\sum_m |\hat{x}_m - x| - \frac{1-\varepsilon}{2M(M-1)}\sum_{m,m'} |\hat{x}_m - \hat{x}_{m'}|$$

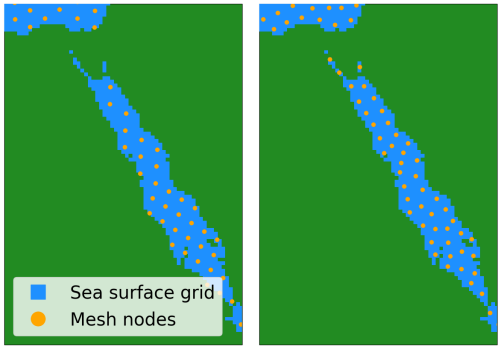
$\varepsilon = (1-\alpha)/M, \quad \alpha = 0.95$

CURRICULUM

Autoencoder	+ KL term	Rollout	+ CRPS	Finetune on analysis
T=1, λ_{KL}=0	λ_{KL}=0.1	T=2	$\lambda_{\mathrm{CRPS}} > 0$	T up to 7

METHOD

A graph adapted to ocean geometry



Icosahedral | **K-means cluster**

Mesh nodes in the northern Red Sea and Suez Canal.

Icosahedral meshes (GraphCast, Graph-EFM) are made for the atmosphere:

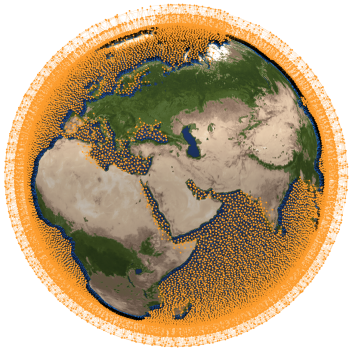
- only discrete sizes, e.g. 6 splits → 28 753 ocean nodes and 7 splits → 115 016 (out of memory)
- masking land breaks connectivity in bays and straits

Cluster meshes place nodes *where the ocean is*:

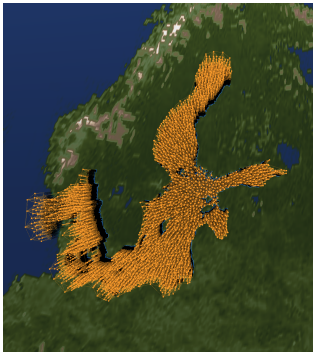
- spherical K-means on area-weighted ocean grid points
- spherical Delaunay edges, land-crossing edges removed
- repeat with fewer clusters to build the hierarchy

METHOD

Cluster meshes, global and regional



Global hierarchy



Baltic Sea hierarchy

- **Any number of nodes**, since the refinement factor is continuous
- Global grid of $N = 676\,736$ points $\rightarrow$ 33 777 nodes ($r_0 = 20$), then $r = 4$ per level
- Euclidean K-means for the Baltic, $r_0 = r = 9$
- Static *distance-to-coast* feature tells the model where the ocean ends

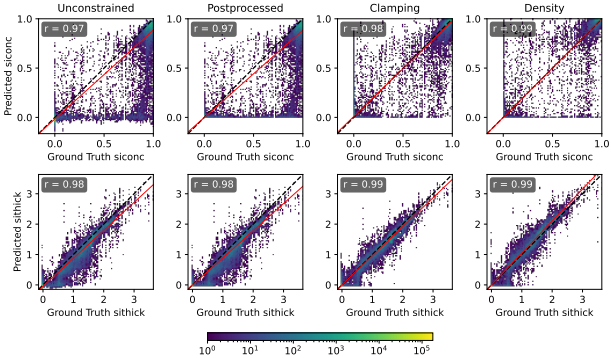
METHOD

Sea ice with physical constraints

- SIC $\in [0, 1]$, SIT ≥ 0, with many values *exactly* at the bounds.
- Rollouts should stay physical for *every sample*.
- Predict a binary **density channel** jointly with the state:

$$d = \mathbb{I}[\text{SIC} > 0]$$

- $\sigma(\hat{d}) < 0.5 \Rightarrow$ all ice variables set to 0; otherwise kept (plus soft clamping)
- Clean ice-free inputs, so no small residual errors accumulate



Predicted vs. true SIC and SIT at 30-day lead (1° ablation). The density channel (right) gives the most physical fields.

METHOD

Scaling to 0.25° global grids

- 676 736 ocean grid points, 23× the Graph-EFM global grid
- Grid↔mesh graphs with $\mathcal{O}(10^6)$ edges dominate memory
- **Flexible GNN layers** with separate widths for grid embeddings, edges and mesh nodes, so edges stay small and the processor can be wide
- Gradient checkpointing between rollout steps

TWO-STAGE RESOLUTION SCHEDULE

1° pretrain → **0.25°** finetune

Same mesh at both resolutions; only the encode/decode edges are swapped.

Sampling takes *3 s* per member and step globally and *1 s* for the Baltic on one AMD MI250X.

03
Experiments

Experimental setup

Model	Domain	Res.	Depths	Training data	Forcing (eval)	Ens.
Njord	Global	0.25°	6 (0–541 m)	GLORYS12 + GLO12 analysis	IFS forecasts	20
Njord-Baltic	Baltic	2 km	5 (1–91 m)	Baltic reanalysis + analysis	IFS + GLO12 boundary	5

EVALUATION

- 52 weekly OceanBench [3] forecasts through 2024, 10-day horizon
- Reanalysis 1993–2021 for training, finetuned on 2023 analysis

BASELINES

- GLO12 (physics), GLONET, WenHai, XiHe and SeaCast [4] globally
- SeaCast and GLO12 for the Baltic, plus persistence for both

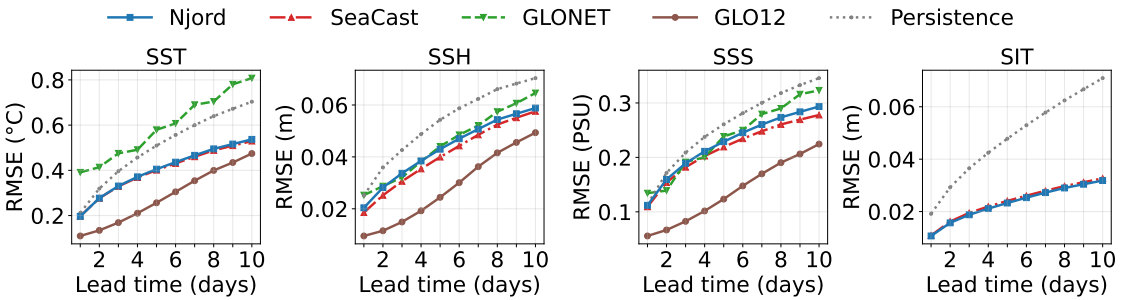
METRICS

RMSE of the ensemble mean $\bar{x} = \frac{1}{M} \sum_m \hat{x}_m$

CRPS how well the whole ensemble distribution matches the data

SSR $\sqrt{\frac{M+1}{M}} \frac{\text{Spread}}{\text{RMSE}}$, calibrated if ≈ 1

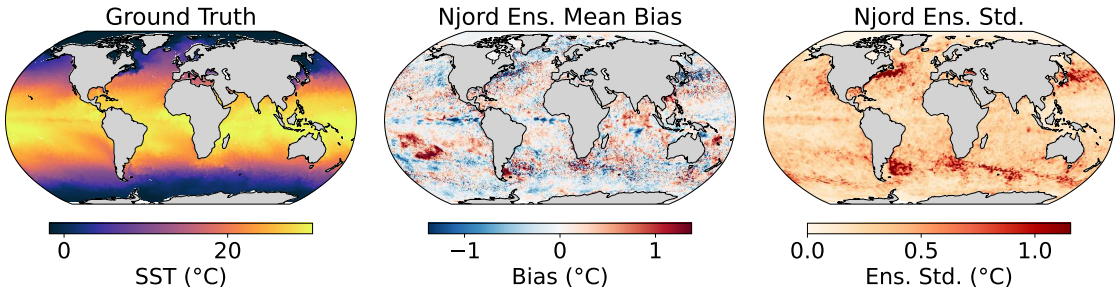
Global skill against GLO12 analysis



RMSE of the ensemble mean per lead time for SST, SSH, SSS and sea ice thickness.

- Competitive with all ML baselines, with *stable error growth* over 10 days.
- GLO12 is a hard reference here, since it produces the analysis targets itself. Observations give a more independent test.

An ensemble forecast of sea surface temperature

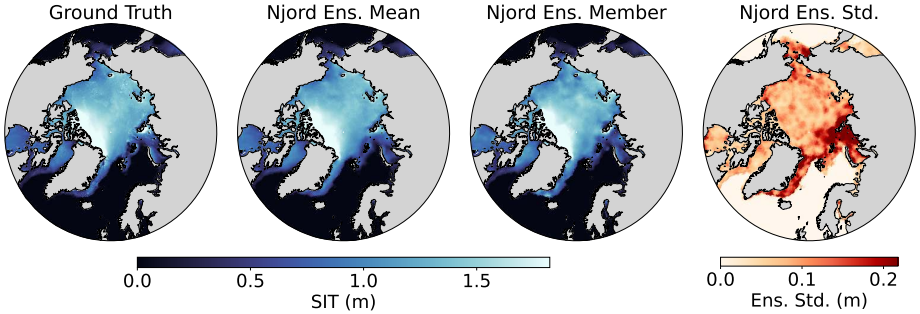


SST at 10-day lead, initialized 2024-01-30. Ground truth is GLO12 analysis.

- Ensemble mean keeps the sharp fronts of the major current systems.
- Spread peaks in dynamically active regions such as the *Gulf Stream*, Brazil–Malvinas Confluence, *Agulhas* Retroflection, *Kuroshio* Extension.

EXPERIMENTS

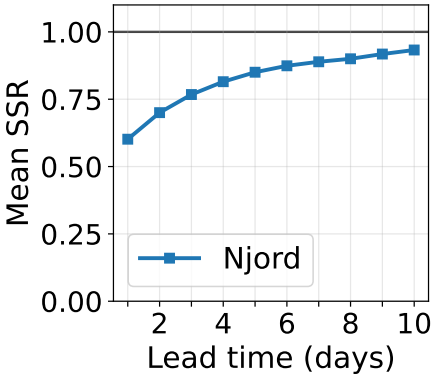
Arctic sea ice



Sea ice thickness at 10-day lead, initialized 2024-01-30. Ground truth is GLO12 analysis.

- Reproduces large-scale thickness gradients and the marginal ice zone.
- Spread peaks in the *Kara Sea* and other dynamic marginal zones.

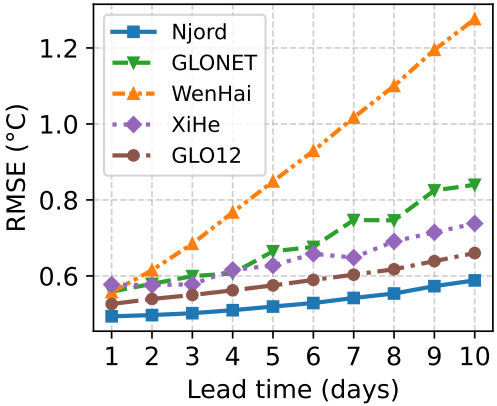
Calibration



SSR averaged over all global ocean variables.

- Spread-skill ratio rises steadily *towards the ideal value of 1*.
- Slightly underdispersed at short lead times.
- Consistent across ocean variables (per-variable CRPS and SSR in the paper).
- So far the only source of uncertainty is Z^t. Perturbing initial conditions and atmospheric forcing is a natural next step.

Against real-world observations: satellite SST



Global SST RMSE against bias-adjusted multi-sensor satellite SST (0.1°).

Njord has the **lowest SST error at every lead time**: below 0.6 °C at day 10, ahead of GLO12 and other ML emulators.

Njord is never trained on observations, but its training targets assimilate them. Likely contributors:

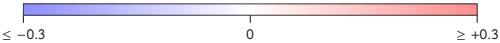
- comprehensive atmospheric forcing
- an expressive GNN backbone
- ensemble averaging

EXPERIMENTS

Against real-world observations: OceanBench in-situ

Floats, drifting buoys and along-track altimetry (IV-TT CLASS-4). Colours show the normalized RMSE difference with respect to GLO12: blue is better.

	Temperature [0-5m] RMSE (°C)					Salinity [0-5m] RMSE (psu)					Zonal current [15m] RMSE (m/s)				
Lead time (days)	1	3	5	7	10	1	3	5	7	10	1	3	5	7	10
GLO12	0.491	0.570	0.865	0.881	0.961	0.251	0.333	0.305	0.275	0.320	0.206	0.219	0.221	0.229	0.236
GLONET	0.592	0.665	0.971	1.028	1.154	0.234	0.361	0.277	0.259	0.313	0.194	0.208	0.210	0.216	0.219
WENHAI	0.526	0.738	1.104	1.260	1.564	0.266	0.346	0.308	0.299	0.340	0.215	0.228	0.224	0.231	0.234
XIHE	0.548	0.605	0.921	0.894	1.016	0.243	0.358	0.297	0.275	0.311	0.210	0.220	0.218	0.220	0.221
NJORD	0.470	0.519	0.844	0.836	0.912	0.245	0.340	0.268	0.261	0.312	0.202	0.211	0.211	0.216	0.220

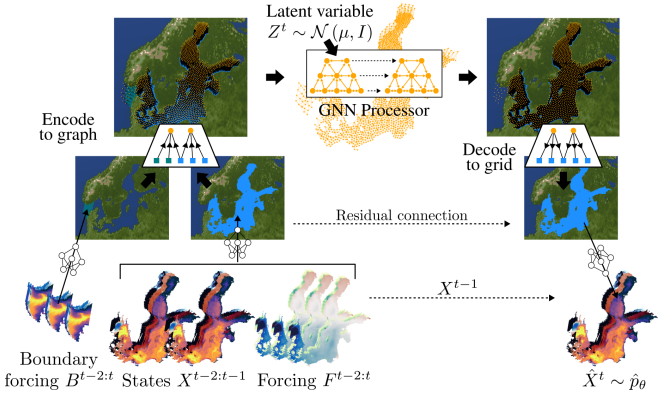


Normalized RMSE Difference w.r.t. GLO12

Lowest 0–5 m temperature RMSE at all lead times, and typically on par with the best model for the other observed variables above 100 m.

04
Regional model: Njord-Baltic

Njord-Baltic: the same framework at 2 km



- Trained on the 2 km NEMO Baltic Sea Physics Reanalysis with 147 701 ocean points, 5 depths (1–91 m), sea ice
- Same latent variable GNN on a regional cluster graph (16 358 → 1 767 → 193 nodes)
- Adds **lateral boundary forcing** $B^{t-2:t}$ from a 1/12° global model (teal, left)

Lateral boundary forcing

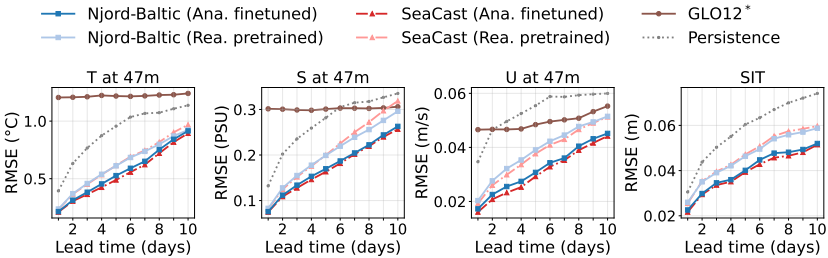
A coarse global model drives the fine regional model: $1/12^\circ$ GLO12 $\rightarrow$ 2 km Njord-Baltic.

$$\hat{X}^t \sim \hat{p}_\theta\left(X^t \mid X^{t-2:t-1}, F^{t-2:t}, B^{t-2:t}\right)$$
$$h_v = \mathrm{MLP}_X\left(X_v^{t-2:t-1}, F_v^{t-2:t}\right), \qquad v \in \mathcal{V}_{\text{grid}}$$
$$h_b = \mathrm{MLP}_B\left(B_b^{t-2:t}\right), \qquad b \in \mathcal{V}_B$$
$$H^{\mathcal{G}_1} \leftarrow \mathrm{GNN}_{\text{G2M}}\left(h_{\mathcal{V}_{\text{grid}} \cup \mathcal{V}_B},\, H^{\mathcal{G}_1}\right)$$

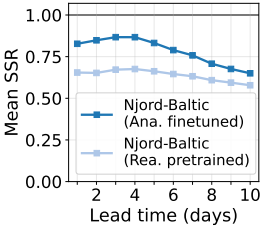
Decoding (M2G) targets only $\mathcal{V}_{\text{grid}}$, so the boundary is an input and never predicted.

- $B^t \in \mathbb{R}^{N_B \times d_B}$: SSH and T, S, U, V at 0, 47, 92 m ($d_B = 13$) on $1/12^\circ$ points $\mathcal{V}_B$ along the open boundary towards the North Sea
- A separate encoder [5] keeps the boundary at its *native* $1/12^\circ$ resolution, with no interpolation to 2 km
- Trained with GLORYS12 reanalysis and tested with GLO12 10-day forecasts, an *independent* operational system rather than the one being emulated

Regional skill and calibration



RMSE of the ensemble mean against NEMO analysis: T, S and U at 47 m, and sea ice thickness.

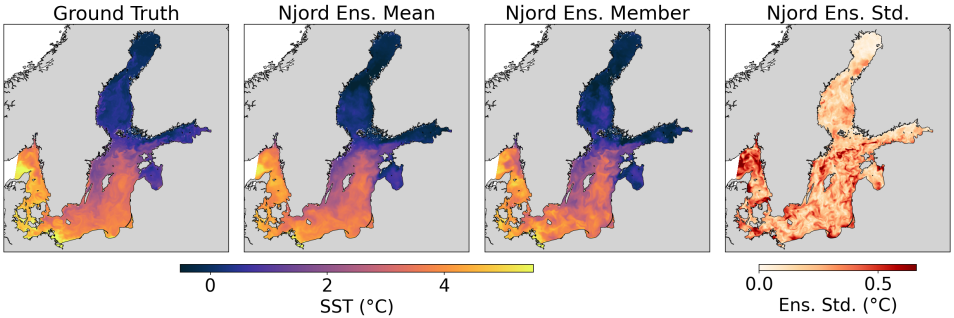


Mean SSR over Baltic variables.

Matches SeaCast [4] in RMSE while additionally providing **ensemble forecasts**.

Finetuning on analysis improves error and calibration; the ensemble is underdispersed at longer leads.

Baltic Sea surface temperature



SST at 10-day lead, initialized 2024-03-05. Ground truth is NEMO analysis.

- The ensemble mean is smooth, while *individual members* recover sharp gradients.
- Spread peaks in the *Danish Straits*, where exchange with the North Sea drives variability, and is zero under the winter ice in the north.

05
Conclusion

Takeaways

- **Generative ensemble forecasting** scales to high-resolution, irregular ocean grids.
- The latent variable GNN needs *one forward pass per sample*, fully parallel across members.
- Cluster meshes fit the coastline; a density channel keeps sea ice physical.
- Matches or beats deterministic ML models, with uncertainty, and gives the *lowest SST error against observations*.

Njord gives *calibrated ensemble ocean forecasts*, globally at 0.25° and regionally at 2 km, in seconds on a single GPU.

CONCLUSION

Limitations and outlook

LIMITATIONS

- Only 5–6 depth levels
- 0.25° rather than the native 1/12° grid
- Spread too low at early (global) and late (Baltic) leads
- Should be able to push error lower than GLO12 on analysis

OUTLOOK

- Subseasonal-to-seasonal (S2S) and climate forecasts
- Include biogeochemistry and ocean waves
- Coupled ocean–atmosphere ensembles
- Perturb initial conditions and forcing as additional sources of uncertainty

Thank you

Code: github.com/deinal/njord
Paper: arXiv:2605.15470
Contact: daniel.holmberg@helsinki.fi

[1] J. Oskarsson et al. Probabilistic weather forecasting with hierarchical graph neural networks. *NeurIPS*, 2024.
[2] S. Lang et al. AIFS-CRPS: Ensemble forecasting using a model trained with a CRPS-based loss. *npj AI*, 2026.
[3] A. El Aouni et al. OceanBench: A benchmark for data-driven global ocean forecasting systems. *NeurIPS*, 2025.
[4] D. Holmberg et al. Accurate Mediterranean Sea forecasting via graph-based deep learning. *Sci. Rep.*, 2025.
[5] S. Adamov et al. Building machine learning limited area models: Kilometer-scale weather forecasting in realistic settings. *arXiv*, 2025.

Code & Paper