Available online www.worldscientificnews.com

World Scientific News

An International Scientific Journal

WSN 221 (2026) 115-138 EISSN 2392-2192

Predicting Temperature–Salinity Profiles at Unobserved Stations and Times in a Glacier-Fed Antarctic Bay: Neighbour Information, Physical Constraints and Leakage-Aware Validation

Howard Chioma Chinagorom[1, *], Love Cherukei Nnoka[2]

[1] Department of Mathematics and Computer Science, University of Africa, Toru-Orua, Bayelsa State, Nigeria;
e-mail: howardchioma@gmail.com. Orcid: 0009-0005-4129-1291

[2] Department of Statistics, Captain Elechi Amadi Polytechnic, Rumuola, Port Harcourt, Rivers State, Nigeria;
e-mail: nnokalove@gmail.com, Orcid: 0009-0000-8381-7603

* Corresponding author: howardchioma@gmail.com

https://doi.org/10.65770/MLDT4430

ABSTRACT

Campaign hydrography in glacier-fed bays is sparse, making prediction at unsampled stations and dates difficult. We evaluated spatial and temporal transfer and physical constraints using 1,982 quality-controlled YSI EXO profiles from 24 Admiralty Bay stations, Antarctica (2018–2022). A graph-attention model with GRU-D-style temporal decay predicted temperature and salinity anomalies; a residual multilayer perceptron added static-stability and temperature–salinity (T–S) hull penalties. We compared five validation schemes with climatology, a spatial generalised additive model (GAM), gradient boosting, inverse-distance weighting (IDW), and persistence. In leave-one-site-out tests with same-day neighbours, IDW achieved a temperature root-mean-square error (RMSE) of 0.110 °C, versus 0.249 °C for climatology. In buffered leave-one-day-out tests (±7-day buffer), the lagged graph model achieved skill scores of 0.45 for temperature and 0.32 for salinity; graph parameters were fixed at leave-one-site-out medians. The apparent advantage of gradient boosting under station withholding disappeared when same-day data were excluded (temperature skill = −0.14), revealing campaign-identifier leakage. On a 12-station subset, increasing the stability-penalty weight from 0 to 3 reduced the share of profiles with density inversions from 37% to 18%, close to the base model's 20%, while increasing temperature RMSE by 0.004 °C. A summer T–S hull excluded 56% of same-day IDW predictions for June–November. Conformal coverage was near nominal across stations but deteriorated under seasonal shift for several models; same-day IDW remained near nominal. The stability penalty improved plausibility relative to the unconstrained residual model but not accuracy; the hull penalty had little effect on leave-one-site-out accuracy. Recent neighbouring profiles contained most of the transferable information.

Keywords: Admiralty Bay, King George Island, temperature–salinity profiles, spatial and temporal transfer, attention message passing, leakage-aware validation.

(Received 10 September 2026; Accepted 21 October 2026; Date of Publication 19 November 2026)

1. INTRODUCTION

Tidewater glaciers reshape the coastal ocean. Subglacial discharge, surface melt, and iceberg melt freshen and cool the water near glacier fronts, and warmer offshore water exchanges heat with the bay (Straneo & Cenedese, 2015). In bays of the West Antarctic Peninsula, these processes create strong gradients within a few kilometres. The observations that constrain them are costly, come mostly from austral summer, and are collected in short campaigns. Monitoring design and gap-filling therefore depend on a question that is rarely tested directly: what can be predicted at a place or time that nobody sampled?

Existing work addresses parts of this question. Process models explain fjord temperature in Greenland (Mas e Braga et al., 2025), graph neural networks (GNNs) now produce global and regional ocean forecasts (Holmberg et al., 2026), and physics-guided learning promises better extrapolation where data are sparse (Karniadakis et al., 2021). For Antarctic Peninsula bays, however, the literature is descriptive. It documents what the campaigns measured but does not test prediction at unobserved stations, days, or seasons. Validation practice adds a second problem. Random or purely spatial cross-validation can overstate skill when observations cluster in campaigns (Meyer & Pebesma, 2021; Roberts et al., 2017). Section 2 reviews these gaps in detail.

We therefore study transfer in Admiralty Bay, King George Island, using 1,982 quality-controlled profiles from 24 stations on 118 campaign days (Section 4). Each station yields an irregularly sampled time series of profiles, with campaigns spaced from days to months apart. We predict profiles at station–date combinations with no observations; we do not forecast future states. We build a station-graph model with attention over neighbouring stations, a GRU-D-style temporal decay, a residual network, and two physical penalties, and test it under five validation schemes against five baselines. We ask four questions. Can a station graph predict profiles at a withheld station better than a spatial generalised additive model (GAM) and gradient boosting? Does that skill persist when no same-day data exist and when the test season differs from the training season? Do static-stability and temperature–salinity (T–S) constraints improve accuracy, plausibility, or calibration? Does the graph structure matter?

The main findings are as follows. Same-day neighbours carry most of the skill: inverse-distance weighting (IDW) reaches a temperature root-mean-square error (RMSE) of 0.110 °C against 0.249 °C for climatology. Without same-day data, the lagged graph model is best on unobserved days (temperature skill +0.45; salinity skill +0.32). Gradient boosting looks strong under station-withheld validation largely because day of year can act as a campaign identifier. In the full leave-one-site-out ablation, the stability penalty reduces inversions relative to the unconstrained residual network, but the inversion rate remains above that of the graph base model; physical constraints do not consistently improve accuracy.

The contributions are of four kinds. **Empirical:** a predictive transfer test for Admiralty Bay profiles, to our knowledge the first, with a quality-controlled profile table. **Methodological:** a validation protocol that separates interpolation from transfer and exposes campaign-identifier leakage. **Physical:** an evaluation of when physical consistency constraints improve plausibility, including a weight sweep and a seasonal failure case. **Practical:** evidence that recent neighbouring profiles carry the transferable signal, which may inform the design of sparse monitoring.

2. RELATED WORK AND RESEARCH GAP

2.1. Hydrography and Meltwater in Glacier-Fed Bays

Admiralty Bay is a well-sampled system. The data descriptor of Osińska et al. (2023a) reports casts at 31 stations in glacial coves, the main bay, and Ezcurra Inlet. The study by Souza et al. (2025) analyses surface-water physicochemical parameters and their link to atmospheric conditions in the same bay. Both describe what the campaigns measured. In Greenland, a reduced-physics model reproduces fjord temperature across many fjords and shows which processes control it (Mas e Braga et al., 2025), while reviews summarise how plumes, icebergs, and shelf exchange set fjord structure (Straneo & Cenedese, 2015). These processes are understood in outline, but, to our knowledge, no study evaluates how well hydrographic structure can be predicted at unsampled locations in Antarctic Peninsula bays.

2.2. Learning from Sparse Hydrographic and Cryosphere Data

Neural networks reconstruct three-dimensional temperature and salinity fields from surface and in situ data in the Gulf Stream (Pauthenet et al., 2022), and graph neural networks forecast the global atmosphere and, more recently, the ocean (Holmberg et al., 2026; Lam et al., 2023). In glaciology, learning methods emulate subglacial hydrology (Verjans & Robel, 2024), model glacier dynamics with physical structure (Bolibar et al., 2020), and transfer runoff models to unseen conditions (Maciel-Seidman et al., 2026). Most of these studies work at regional or larger scale, on gridded or modelled data. Water-quality time series are another established target. Howard and Apeli (2026) compared a hybrid CNN–LSTM–attention model with LSTM, GRU, and ARIMA baselines for seven contaminants at three flow stations, reporting lower error at a seven-day horizon under chronological, walk-forward, and rolling-origin validation. Howard and Howard (2026) also evaluated hybrid neural-network approaches for produced-water quality forecasting and management in Nigerian oil fields. These studies forecast monitored water-quality series, whereas our task is spatial and temporal transfer of hydrographic profiles. None of the studies above, to our knowledge, evaluates a station graph on raw casts from a small glacier-fed embayment.

2.3. Physical Constraints in Learned Models

Constraints enter learning through the architecture, the loss, or post-processing (Beucler et al., 2021; Karniadakis et al., 2021). Pauthenet et al. (2022) remove static instabilities from reconstructed profiles. Whether constraints improve accuracy depends on whether they encode information that the data lack, since otherwise they add nothing. Two points are less studied: how constraint weight trades accuracy against plausibility, and whether a constraint estimated from one season carries to another.

2.4. Validation and Uncertainty Under Dependence

Clustered observations break the independence assumption of random cross-validation, and blocked schemes address it (Meyer & Pebesma, 2021; Roberts et al., 2017). Campaign data add a temporal version of the same problem: any covariate that nearly identifies a campaign lets a flexible model recover campaign-level values from the training set. Conformal prediction gives finite-sample coverage under exchangeability, which fails under covariate shift (Angelopoulos & Bates, 2023; Tibshirani et al., 2019; Vovk et al., 2005).

2.5. Research Gap

Table 1 summarises the gaps. In short, no study to our knowledge combines a station-graph model, physical constraints, and campaign-aware validation for a glacier-fed bay.

Table 1. Research Gaps and How This Study Addresses Them.

Gap	What exists	What this study adds
Empirical: Antarctic bays	Campaign-based descriptions of Admiralty Bay hydrography (Osińska et al., 2023a; Souza et al., 2025).	A predictive test at unobserved stations, days and seasons on 1,982 quality-controlled profiles.
Methodological: validation	Guidance on spatial and temporal blocking (Meyer & Pebesma, 2021; Roberts et al., 2017); chronological, walk-forward, and rolling-origin validation for water-quality series (Howard & Apeli, 2026).	A documented campaign-identifier leak through day of year, and a buffered leave-one-day-out test.
Physical: constraints	Stability adjustment and analytic constraints in profile reconstruction (Beucler et al., 2021; Pauthenet et al., 2022).	A weight sweep, ablations, and a seasonal test of whether constraints raise accuracy, plausibility, or both.
Uncertainty	Conformal methods and their shift limits (Angelopoulos & Bates, 2023; Tibshirani et al., 2019).	Cross-fitted conformal coverage under five spatial and temporal shifts.
Context: small embayments	Basin-scale ocean graph models and fjord-transfer proposals (Holmberg et al., 2026).	A small station graph with a learned metric and temporal decay for a glacier-fed bay.

3. PHYSICAL FRAMEWORK AND HYPOTHESES

3.1. Physical Basis

Four physical ideas motivate the model, and each sets an expectation. Surface heat exchange, winter cooling, and meltwater input act on the whole bay, so anomalies relative to the seasonal mean should be spatially coherent over kilometres and persist for days to weeks. This motivates aggregating neighbour anomalies, with a decay towards the seasonal mean. Outside transient events, density should not decrease with depth, and a penalty on density inversions encodes this static stability. Waters formed by mixing conserved properties lie in the convex hull of their end-members in temperature–salinity space. The hull is necessary but not sufficient: temperature is not a conservative tracer during ice-melt mixing (Gade, 1979) or surface exchange, and the hull depends on the end-members present in the training data. Sensor noise also sets a floor on detectable inversions, so the stability penalty needs a tolerance.

3.2. Hypotheses

We organise five expectations as hypotheses. H1 predicts that same-day neighbours carry most of the transferable information, so neighbour-based models should outperform covariate-only models. H2 predicts that, on unobserved days, a model using lagged neighbour anomalies will outperform covariate-only models. H3 is the original expectation that constraints improve accuracy where data are sparse; H3b predicts that they improve plausibility. H4 predicts that a hull learned in one season will fail in another. H5 predicts that learned edges will outperform distance-based edges.

The original study plan asked whether the graph could predict profiles at withheld stations and whether constraints could improve accuracy. The separate same-day and lagged comparisons (H1 and H2), seasonal hull analysis (H4), graph-structure test (H5), and buffered leave-one-day-out test were added during analysis. These specific comparisons are therefore exploratory. Table 12 summarises each expectation and its result.

4. STUDY AREA AND DATA

4.1. Observations

We use the Admiralty Bay data set published by Osińska et al. (2023b). A YSI EXO multiparameter sonde recorded casts at 31 stations in glacial coves, the main bay, and Ezcurra Inlet between December 2018 and January 2022. Sampling ran at 0.2 Hz until 30 December 2019 and at 1 Hz afterwards. The cable limits casts to about 100 m. The authors state that no formal quality assurance was applied, which motivates the pipeline in Section 4.2.

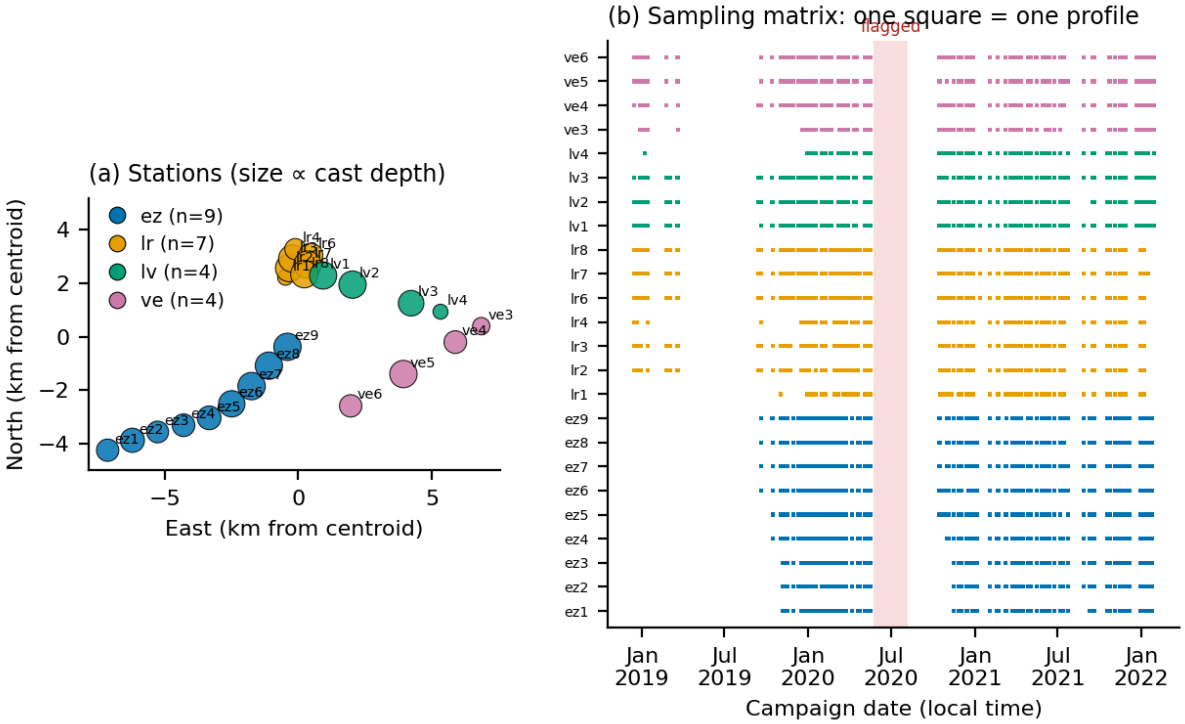


Figure 1. Study Stations and Sampling Coverage. (A) The 24 Profiling Stations by Station-Code Group; Marker Size Scales With 95th-Percentile Cast Depth; Positions Are Kilometres East and North of the Centroid. (B) Sampling Matrix: One Square Is One Quality-Controlled Profile. The Shaded Interval Marks the Flagged 2020 Salinity Episode.

4.2. Quality Control and Profile Construction

We reduced raw logs to site-day profiles in several steps (Table 2). Salinity is practical salinity, in PSU. We removed records with missing values or depth below 0.5 m and split each station log into casts at time gaps above 120 s. We kept the downcast from the start to the deepest point, despiked with a rolling-median test, and averaged into 2 m bins centred at 1, 3, . . . , 99 m, interpolating interior gaps of up to three bins. The rolling window was shorter for 0.2 Hz casts. We defined campaign days in local time (UTC−3) so that evening casts stay with their campaign. When a station was cast more than once in a day, we kept the deepest cast. Density follows the EOS-80 one-atmosphere formula (Millero & Poisson, 1981):

$$\sigma_t(T,S) = \rho_w(T) + A(T)\,S + B(T)\,S^{3/2} + C\,S^2 - 1000 \qquad (1)$$

Pressure effects on σ_t are nearly uniform over the upper 100 m, so we use σ_t differences between adjacent bins as a stability diagnostic. TEOS-10 density (Intergovernmental Oceanographic Commission et al., 2010) could give different inversion rates; we did not evaluate this sensitivity.

Table 2. Quality-Control and Profile-Construction Log.

Step	Count	Comment
Raw records	290,436	YSI EXO logs, 31 stations, Dec 2018–Jan 2022
Missing temperature, salinity, or depth	14,548	Removed
Depth < 0.5 m (in air / surface soak)	3,112	Removed
Records in downcasts	267,072	2,223 casts; segmented at gaps > 120 s; start to deepest point
Spike values removed	152	Rolling-median test, 0.25 threshold; 3-point window for 0.2 Hz casts, 9-point for 1 Hz
Site-day profiles	2,216	2 m bins, gaps ≤ 6 m interpolated; 330 profiles from 0.2 Hz sampling
Profiles at 24 profiling stations	2,136	7 stations with shallow or point-like casts excluded
Outlier profiles removed	17	Profile-median S or T departed from that day's median by >0.5 PSU or >1.5 °C, respectively
Flagged salinity episode removed	139	24 May–6 Aug 2020 (2 profiles also met the outlier rule)
Analysis set	**1,982**	**24 stations, 118 campaign days, 68,926 bin observations**

Counts come from our pipeline. They differ from the 3,045 measurements in the data descriptor because we excluded the seven shallow stations and because casts under 20 records or 3 m cannot form a profile.

4.3. Profiling Stations and the Flagged Salinity Episode

Seven stations log only about 7–17 records a day and cast to a few metres. They behave as near-surface point measurements, so we excluded them. This leaves 24 profiling stations in four code groups (Table 3; ez corresponds to the nine Ezcurra Inlet stations). Ezcurra stations were first sampled in September 2019, so the December–May training block has no Ezcurra profiles before December 2019.

Table 3. Profiling Stations by Station-Code Group.

Station group	Stations	Profiles	95th-percentile cast depth (m)	First to last profile
ez	9	722	56–98	2019-09-19 to 2022-01-24
lr	7	601	21–97	2018-12-15 to 2022-01-15
lv	4	339	18–97	2018-12-15 to 2022-01-28
ve	4	320	30–97	2018-12-15 to 2022-01-28
All	**24**	**1,982**	**18–98**	**2018-12-15 to 2022-01-28**

Between 17 and 24 May 2020, median salinity rose by about 0.55 PSU at all stations at once. It stayed near 35.1 until early August and returned to about 34.4 in October. The spread across stations that week was only 0.04 PSU. Strong winter brine rejection and an unrecorded conductivity calibration shift both fit this pattern, and the data cannot separate them. We excluded the 139 profiles from 24 May to 6 August 2020 from the main analysis and used them in a sensitivity run (Section 6.7). We also removed 17 outlier profiles whose median salinity or temperature departed from the day's median by more than 0.5 PSU or 1.5 °C.

Vertical structure is weak at the bin scale. Salinity changes by a few tenths of a PSU over 100 m. Adjacent-bin inversions larger than 0.02 kg m^{-3} occur in 0.57% of bin pairs and in 12.4% of profiles. That rate is the reference for the plausibility metrics.

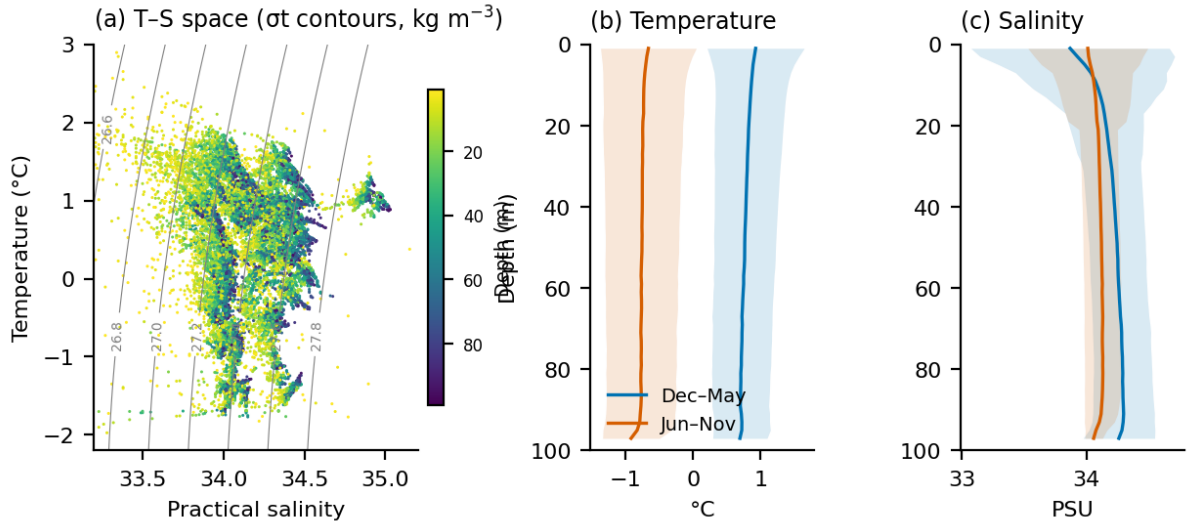


Figure 2. Structure of the Analysis Data. (A) T–S Diagram Of 25,000 Randomly Drawn Bin Observations Coloured By Depth, With σ_t Contours. (B, C) Mean Temperature and Salinity Profiles ±1 Standard Deviation for Dec–May and Jun–Nov. Winter Water Is Colder and Lies Largely Outside the T–S Range of Summer.

5. METHODOLOGY

5.1. Prediction Tasks

The target is the profile of temperature T and salinity S on 50 depth levels at station s on campaign day d. Density σ_t follows from the prediction, and the depth extent of the target profile is known. We distinguish two information settings. In **Task A (same-day)** the model may use profiles from other stations on the same day, which is spatial infilling. In **Task B (lagged)** it may use only the latest earlier profile at each other station (median lag 8 days, 90th percentile 20 days). Task B is the harder transfer claim because the model must carry information across time. In time-series terms, each station is an irregularly sampled series of profiles. We handle the irregular gaps through the decay of lagged anomalies (Section 5.2), and the task is prediction at unobserved stations and dates (gap-filling and transfer), not forecasting. Figure 3 summarises the framework.

5.2. Graph Base Model

We compute a seasonal climatology per level and variable with a circular Gaussian kernel in day of year (bandwidth h = 8 days), using training profiles only. The bandwidth was chosen on pooled leave-one-site-out error, which slightly favours the baseline (Section 6.8).

$$c_v(k,d) = \frac{\sum_i w\,(d,d_i)\,y_{ivk}}{\sum_i w\,(d,d_i)}, \qquad w(d,d_i) = \exp\left[-\frac{1}{2}\left(\frac{\Delta_{\mathrm{circ}}(d,d_i)}{h}\right)^2\right] \tag{2}$$

Each source profile j contributes an anomaly relative to the climatology at its own date:

$$\delta_{jvk} = y_{jvk} - c_v(k,d_j) \tag{3}$$

Stations are nodes with static features φ = (east, north, cast depth). Attention over source stations uses a learned diagonal metric with length scales ℓ_m. It is masked to sources with valid data at level k ($u_{jk} \in \{0,1\}$). Elapsed time Δt since the source observation enters through an exponential decay in the style of GRU-D, a recurrent network for irregularly sampled data (Che et al., 2018). The decay relaxes anomalies towards the climatology with learned rates ω_v:

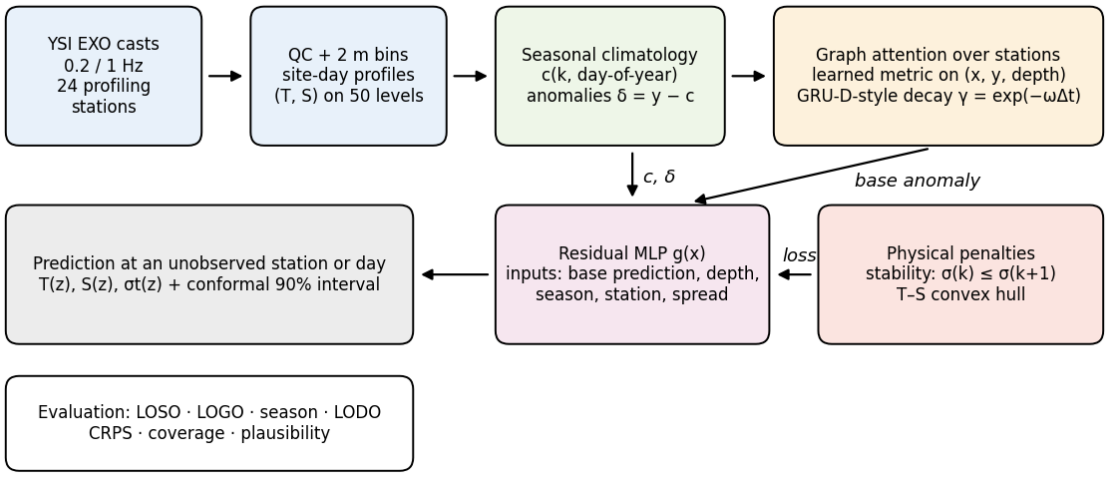


Figure 3. Framework. Quality-Controlled Profiles Feed a Seasonal Climatology and a Graph-Attention Base Model With Temporal Decay. A Residual Network Corrects the Base Prediction Under Stability and T–S Penalties. Five Validation Schemes Evaluate the Result.

$$e_{sj} = -\sum_{m=1}^{3}\left(\frac{\phi_{sm} - \phi_{jm}}{\ell_m}\right)^2, \qquad a_{sjk} = \frac{u_{jk}\exp(e_{sj})}{\sum_{j'} u_{j'k}\exp(e_{sj'})} \tag{4}$$

$$\gamma_{jv} = \exp(-\omega_v \Delta t_j) \tag{5}$$

The base prediction is the climatology plus the decayed, attention-weighted anomaly, scaled by a learned shrinkage β_v:

$$\hat{y}_{svk}^{\text{base}} = c_v(k, d_s) + \beta_v \sum_{j \neq s} a_{sjk}\ \gamma_{jv}\ \delta_{jvk} \tag{6}$$

The model has seven parameters: three length scales, two decay rates and two shrinkage factors. We fit them by L-BFGS-B (at most 30 iterations, on up to 500 random training profiles) on standardised squared error. Training targets exclude their own station from the sources, which mimics withholding a station. In Task A, $\Delta t = 0$ and the decay is inactive.

5.3. Residual Network and Physical Constraints

A one-hidden-layer network (multilayer perceptron, MLP) g_θ with 24 tanh units, shared across depth, adds a residual to the base prediction. Its inputs are the base prediction, climatology, target depth, seasonal encoding, station coordinates, cast depth, attention-weighted spread of source anomalies, and number of valid sources. The symbol s_v is the training standard deviation of the anomaly.

$$\hat{y}_{svk} = \hat{y}_{svk}^{\text{base}} + s_v\, g_\theta(x_{sk})_v \tag{7}$$

Training minimises a data term plus two penalties. The stability penalty charges density inversions above a tolerance $\tau = 0.02$ kg m^{-3} per 2 m bin, chosen so that about 0.6% of observed bin pairs exceed it. Here $\mathcal{P}$ is the set of adjacent valid bin pairs and $\tau_0 = 0.02$ sets the penalty scale. The mixing penalty charges the squared distance of each standardised (S, T) pair to the convex hull $\mathcal{H}$ of training observations, trimmed at the 0.25th and 99.75th percentiles of each axis.

$$\mathcal{L} = \frac{1}{2n}\sum_{s,k,v}\left(\frac{\hat{y}_{svk} - y_{svk}}{s_v}\right)^2 + \lambda_s P_{\text{stab}} + \lambda_m P_{\text{mix}} \tag{8}$$

$$P_{\text{stab}} = \frac{1}{|\mathcal{P}|\,\tau_0^2}\sum_{(s,k)\in\mathcal{P}}\left[\max\left(0,\ \sigma_{s,k} - \sigma_{s,k+1} - \tau\right)\right]^2 \tag{9}$$

$$P_{\text{mix}} = \frac{1}{n\,d_0^2}\sum_{s,k}\text{dist}^2\left(\mathbf{z}_{sk}, \mathcal{H}\right), \qquad \mathbf{z} = \left(\frac{S - \mu_S}{s_S}, \frac{T - \mu_T}{s_T}\right) \tag{10}$$

We fixed $\lambda_s = \lambda_m = 0.3$ and $d_0 = 0.15$ and later varied the weights (Section 6.8). We trained with Adam for 80 full-batch epochs on at most 800 training profiles. Deep-learning libraries were not available in our computing environment, so the model is a NumPy implementation with hand-written gradients, including analytic derivatives of the equation of state. It is a one-layer attention message-passing model, smaller than the architectures in the GNN literature (Battaglia et al., 2018; Veličković et al., 2018).

5.4. Baselines

We compare against five baselines. The climatology is the reference for skill. The spatial GAM is a ridge-penalised spline model with tensor-product interactions between depth, day of year (periodic), and position, used because standard GAM software could not be installed in our computing environment (Wood, 2017). Gradient boosting uses depth, position, bathymetry, and season (Friedman, 2001), with a per-station variant for the seasonal split. Inverse-distance weighting averages neighbouring profiles (power 2) with the same source rule as the graph. Own-site persistence repeats the previous cast and is used in the unobserved-day test. We did not evaluate zero-shot foundation models because no pretrained weights could be obtained.

5.5. Validation design

We use five validation schemes. In leave-one-site-out (LOSO), each station is withheld in turn from training and from the source set. In leave-group-out (LOGO), each code group is withheld, which removes the nearest neighbours of every test station. In the seasonal split (SEAS), we train on December–May profiles and test on June–November profiles, and the climatology, hull, and all parameters use training months only. The combined scheme, LOSO×SEAS, withholds a station and restricts the test to June–November. In the buffered leave-one-day-out test (LODO), we withhold all profiles of one campaign day together with every training profile within ±7 days, while lagged neighbours remain available as inputs. For tractability, LODO evaluates the base graph model with parameters fixed at their LOSO medians.

Time-ordered schemes such as chronological splits, walk-forward, and rolling-origin validation are used for forecasting water-quality series (Howard & Apeli, 2026). They withhold the future. Our tests instead withhold stations, seasons, or a buffered campaign day, because our task is gap-filling and transfer, not forecasting.

Leakage notes. In LOSO and LOGO, models without neighbour inputs (climatology, GAM, gradient boosting) still train on the other stations' profiles from the *same campaign day*. Day of year is almost unique to each campaign, so tree models can identify the campaign. Their LOSO and LOGO scores therefore describe campaign-informed models, not lagged ones. We mark them with † and use SEAS and LODO, which exclude same-day training data, for the cleanest lagged comparisons. In LODO, however, graph parameters are fixed at their LOSO medians, so this test is not fully independent of the withheld days.

5.6. Metrics

We pool errors over test bins and report RMSE for T, S and σ_t, plus skill against climatology based on mean squared error (MSE). Confidence intervals are 95% percentile intervals from 1,000 block-bootstrap resamples of campaign days, and skill differences use the same resamples.

$$\mathrm{SS} = 1 - \frac{\mathrm{MSE}_m}{\mathrm{MSE}_{\mathrm{clim}}} \tag{11}$$

Plausibility is the share of profiles and bin pairs with an inversion above τ, and the share of predicted (S, T) points outside the training hull. For probabilistic skill we use the Gaussian continuous ranked probability score (CRPS; Gneiting & Raftery, 2007). Its scale is the root-mean-square cross-fitted residual in three depth bands (0–20, 20–50 and 50–100 m). We also use split-conformal 90% intervals (Angelopoulos & Bates, 2023; Vovk et al., 2005), calibrated by three-fold cross-fitting over training stations (cf. Barber et al., 2021).

$$\mathrm{CRPS}(\mathcal{N}(\mu,\sigma^2), y) = \sigma\left[z\,(2\Phi(z) - 1) + 2\varphi(z) - \frac{1}{\sqrt{\pi}}\right], \quad z = \frac{y-\mu}{\sigma} \tag{12}$$

$$q_{vb} = \mathrm{Quantile}_{\lceil (n+1)(1-\alpha)\rceil/n}(|r_i|), \qquad \hat{y} \pm q_{vb} \tag{13}$$

Intervals carry no guarantee under spatial or seasonal shift (Tibshirani et al., 2019), so we report empirical coverage for each scheme. In the cross-fit, base-model parameters come from the full training fit.

6. RESULTS

6.1. Same-Day Information Dominates

With same-day neighbours (Task A), neighbour-based models clearly beat models that predict from position and season (Table 4, Figure 4). IDW reaches a temperature RMSE of 0.110 [0.101–0.119] °C against 0.249 °C for climatology. The graph base model is close (0.122 °C; skill +0.76). Its learned metric tracks local geometry, with length scales of 0.9 km east–west, 1.3 km north–south, and 49 m in cast depth (Figure 8a). The residual network and constraints leave temperature RMSE unchanged (0.122 °C). Salinity is harder: no model exceeds skill of +0.32, and intervals are wide because a few stations (ez4, ez6, lv4) have station-specific offsets that no model reproduces (Figure 12). Errors are largest in the upper 20 m (Figure 5).

Table 4. Leave-One-Site-Out, Same-Day (Task A) Accuracy; 1,982 Profiles. Bold Marks the Lowest Value In Each Error Column. † Climatology, GAM and Gradient Boosting Use No Neighbour Inputs but Train on Same-Day Profiles of Other Stations (Section 5.5).

Model	RMSE T (°C) [95% CI]	RMSE S (PSU) [95% CI]	RMSE σ_t (kg m^{-3})	Skill T	Skill S
Climatology†	0.249 [0.230–0.266]	0.301 [0.205–0.400]	0.245	+0.00	+0.00
Spatial GAM (ridge splines) †	0.251 [0.233–0.268]	0.308 [0.217–0.404]	0.250	−0.01	−0.05
Gradient boosting†	0.138 [0.127–0.147]	0.275 [0.122–0.409]	0.223	+0.69	+0.16
IDW of neighbours	**0.110 [0.101–0.119]**	**0.249 [0.090–0.378]**	**0.201**	+0.80	+0.32
Graph base model	0.122 [0.112–0.132]	0.262 [0.103–0.397]	0.212	+0.76	+0.24
Graph + MLP	0.124 [0.114–0.133]	0.258 [0.113–0.384]	0.209	+0.75	+0.26
Graph + MLP + both	0.122 [0.113–0.131]	0.253 [0.114–0.374]	0.205	+0.76	+0.29

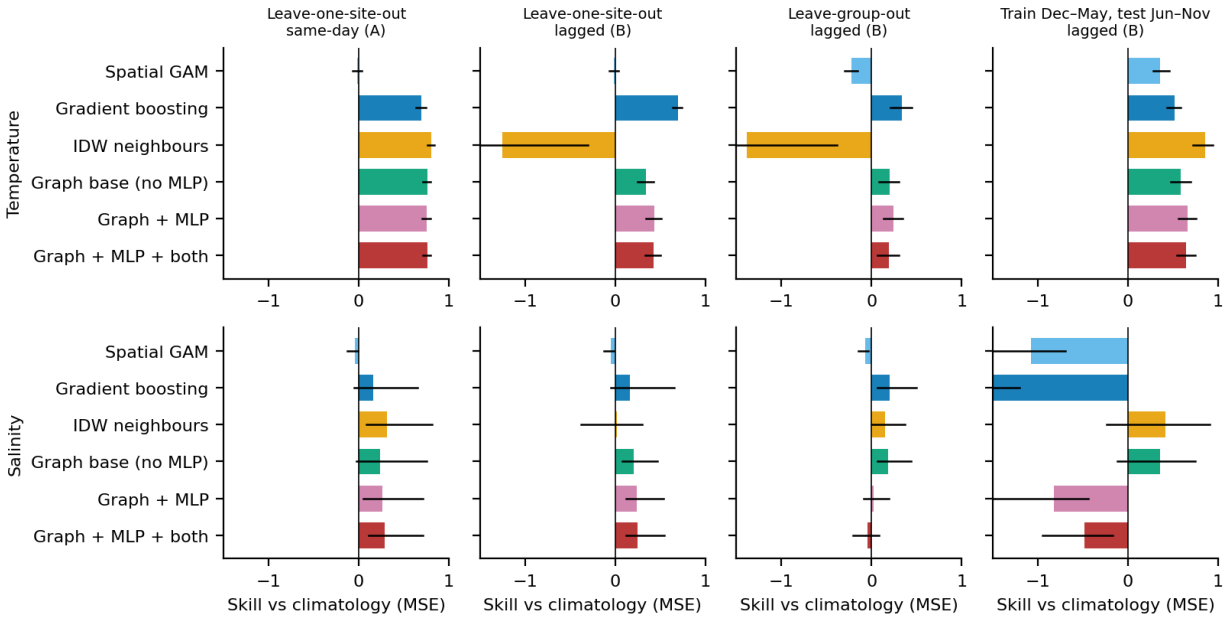


Figure 4. Skill Against Climatology (MSE) for Temperature (Top) and Salinity (Bottom) Across Four Validation Settings. Horizontal Lines Are 95% Day-Block Bootstrap Intervals. The Lagged LOSO and LOGO Panels Include Campaign-Informed Baselines (Section 5.5); the Seasonal Panel Does Not.

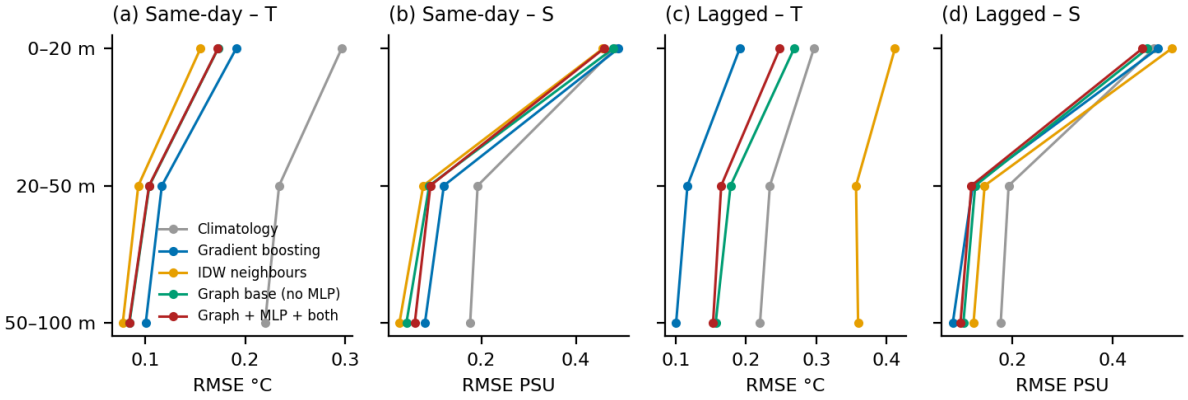


Figure 5. Depth-Resolved RMSE for Leave-One-Site-Out, Same-Day (A, B) and Lagged (C, D) Inputs.

6.2. Lagged Transfer and the Leakage Problem

With lagged inputs in leave-one-site-out, gradient boosting looked strongest for temperature (RMSE 0.138 °C, skill +0.69), and the lagged graph base model reached 0.202 °C (skill +0.34). Lagged IDW without shrinkage was worse than climatology (skill −1.26), probably because raw values from a campaign days or weeks earlier carry the wrong seasonal state. The gradient-boosting result is not a lagged-transfer result. Its training set contains the other stations on the same campaign day. The buffered leave-one-day-out test shows this (Table 5, Figure 6).

Table 5. Buffered Leave-One-Campaign-Day-Out Test (Task B, ±7 D Buffer; 1,982 Profiles, 118 Days). No Model Sees Same-Day or Near-Same-Day Training Profiles. The Graph Model Uses Lagged Neighbours As Inputs, With Parameters Fixed at Their LOSO Medians.

Model	RMSE T (°C) [95% CI]	RMSE S (PSU) [95% CI]	Skill T [95% CI]	Skill S [95% CI]
Climatology	0.308 [0.282–0.334]	0.334 [0.251–0.426]	+0.00 [0.00, 0.00]	+0.00 [0.00, 0.00]
Spatial GAM (ridge splines)	0.271 [0.247–0.294]	0.325 [0.242–0.420]	+0.23 [0.19, 0.27]	+0.06 [0.01, 0.12]
Gradient boosting	0.328 [0.294–0.362]	0.384 [0.310–0.470]	−0.14 [−0.29, 0.02]	−0.32 [−0.64, −0.16]
Own-site persistence	0.392 [0.301–0.502]	0.366 [0.240–0.482]	−0.62 [−1.67, 0.04]	−0.20 [−1.07, 0.33]
IDW of neighbours	0.374 [0.283–0.489]	0.298 [0.194–0.400]	−0.47 [−1.53, 0.14]	+0.20 [−0.10, 0.48]
Graph base model	**0.228 [0.209–0.246]**	**0.276 [0.168–0.379]**	+0.45 [0.36, 0.53]	+0.32 [0.19, 0.56]

In this clean test the ordering reverses. Gradient boosting has no skill (temperature −0.14, 95% CI −0.29 to 0.02; salinity −0.32). The spatial GAM is modest (+0.23 for temperature), because its smooth seasonal fit cannot memorise campaigns. The lagged graph model is best (temperature skill +0.45, salinity +0.32). Own-site persistence also has negative skill. The benefit therefore does not come from remembering a station's own last cast. It comes from anomalies that neighbouring stations share. For evaluation practice, the lesson is direct: a covariate model with fine day-of-year resolution can pass for a transfer model under spatial cross-validation when campaigns are the unit of variability.

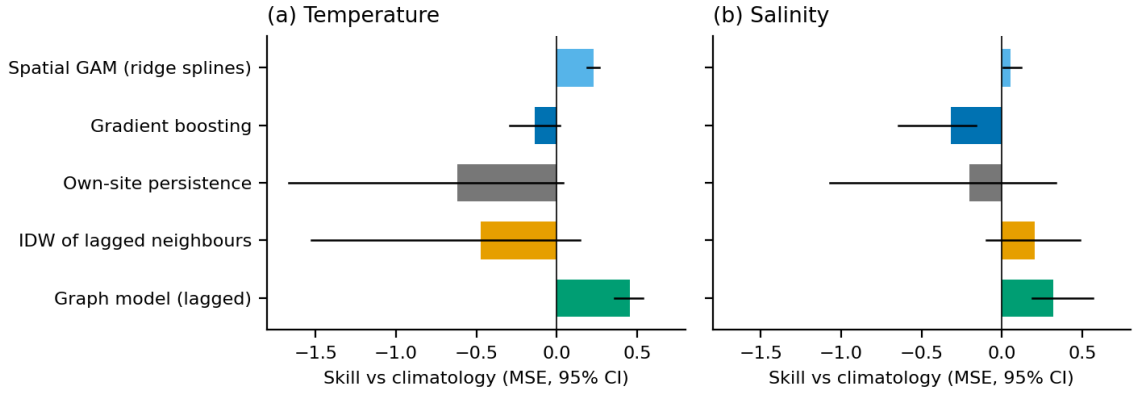


Figure 6. Skill Against Climatology in the Buffered Leave-One-Day-Out Test, with 95% Day-Block Bootstrap Intervals. Only the Lagged Graph Model Has Clearly Positive Skill for Both Variables.

6.3. Withheld Groups and Seasonal Transfer

Withholding a whole station group lowers skill but keeps the ordering for same-day inputs (LOGO, Task A: IDW 0.181 °C, graph 0.183 °C, boosting 0.213 °C, climatology 0.261 °C; Tables 6 and 7). Seasonal transfer is far harder for models that rely on position and season. Trained on December–May, climatology reaches 1.323 °C RMSE on June–November, the GAM 1.061 °C, gradient boosting 0.919 °C and the per-station boosting model 0.940 °C. With same-day neighbours the error falls to 0.097 °C (IDW) and 0.154 °C (graph), because the neighbours supply the winter state. With lagged inputs the graph base model keeps +0.59 temperature skill (RMSE 0.851 °C) and +0.36 salinity skill. The GAM and gradient boosting have strongly negative salinity skill (−1.08 and −2.28). Lagged IDW is best for temperature (0.499 °C), because the lag is short compared with the seasonal change. Errors of lagged models peak in July–September, when the winter transition is fastest and campaign days are few (Figure 7). Withholding station and season together gives the same ordering as the seasonal split (LOSO×SEAS columns).

Table 6. Temperature RMSE (°C) by Validation Scheme. A = Same-Day, B = Lagged Inputs; LS = Leave-One-Site-Out Within the Seasonal Split. † Baselines Without Neighbour Inputs Train on Same-Day Profiles of Other Stations in LOSO and LOGO (Section 5.5).

Model	LOSO-A	LOSO-B	LOGO-A	LOGO-B	SEAS-A	SEAS-B	LS-A	LS-B
Climatology	0.249†	0.249†	0.261†	0.261†	1.323	1.323	1.335	1.335
Spatial GAM	0.251†	0.251†	0.289†	0.289†	1.061	1.061	1.058	1.058
Gradient boosting	0.138†	**0.138†**	0.213†	**0.213†**	0.919	0.919	0.925	0.925
IDW of neighbours	**0.110**	0.374	**0.181**	0.403	**0.097**	**0.499**	**0.097**	**0.499**
Graph base model	0.122	0.202	0.183	0.234	0.154	0.851	0.153	0.820
Graph + MLP + both	0.122	0.189	0.194	0.235	0.137	0.788	0.167	0.850

Table 7. Salinity RMSE (PSU) by Validation Scheme. † As In Table 6.

Model	LOSO-A	LOSO-B	LOGO-A	LOGO-B	SEAS-A	SEAS-B	LS-A	LS-B
Climatology	0.301†	0.301†	0.303†	0.303†	0.348	0.348	0.349	0.349
Spatial GAM	0.308†	0.308†	0.314†	0.314†	0.502	0.502	0.491	0.491
Gradient boosting	0.275†	0.275†	0.272†	**0.272†**	0.631	0.631	0.598	0.598
IDW of neighbours	**0.249**	0.298	**0.249**	0.279	**0.167**	**0.267**	**0.167**	**0.267**
Graph base model	0.262	0.268	0.253	0.274	0.177	0.279	0.179	0.285
Graph + MLP + both	0.253	**0.261**	0.259	0.311	0.289	0.424	0.247	0.388

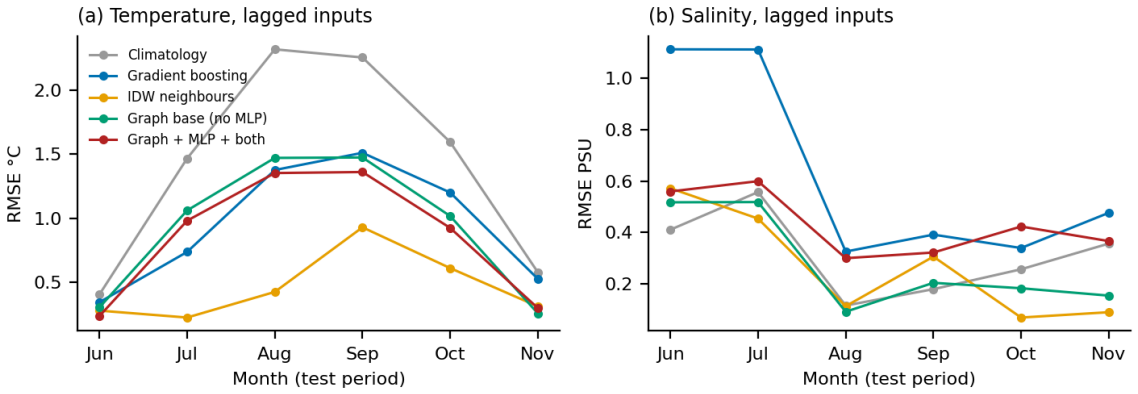


Figure 7. Monthly RMSE in the Seasonal Split (Training December–May, Lagged Inputs) for Temperature (A) and Salinity (B).

6.4. What the Learned Graph Looks Like

In Task A, the fitted metric is local (Figure 8a). Weights fall off within about a kilometre horizontally and with a depth scale near 50 m, so a station is predicted mainly from its closest neighbours with a similar cast depth. In Task B, the fitted metric stretches to the upper bounds of the search range (ℓ_x = 60 km, $\ell_z \approx 400$ m; Figure 8b), which makes the weights nearly uniform. The decay rate for temperature sits at its lower bound, so anomalies persist. The salinity rate is 0.007 d^{-1}, a half-life of about 100 days. Consistent with this, replacing the learned graph with a uniform graph or a distance-only graph does not change lagged accuracy (Table 8). In the lagged setting, the model uses the bay-wide seasonal anomaly, not the station geometry. For same-day inputs, distance-only IDW is at least as accurate as the learned graph, so the learned metric has not shown an advantage in either setting.

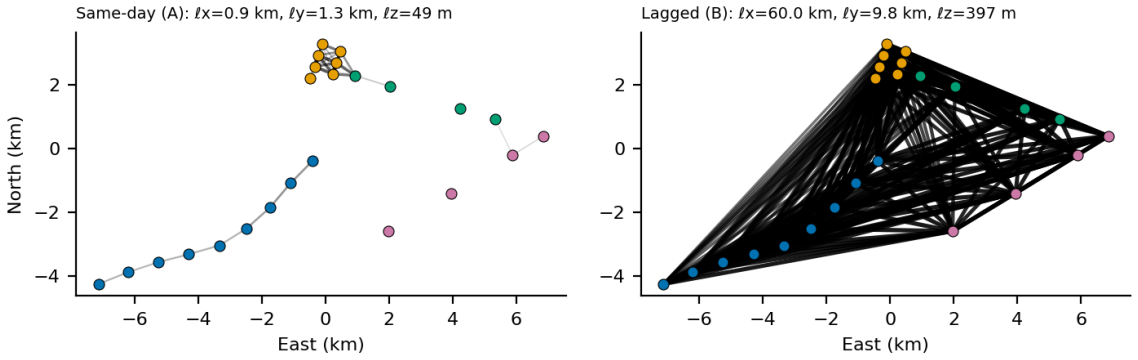


Figure 8. Learned Station Graph (Median Parameters Over Leave-One-Site-Out Folds). Line Width and Opacity Scale with Attention Weight Between Station Pairs. (A) Same-Day Inputs. (B) Lagged Inputs, Where the Metric Saturates and the Graph Is Nearly Uniform.

6.5. Physical Constraints

The stability penalty improves plausibility but neither physical penalty consistently improves accuracy (Table 8, Figure 9). The base graph model produces density inversions in 18% of profiles, against 12.4% observed. The unconstrained residual network raises this to 36%. The stability penalty lowers it to 27%, with nearly unchanged temperature RMSE (0.189 °C against 0.188 °C). The hull penalty has no measurable effect in leave-one-site-out, because only 0.6% of predictions fall outside the hull to begin with. The spatial GAM never produces an inversion, but it also has no skill. Across schemes, constraint effects on RMSE are small and of mixed sign. In the seasonal split with same-day inputs, constraints lowered temperature RMSE (0.172 to 0.137 °C) but raised salinity RMSE (0.260 to 0.289 PSU). With lagged inputs, the salinity pattern reversed (0.471 to 0.424 PSU). Under seasonal shift, the residual network makes salinity worse than the base model (0.471 PSU against 0.279 PSU; Table 7).

The hull exposes a general limitation. Built from December–May observations, it excludes 56% of the points that same-day IDW predicts for June–November. IDW closely reproduces the observations, so this result indicates that many winter observations lie outside the summer T–S envelope (Figure 2). A constraint estimated from one season may therefore mis specify another. A hull on T and S alone is also a weak statement of conservative mixing: it does not account for latent-heat exchange during meltwater mixing (Gade, 1979) and does not separate end-members that share a density.

Table 8. Ablation in Leave-One-Site-Out with Lagged Inputs. Plausibility Is Judged Against 12.4% of Observed Profiles Having an Inversion Above 0.02 Kg M^{-3}.

Variant	RMSE T (°C)	RMSE S (PSU)	RMSE σ_t	Profiles with inversion	Inverted bin pairs	Points outside T–S hull
Graph base model	0.202	0.268	0.217	18%	0.7%	0.1%
Graph + MLP	**0.188**	0.262	0.213	36%	1.6%	0.6%
Graph + MLP + stability	0.189	**0.261**	0.212	27%	1.1%	0.5%
Graph + MLP + T–S hull	**0.188**	0.262	0.213	36%	1.6%	0.5%
Graph + MLP + both	0.189	**0.261**	0.212	27%	1.1%	0.3%
Both, uniform graph	**0.188**	**0.261**	**0.211**	26%	1.0%	0.3%
Both, distance-only graph	0.189	**0.261**	0.212	27%	1.1%	0.3%
Both, no temporal decay	0.190	0.262	0.212	24%	1.0%	0.5%

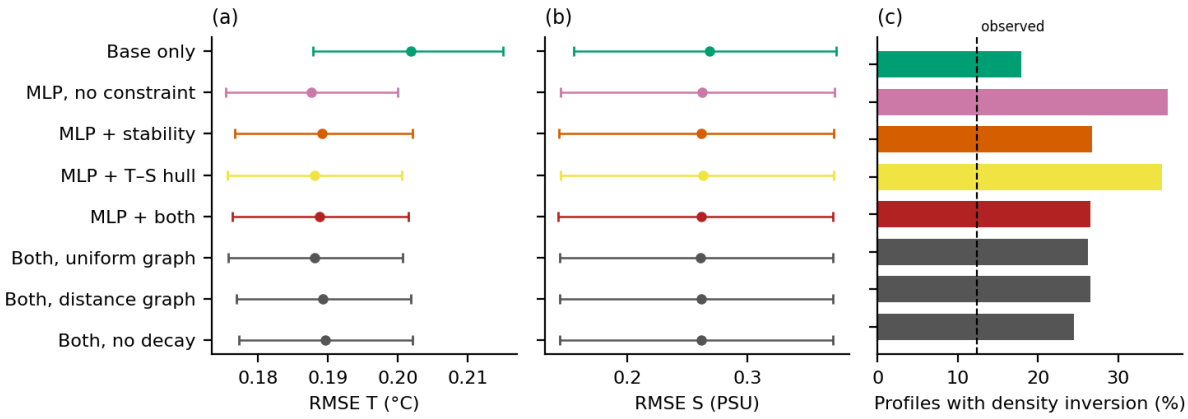


Figure 9. Ablation (Leave-One-Site-Out, Lagged Inputs). (A, B) RMSE With 95% Intervals. (C) Share of Profiles with a Density Inversion; the Dashed Line Shows the Observed Share.

6.6. Uncertainty

Cross-fitted conformal intervals were close to nominal across stations. In leave-one-site-out, 90% intervals covered 0.90–0.95 of temperature and salinity bins for the graph models, in both information settings (Table 9, Figure 11). Under seasonal shift, they failed for models whose errors grow outside the training season. Coverage was 0.25 (temperature) for climatology, 0.18 for gradient boosting, and 0.37 for the GAM, and 0.48 (temperature) and 0.40 (salinity) for the lagged graph model with constraints. Same-day IDW stayed near nominal (0.97 and 0.96), because its error does not depend on the training season. CRPS follows the accuracy ordering. For same-day temperature, it is 0.054 (IDW), 0.060 (graph with constraints), and 0.071 (gradient boosting), against 0.139 for climatology. Calibration on one season should not be trusted for another.

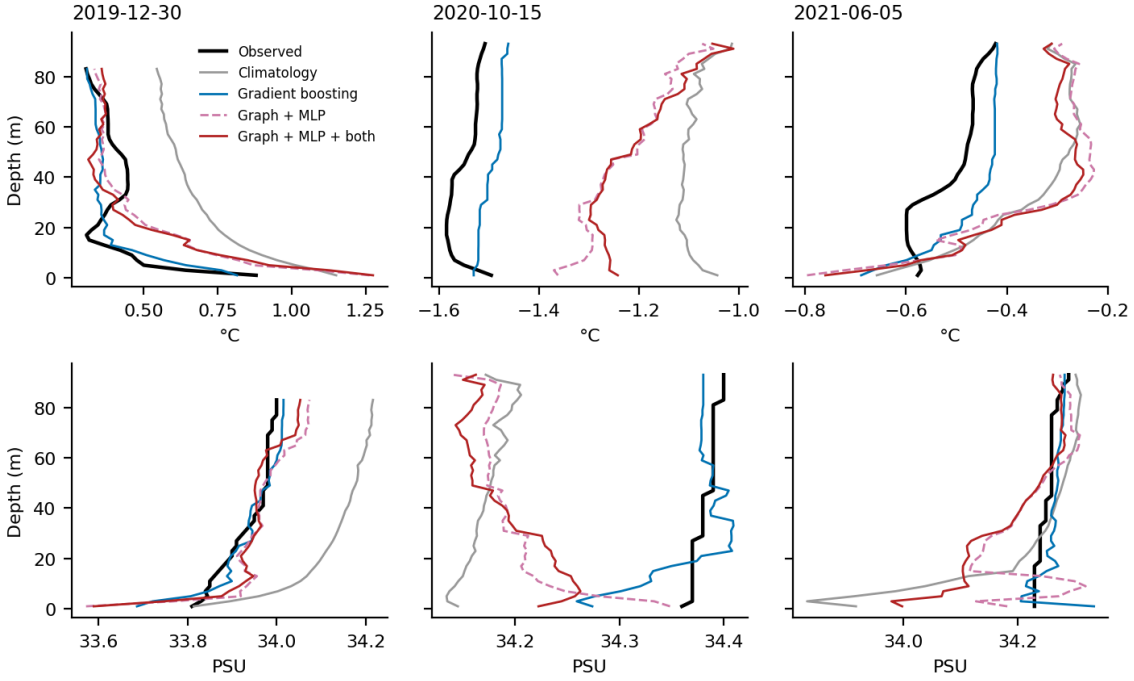


Figure 10. Example Profiles at Held-Out Station Lr7 (Leave-One-Site-Out, Lagged Inputs), Taken at the 25th, 50th, and 75th Percentiles of Its Profile List, Not Selected for Fit. The Graph Model Captures The Seasonal Level Better Than Climatology But Can Miss The Vertical Shape.

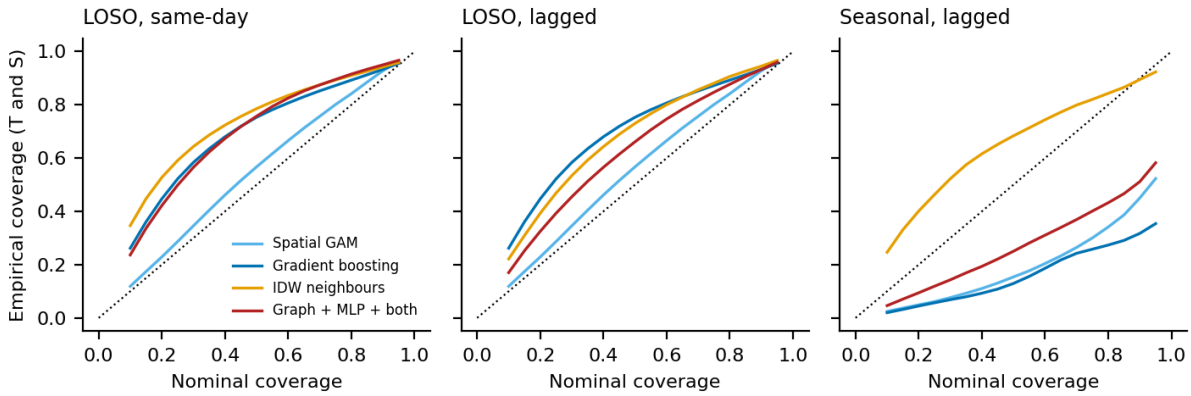


Figure 11. Reliability of Gaussian Predictive Distributions (Pooled Over Temperature and Salinity) for Leave-One-Site-Out and Seasonal Splits. Curves Above the Diagonal Are Under-Confident; Curves Below Are Over-Confident.

6.7. Sensitivity to the Flagged 2020 Episode

Including the 139-episode profiles gives 810 test profiles over 49 days instead of 693 over 40. The ordering of methods in the seasonal split does not change (Table 10). Same-day neighbour models still reach about 0.1 °C. Climatology salinity error rises from 0.348 to 0.523 PSU. The residual network with constraints is again no better than the base model for lagged salinity (0.541 against 0.423 PSU). The episode mainly enlarges salinity errors for every method. Its cause should be resolved before salinity is used for water-mass analysis.

Table 9. Probabilistic Skill and Calibration. Cover = Empirical Coverage of 90% Conformal Intervals; Crps Uses a Gaussian with Cross-Fitted Scale. † Campaign-Informed Baselines (Section 5.5).

Scheme	Model	CRPS T (°C)	CRPS S (PSU)	Cover T (90%)	Cover S (90%)	Profiles with inversion
LOSO, same-day	Climatology†	0.139	0.120	0.90	0.90	16%
LOSO, same-day	Spatial GAM (ridge splines) †	0.141	0.126	0.91	0.91	0%
LOSO, same-day	Gradient boosting†	0.071	0.077	0.91	0.90	15%
LOSO, same-day	IDW of neighbours	0.054	0.054	0.90	0.90	21%
LOSO, same-day	Graph + MLP	0.061	0.067	0.91	0.93	40%
LOSO, same-day	Graph + MLP + both	0.060	0.069	0.91	0.95	28%
LOSO, lagged†	Climatology†	0.139	0.120	0.90	0.90	16%
LOSO, lagged†	Spatial GAM (ridge splines) †	0.141	0.126	0.91	0.91	0%
LOSO, lagged†	Gradient boosting†	0.071	0.077	0.91	0.90	15%
LOSO, lagged†	IDW of neighbours	0.187	0.094	0.90	0.90	25%
LOSO, lagged†	Graph + MLP	0.101	0.083	0.91	0.91	36%
LOSO, lagged†	Graph + MLP + both	0.101	0.082	0.91	0.91	27%
Seasonal, same-day	Climatology	0.944	0.185	0.25	0.68	39%
Seasonal, same-day	Spatial GAM (ridge splines)	0.715	0.319	0.37	0.36	0%
Seasonal, same-day	Gradient boosting	0.690	0.314	0.18	0.21	34%
Seasonal, same-day	IDW of neighbours	0.044	0.053	0.97	0.96	27%
Seasonal, same-day	Graph + MLP	0.093	0.125	0.77	0.64	71%
Seasonal, same-day	Graph + MLP + both	0.068	0.127	0.89	0.74	60%
Seasonal, lagged	Climatology	0.944	0.185	0.25	0.68	39%
Seasonal, lagged	Spatial GAM (ridge splines)	0.715	0.319	0.37	0.36	0%
Seasonal, lagged	Gradient boosting	0.690	0.314	0.18	0.21	34%
Seasonal, lagged	IDW of neighbours	0.234	0.092	0.83	0.95	32%
Seasonal, lagged	Graph + MLP	0.475	0.291	0.45	0.34	55%
Seasonal, lagged	Graph + MLP + both	0.478	0.244	0.48	0.40	59%

Table 10. Seasonal-Split RMSE with the Flagged Episode Included (T in °C, S in PSU).

Model	Same-day T	Same-day S	Lagged T	Lagged S
Climatology	1.301	0.523	1.301	0.523
Spatial GAM (ridge splines)	1.005	0.668	1.005	0.668
Gradient boosting	0.893	0.802	0.893	0.802
IDW of neighbours	**0.106**	**0.315**	**0.426**	**0.383**
Graph base model	0.120	0.331	0.774	0.423
Graph + MLP + both	0.142	0.319	0.795	0.541

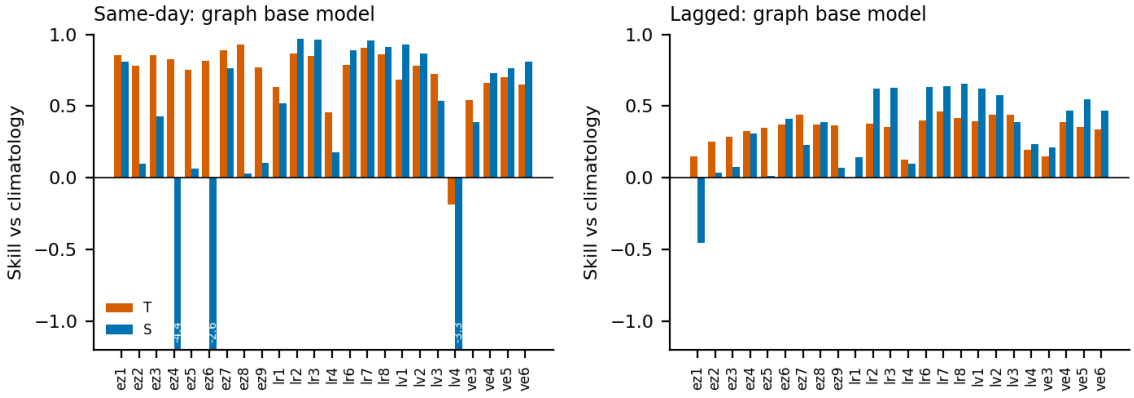


Figure 12. Station-Level Skill of the Graph Base Model Against Climatology (Leave-One-Site-Out). Values Below −1.2 Are Truncated and Labelled. Salinity Skill Is Strongly Negative at Ez4, Ez6 and Lv4.

6.8. Robustness to Constraint Weight, Tolerance, Seeds, and Reference Bandwidth

We tested four settings that the main results fixed in advance (Table 11, Figure 13). The constraint-weight sweep used 12 stations (every second station, 985 profiles) with lagged inputs and base parameters fixed at the fitted values.

Raising the constraint weight λ from 0 to 3 lowers the share of inverted profiles monotonically, from 37% to 18%, which is close to the base model's 20% on this subset and approaches the observed 12%. Temperature RMSE rises only from 0.194 to 0.198 °C, and salinity RMSE stays at 0.288–0.291 PSU. The constraint therefore works as a plausibility control with a small accuracy cost, and it does not add skill. Conclusions about plausibility also depend on the tolerance, as expected for noisy sensors. The constrained model is close to the observed inversion rate at $\tau = 0.05$ (5.7% against 7.5% observed) and above it at $\tau = 0.02$ and 0.01.

Three random seeds at $\lambda = 0.3$ give temperature RMSE of 0.191–0.195 °C, salinity RMSE of 0.288–0.291 PSU and 27–28% inverted profiles, so seed variation is small relative to the differences between model families. Climatology RMSE for temperature rises from 0.242 to 0.302 °C as the bandwidth increases from 5 to 25 days. The default of 8 days sits near the best value, so the baseline is not weak, and the lagged graph models (0.188–0.202 °C) and the same-day neighbour models still beat the best climatology (0.242 °C).

Table 11. Robustness to Constraint Weight and Seeds (Leave-One-Site-Out, Lagged Inputs, 12 Stations).

Setting (12 stations, 985 profiles, lagged inputs)	RMSE T (°C)	RMSE S (PSU)	Profiles with inversion	Points outside hull
Climatology	0.254	0.325	18%	n/a
Graph base model (no MLP)	0.207	0.297	20%	0.1%
MLP, λ = 0	0.194	0.290	37%	0.6%
MLP, λ = 0.1	0.194	0.289	32%	0.6%
MLP, λ = 0.3 (seed 0)	0.195	0.288	28%	0.3%
MLP, λ = 1	0.195	0.290	25%	0.2%
MLP, λ = 3	0.198	0.291	18%	0.1%
MLP, λ = 0.3, three seeds (range)	0.191–0.195	0.288–0.291	27–28%	0.3%

7. DISCUSSION

7.1. Expected and Observed Results

Table 12 compares each hypothesis with the outcome. The same-day and lagged comparisons (H1 and H2), the seasonal hull analysis (H4), and the graph-structure test (H5) were exploratory. Results support H1, H2, H3b (the stability constraint improves plausibility), and H4 (the summer hull does not transfer to winter). The original expectation that constraints improve accuracy (H3) was not supported. The formal edge ablation did not support H5 for lagged inputs; the same-day comparison with IDW was not a matched edge ablation, so H5 remains untested for same-day inputs.

7.2. Why Neighbours Dominate

When same-day neighbours exist, a station is largely determined by its neighbours, and a simple distance-based weighting reproduces it well. This agrees with the physical expectation of a spatially coherent anomaly (Section 3.1). Without same-day data, the useful information is the bay-wide anomaly relative to the seasonal state, which a lagged graph model captures and a covariate model cannot, since a covariate model has no access to the current state. We did not measure the drivers, such as meltwater discharge, wind, or air temperature, so we cannot say which process creates the shared anomaly. The weak vertical structure and the strong seasonal cycle in temperature suggest that surface exchange and seasonal cooling matter, but this is an inference, not a result.

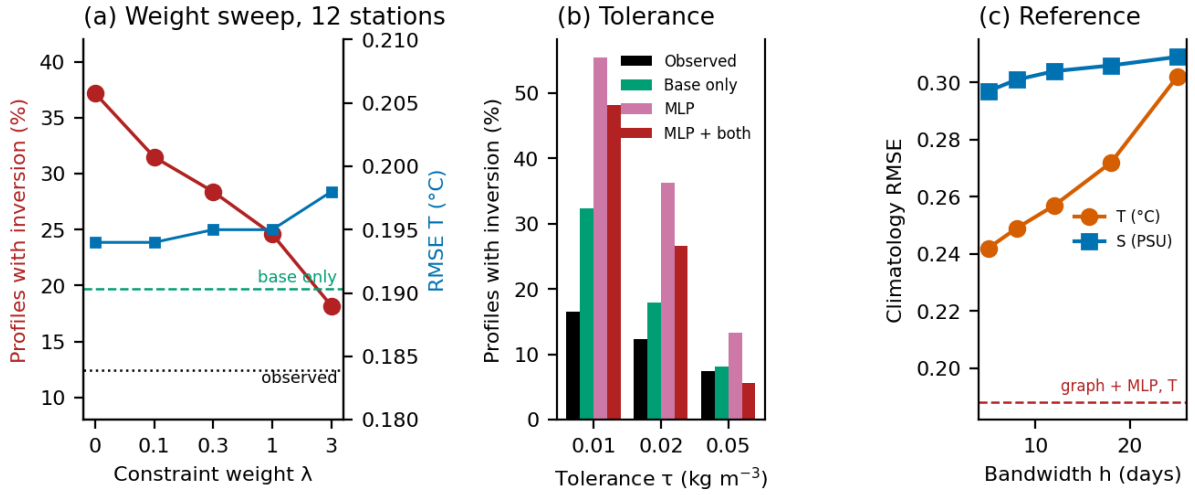


Figure 13. Robustness. (A) Constraint Weight Against Inversion Share and Temperature RMSE. (B) Inversion Share at Three Tolerances. (C) Climatology RMSE Against Kernel Bandwidth, with the Graph-Plus-MLP Temperature RMSE for Reference.

Table 12. Hypotheses, Expectations, and Outcomes.

Hypothesis	Expected result	Observed result	Verdict
H1 (exploratory) Same-day neighbours carry most transferable information	Neighbour-based models beat covariate-only models	IDW 0.110 °C; boosting 0.138 °C; GAM 0.251 °C (campaign-informed)	Supported
H2 (exploratory) Lagged neighbour anomalies beat covariate-only models on unobserved days	Positive skill for the lagged graph; none for covariate models	Graph skill +0.45 (T), +0.32 (S); boosting −0.14; GAM +0.23	Supported
H3 Constraints improve accuracy where data are sparse (original project expectation)	Lower RMSE with constraints	RMSE unchanged or of mixed sign (T 0.188 to 0.189 °C); small cost at high weight	Not supported
H3b Stability constraints improve plausibility	Fewer density inversions	Inverted profiles fall from 37% to 18% as constraint weight rises on the 12-station subset	Supported
H4 A hull learned in one season fails in another (exploratory)	Many winter points outside the summer hull	56% of same-day IDW predictions for June–November fall outside	Supported
H5 Learned edges beat distance edges (exploratory)	Lower RMSE with learned metric	No difference in lagged edge ablation; same-day edge ablation not run	Not supported for lagged inputs; untested for same-day inputs

7.3. Why Constraints Did Not Help Accuracy

A constraint derived from the training data cannot add information the data lack. The base model averages real profiles, which are mostly stable, so the stability constraint has little to correct. The residual network introduces inversions (36% of profiles, against 18% for the base model), and the penalty removes only some of them (27%). This is consistent with the penalty improving plausibility without improving accuracy. The hull penalty is inactive within a season because the training hull is wide. Across seasons, it is mis specified, since about half of the new season's observations lie outside it. Our seasonal runs combine both penalties, so we cannot separate the hull's own effect. These findings are consistent with constraints that help plausibility reliably and accuracy only when they carry information beyond the data, although we did not test that general claim. The weight sweep (Section 6.8) shows the trade-off explicitly.

7.4. Leakage and Evaluation Practice

The leak in station-withheld splits is not specific to this bay. Whenever observations come in campaigns, a covariate that is nearly unique to a campaign lets flexible models recover the campaign mean from the training set. Day of year and any continuous time stamp can do this, and spatial cross-validation does nothing to prevent it. Buffered temporal blocking, as in our leave-one-day-out test, or models that cannot resolve time finely, address the problem. We recommend reporting both an interpolation setting and a buffered transfer setting.

7.5. Relation to Earlier Work

Our plausibility result is consistent with the use of static-stability adjustments in profile reconstruction (Pauthenet et al., 2022). Our setting differs in scale from basin-wide graph forecasts (Holmberg et al., 2026). In a bay a few kilometres wide, a uniform or distance-based graph did as well as the learned one in our tests, and we make no claim about larger domains. The finding that transfers to unseen conditions needs informative inputs is consistent with glaciological transfer studies (Maciel-Seidman et al., 2026), although those address different variables and cannot be compared numerically. Related produced-water studies have applied both CNN–LSTM–attention forecasting (Howard & Apeli, 2026) and hybrid neural-network approaches (Howard & Howard, 2026). Their monitored water-quality series differ from our hydrographic profiles, so we make no numerical comparison. In our tests, a seasonal reference plus neighbour anomalies carried most of the skill (Sections 6.3 and 6.8), and we did not test deep temporal models.

7.6. Implications for Monitoring And Modelling

The results suggest two practical points. Recent profiles at neighbouring stations carry most of the transferable information, so a campaign that samples a subset of stations on each visit retains much of the bay-wide signal, at least for temperature. We did not test monitoring-network design directly, so this is a hypothesis for future work. Constraints can also serve as quality checks and plausibility controls in gap-filled products, even when they do not improve accuracy. Modellers who evaluate transfer should use buffered validation and should report whether baselines can see same-campaign data.

7.7. Limitations

Several limitations apply. The graph model is a small NumPy implementation of attention message passing, not a deep GNN, so deeper architectures or dynamic edges may behave differently. We did not evaluate zero-shot foundation models; the GAM is a custom ridge-spline model. ERA5 reanalysis meteorology and glacier-terminus positions were not available for this analysis, so Task B uses lagged neighbours and season only. The hull lives in T–S space only, and the main runs fixed λ and d_0. The climatology bandwidth was chosen on pooled leave-one-site-out error. Base-model parameters were fixed at the full-training fit in the cross-fit and at LOSO medians in the leave-one-day-out test, which leaks a few parameters from the withheld day into LODO.

The seasonal test has 693 profiles on 40 days, and its winter part is thinner still. One bay, 24 stations, and one sensor type limit generality. The records are irregular campaign series, not continuous ones. We test gap-filling and transfer, not forecasting. Our only own-site temporal baseline is persistence of the previous cast, and we did not run dedicated time-series models such as seasonal ARIMA. The 2020 salinity episode is also unresolved, seven shallow stations are excluded, and sensor accuracy limits any buoyancy-frequency analysis.

7.8. Future Research

Several extensions follow. Forcing data (ERA5, glacier-terminus position) could be added to test whether they let Task B predict the bay-wide anomaly instead of inheriting it from neighbours. Dedicated time-series baselines (seasonal naive, SARIMA, Holt–Winters) could be added for stations with denser records, and the task could be extended from gap-filling to forecasting. The decay-based temporal encoder could be replaced with a recurrent or attention encoder, such as the CNN–LSTM–attention design of Howard and Apeli (2026), and compared with hybrid neural-network approaches used for produced-water quality forecasting (Howard & Howard, 2026) under the buffered protocol. The models should also be run on other bays and fjords, and with deeper graph architectures and zero-shot foundation models, under the same buffered protocol. Monitoring-network design can be tested directly by dropping stations or campaigns and measuring the loss of skill. The 2020 salinity episode should be resolved with the data providers, with TEOS-10 and conductivity-calibrated data used for water-mass analysis. The T–S hull could also be replaced with end-member mixing models that include latent heat, and constraint weights could be tuned on held-out stations.

8. CONCLUSIONS

In Admiralty Bay, transfer of profile structure to unobserved stations is largely a matter of neighbour information. Same-day neighbours reduce temperature MSE by about 80% relative to climatology. Lagged neighbours with a seasonal reference reduce temperature MSE by 45% and salinity MSE by 32% on unobserved days. A covariate model that looks better under station-withheld validation loses its advantage once same-day training data are removed. In the full leave-one-site-out ablation, the stability penalty reduces inversions relative to the unconstrained residual network (36% to 27%), but the inversion rate remains above that of the graph base model (18%), at a small accuracy cost. The 12-station weight sweep shows that a stronger penalty can reduce the inversion rate to 18%, close to the base model's 20%. The hull penalty has little effect in leave-one-site-out evaluation; neither constraint consistently improves accuracy, calibration, or seasonal transfer. Conformal coverage is near nominal across stations but deteriorates under seasonal shift for several models; same-day IDW remains near nominal.

The contributions are a leakage-aware, buffered evaluation protocol for campaign-based ocean data, a transparent station-graph baseline, and a quality-controlled profile table. The practical implication is that recent profiles at neighbouring stations are among the most valuable measurements to collect, while physical constraints are best treated as plausibility controls. Section 7.8 outlines next steps. The analysis pipeline and processed profile table are available from the corresponding author so that other stations, forcing data, and deeper models can be tested under the same protocol.

DECLARATIONS

Data availability. The YSI EXO sonde measurements from Admiralty Bay, collected from 15 December 2018 to 28 January 2022, are publicly available in PANGAEA (Osińska et al., 2023b). The quality-controlled profile table generated during this study is available from the corresponding author on reasonable request.

Ethics. The study uses published environmental measurements and involves no human participants or animals.

Funding. The authors received no funds, grants, or other support during the preparation of this manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Author contributions. Conceptualization: Howard Chioma Chinagorim; Methodology: Howard Chioma Chinagorim; Validation: Love Cherukei Nnoka; Formal analysis: Howard Chioma Chinagorim; Investigation: Love Cherukei Nnoka; Data curation: Howard Chioma Chinagorim; Writing–original draft: Howard Chioma Chinagorim; Writing–review and editing: Love Cherukei Nnoka; Supervision: Love Cherukei Nnoka. All authors approved the final manuscript.

Acknowledgements. The authors thank the researchers who collected and archived the Admiralty Bay sonde data in PANGAEA (Osińska et al., 2023b).

References

[1] Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494–591. https://doi.org/10.1561/2200000101

[2] Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2021). Predictive inference with the jackknife+. *The Annals of Statistics, 49*(1), 486–507. https://doi.org/10.1214/20-AOS1965

[3] Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G. E., Vaswani, A., Allen, K., Nash, C., Langston, V., . . . Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1806.01261

[4] Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Physical Review Letters, 126*, 098302. https://doi.org/10.1103/PhysRevLett.126.098302

[5] Bolibar, J., Rabatel, A., Gouttevin, I., Galiez, C., Condom, T., & Sauquet, E. (2020). Deep learning applied to glacier evolution modelling. *The Cryosphere, 14*, 565–584. https://doi.org/10.5194/tc-14-565-2020

[6] Che, Z., Purushotham, S., Cho, K., Sontag, D., & Liu, Y. (2018). Recurrent neural networks for multivariate time series with missing values. *Scientific Reports, 8*, Article 6085. https://doi.org/10.1038/s41598-018-24271-9

[7] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[8] Gade, H. G. (1979). Melting of ice in sea water: A primitive model with application to the Antarctic ice shelf and icebergs. *Journal of Physical Oceanography, 9*(1), 189–198. https://doi.org/10.1175/1520-0485(1979)009<0189:MOIISW>2.0.CO;2

[9] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. https://doi.org/10.1198/016214506000001437

[10] Holmberg, D., Oskarsson, J., Wikingsson, E., Lindsten, F., & Roos, T. (2026). *Njord: A probabilistic graph neural network for ensemble ocean forecasting* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2605.15470

[11] Howard, C. C., & Apeli, A. (2026). Hybrid CNN–LSTM–attention forecasting of toxic contaminants in produced water: Evidence from Niger Delta flow stations. *World Scientific News, 219*, 21–39. https://doi.org/10.65770/ZIWD5587

[12] Howard, C. C., & Howard, I. C. (2026). Hybrid neural network approaches to produced water quality forecasting and management in Nigerian oil fields. *International Journal of Mathematics, Statistics, and Computer Science, 4*, 527–543. https://doi.org/10.59543/7fjxt923

[13] Intergovernmental Oceanographic Commission, Scientific Committee on Oceanic Research, & International Association for the Physical Sciences of the Oceans. (2010). *The international thermodynamic equation of seawater–2010: Calculation and use of thermodynamic properties* (IOC Manuals and Guides No. 56). UNESCO. https://doi.org/10.25607/OBP-1338

[14] Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics, 3*, 422–440. https://doi.org/10.1038/s42254-021-00314-5

[15] Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., & Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science, 382*(6677), 1416–1421. https://doi.org/10.1126/science.adi2336

[16] Maciel-Seidman, M. L., Ryan, J. C., Esenther, S. E., Rennermalm, Å. K., Smith, L. C., & Fettweis, X. (2026). *A transfer learning approach for simulation of Greenland Ice Sheet meltwater runoff* [Preprint]. ESS Open Archive. https://doi.org/10.22541/essoar.15008734/v1

[17] Mas e Braga, M., Cowton, T., Slater, D., Inall, M., Johnstone, E., & Fraser, N. (2025). Controls on fjord temperature throughout Greenland in a reduced-physics model. *Geophysical Research Letters, 52*(17), e2025GL116902. https://doi.org/10.1029/2025GL116902

[18] Meyer, H., & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution, 12*(9), 1620–1633. https://doi.org/10.1111/2041-210X.13650

[19] Millero, F. J., & Poisson, A. (1981). International one-atmosphere equation of state of seawater. *Deep-Sea Research Part A: Oceanographic Research Papers, 28*(6), 625–629. https://doi.org/10.1016/0198-0149(81)90122-9

[20] Osińska, M., Wójcik-Długoborska, K. A., & Bialik, R. J. (2023a). Annual hydrographic variability in Antarctic coastal waters infused with glacial inflow. *Earth System Science Data, 15*(2), 607–616. https://doi.org/10.5194/essd-15-607-2023

[21] Osińska, M., Wójcik-Długoborska, K. A., & Bialik, R. J. (2023b). *Water conductivity, salinity, temperature, turbidity, pH, fluorescent dissolved organic matter (fDOM), optical dissolved oxygen (ODO), chlorophyll a and phycoerythrin measurements in Admiralty Bay, King George Island, from Dec 2018 to Jan 2022* [Data set]. PANGAEA. https://doi.org/10.1594/PANGAEA.947909

[22] Pauthenet, E., Bachelot, L., Balem, K., Maze, G., Tréguier, A.-M., Roquet, F., Fablet, R., & Tandeo, P. (2022). Four-dimensional temperature, salinity and mixed-layer depth in the Gulf Stream, reconstructed from remote-sensing and in situ observations with neural networks. *Ocean Science, 18*(4), 1221–1244. https://doi.org/10.5194/os-18-1221-2022

[23] Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography, 40*(8), 913–929. https://doi.org/10.1111/ecog.02881

[24] Souza, A. C. de, Martins, C. C., Combi, T., & Lourenço, R. A. (2025). Spatial and temporal dynamics of surface water physicochemical parameters and the influence of atmospheric conditions in Admiralty Bay, Antarctic Peninsula. *Antarctic Science, 37*, 309–319. https://doi.org/10.1017/S0954102025000148

[25] Straneo, F., & Cenedese, C. (2015). The dynamics of Greenland's glacial fjords and their role in climate. *Annual Review of Marine Science, 7*, 89–112. https://doi.org/10.1146/annurev-marine-010814-015714

[26] Tibshirani, R. J., Barber, R. F., Candès, E. J., & Ramdas, A. (2019). *Conformal prediction under covariate shift* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1904.06019

[27] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). *Graph attention networks* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1710.10903

[28] Verjans, V., & Robel, A. A. (2024). Accelerating subglacial hydrology for ice sheet models with deep learning methods. *Geophysical Research Letters, 51*(2), e2023GL105281. https://doi.org/10.1029/2023GL105281

[29] Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. Springer. https://doi.org/10.1007/b106715

[30] Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279