https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

A Deep Learning Approach for Lake Ice Cover Forecasting

Samuel J. Johnston[1], Justin Murfitt[1], Claude R. Duguay[1,2],

[1]H2O Geomatics Inc., Kitchener, N2G 4X8, Canada
[2]Department of Geography and Environmental Management, University of Waterloo, Waterloo, N2L 3G1, Canada

Correspondence to: Samuel J. Johnston (samuel.johnston@h2ogeomatics.com)

Abstract. Lakes cover a significant proportion of the high-latitude landscape and exert a strong influence on local weather and climate. Their seasonal lake ice cover (LIC) further impacts lake-atmosphere interactions, while also providing key socioeconomic services ~~for northern communities~~. Climate change is impacting LIC and its thickness, two thematic products of Lakes as an Essential Climate Variable (ECV). Accurate prediction of LIC improves numerical weather prediction (e.g. lake-effect snowfall and thermal moderation) and is crucial for anticipating the impacts of climate change in lake-rich regions of the Northern Hemisphere.

This paper introduces ~~LIF-DL (Lake Ice Forecasting using Deep Learning)~~, a novel data-driven model for forecasting LIC extent across entire lake surfaces. LIF-DL uses Spatial-Temporal Transformer Networks (STTN) to capture relationships between lake conditions (ice and open water), lake depth and atmospheric forcings. The study focuses on five large Canadian lakes with pronounced ice phenology: Great Slave Lake, Great Bear Lake, Lake Winnipeg, Lake Athabasca, and Reindeer Lake. Data sources included ice cover observations from the Interactive Multi-Sensor Snow and Ice Monitoring System (IMS), atmospheric reanalysis from the European Centre for Medium-Range Weather Forecasts (ECMWF) 5th generation of European ReAnalysis (ERA5 and ERA5-Land), and Canadian Ice Service (CIS) records for external validation. To benchmark the proposed approach against a traditional physics-based model, the widely used Freshwater Lake (FLake) model embedded in ERA5 and ERA5-Land was employed. LIF-DL was trained to produce one-week forecasts using data from 2004–2017 and then deployed auto-regressively to predict ice cover during the 2018–2022 holdout period. Forecasts were evaluated against IMS and CIS observations and compared with those from FLake.

Across all evaluations—phenology timing, ice cover fraction, and spatial patterning—LIF-DL consistently outperformed FLake. Freeze-up and break-up events were predicted within 3–9 days of observations (versus 5–22 days for FLake), and ice cover fraction (range 0–1) root mean squared errors were reduced (0.06–0.16 versus 0.1–0.2). A key advantage of LIF-DL was its capacity to represent spatial dependencies across lake surfaces, producing coherent freeze-up and break-up dynamics and realistic spatial clustering of early and late ice timing compared to the fragmented patterns of FLake. These improvements reduced extreme timing biases—from as much as 30 days to only 4–6 days—particularly for large, deep lakes. Variable importance analysis indicated sensitivity to physically meaningful drivers, including air temperature, accumulated degree days,

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

solar radiation, and lake depth, suggesting that LIF-DL learned relevant physical processes rather than statistical artifacts. Finally, the model maintained stable performance when iteratively forecasting over a four-year period, demonstrating robustness under varying atmospheric conditions.

The demonstrated accuracy, robustness, and physical interpretability of LIF-DL highlight the potential of deep learning f advancing lake ice modelling. Future research should focus on integrating physical constraints to develop hybrid physics-machine learning frameworks, improving model interpretability, and expanding to new predictive variables such as ice thickness and snow cover. Leveraging emerging high-resolution satellite datasets will further enhance spatial fidelity and enable application to smaller lakes. Ultimately, spatiotemporal deep learning represents a transformative step toward next-generation, spatially resolved lake ice forecasts that can improve weather and climate prediction, inform northern transportation planning, and support climate change adaptation in lake-rich regions of the Northern Hemisphere.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

1 Introduction

Freshwater lakes cover a significant proportion of the Earth's surface across the Northern Hemisphere. For example, within the Arctic and Subarctic regions of North America, lakes are estimated to represent between 15% and 40% of the land area, depending on the location (Brown and Duguay, 2010, 2011). This large spatial coverage, coupled with the capacity of lakes to transmit and absorb energy, makes them a vital component of the cryosphere (Brown and Duguay, 2010; Rouse et al., 2005). Lakes are known to impact regional energy and water balances, and the presence (or absence) of ice cover during winter months further affects these lake-climate interactions (Rouse et al., 2005, 2008).

Lake ice phenology, defined as the seasonal timing (freeze-up/break-up) and duration of lake ice cover, is sensitive to climate variability, and climate change is stimulating later freeze-up (ice formation) and earlier break-up (ice melt) dates, effectively shortening the duration of ice cover (Brown and Duguay, 2011; Dauginis and Brown, 2021; Duguay et al., 2006). Shorter lake ice seasons will significantly enhance energy and moisture exchange over lakes, impacting regional weather patterns and events (Ménard et al., 2002). Shortening ice cover seasons also present substantial socio-economic challenges for northern communities, who rely on ice cover for transportation (ice roads), recreation, and ecosystem health (Derksen et al., 2012). Therefore, lake ice monitoring and modelling have become essential tools for understanding how lakes respond to climate change.

Long-term lake ice records, which capture the spatial and temporal responses of lakes to climate change, are crucial for assessing the impacts of climate change (Brown and Duguay, 2010, 2011). Both lake ice thickness and cover are included as thematic variables of "Lakes" as an Essential Climate Variable (ECV), due to their importance as climate change indicators (World Meteorological Organization, 2022). Over the last decade, the monitoring of lake ice has shifted from traditional in-situ observation networks to a greater reliance on satellite remote sensing, facilitating the mapping of ice cover extent for a significantly larger number of lakes globally (Murfitt and Duguay, 2021). Remote sensing methods represent a significant improvement in both efficiency and information quantity, enabling the production of daily LIC maps across entire lake surfaces (Carrea et al., 2023).

The need to predict ice cover under different climate regimes, as well as to simulate lake-atmosphere interactions, has led to the development of physically based lake models, including the Canadian Lake Ice Model (CLIMo, Duguay et al., 2003) and the Freshwater Lake model (FLake; Mironov, 2008). CLIMo predicts the timing of freeze-up and break-up as well as ice thickness and composition. FLake models lakes more wholly, predicting the vertical temperature structure and mixing conditions in lakes using a two-layer parametric representation. It further includes a module for simulating the formation and melting of ice (Mironov, 2008). FLake is ~~intended for use~~ as a lake parameterization module in numerical weather prediction,

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

and multiple studies have demonstrated that its integration enhances the accuracy of regional and global forecasts, particularly in areas with substantial lake cover (Balsamo et al., 2012; Mironov et al., 2012; Mironov et al., 2010).

A primary drawback to the physical approach of lake ice modelling is its reliance on one-dimensional equations, which assume that lake conditions are uniform horizontally and only vary with depth (Brown and Duguay, 2012; Pour et al., 2012). This presents a challenge when modelling large lakes, which can display high spatial heterogeneity in their ice cover (Dauginis and Brown, 2021). One approach to addressing this has been to simulate multiple points across a gridded lake surface (Balsamo et al., 2012; Kheyrollah Pour et al., 2024). However, horizontal mixing mechanisms coupled with spatial morphometric features such as bays, inlets, outlets and open lake areas can have a significant influence on the dynamics of freeze-up and break-up across a single lake (Brown and Duguay, 2010; Pour et al., 2012). The gridded application of one-dimensional lake models is therefore limited in its capacity to model the spatial patterns of lake ice.

The recent growth in lake ice cover records from satellite observations, coupled with the development of global atmospheric reanalysis datasets such as ERA5 (Hersbach et al., 2020), presents a new opportunity for addressing these limitations. Data-driven, deep learning approaches are being increasingly applied for environmental modelling, due to the growing abundance of environmental monitoring data (Haupt et al., 2022). Additionally, recent developments in deep learning have produced modelling frameworks capable of leveraging high-dimensional spatial-temporal datasets, such as the Spatial-Temporal Transformer Network (STTN, Zeng et al., 2020), which was developed for video inpainting. In the case of lakes, data-driven models have already shown success in modelling lake-temperature profiles (Jia et al., 2021) and in automating the classification of ice cover from satellite imagery (Tom et al., 2020; Wu et al., 2021). This establishes the potential for applying a spatial-temporal deep learning framework to the task of forecasting lake ice cover extent, towards a lake ice model which considers spatial interactions across lake surfaces.

The objective of this study was to develop a spatial-temporal lake ice cover forecasting model using deep learning and benchmark its performance to assess the potential of such approaches within the field of lake ice modelling. To achieve this, we developed the ~~LIF-DL (Lake Ice Forecasting with Deep Learning)~~, which was adapted from the Spatial-Temporal Transformer Network (STTN)~~, first developed by Zeng et al. (2020) for video inpainting~~. The LIF-DL utilizes gridded atmospheric forcing inputs to produce short to long-term forecasts of lake ice cover extent across entire lake surfaces and represents a first-of-its-kind approach to lake ice cover modelling.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

2 Materials and Methods

2.1 Data Sources

2.1.1 Ice Observation Data

Gridded ice cover observations were sourced from the Interactive Multisensor Snow and Ice Monitoring System (IMS) produced by the National Snow and Ice Data Center (NSIDC) (U.S. National Ice Center, 2004). The product is cloudless and includes daily estimates of snow and ice cover for the Northern Hemisphere, produced by ice/snow analysts who combine a variety of data products, including multiple satellite imagery sources and in situ data (U.S. National Ice Center, 2004). It is available in a polar stereographic projection (EPSG:3411), centred at the north pole, and at three grid spacings (24 km, 4 km, 1 km). The temporal availability of the data is shorter for finer spatial resolutions. To balance data availability with grid size, the 4 km NetCDF product was selected, spanning the temporal range from February 24th, 2004, to December 31st, 2022. At times, IMS applies persistence to ice boundaries when insufficient sensor data is available, which can result in ice edges appearing unchanged for a day or more, even though they may be evolving (U.S. National Ice Center, 2004). Regardless, IMS has been extensively validated and used in numerous lake ice studies, both for assessing trends/variability and for validating models (Brown and Duguay, 2012; Dauginis and Brown, 2021).

The Canadian Ice Service (CIS) weekly ice cover product was ~~obtained~~ for independent validation of LIF-DL predictions of ice cover. CIS ice cover records are produced by ice analysts via visual interpretation of synthetic aperture radar (SAR) and optical satellite imagery (Hoekstra et al., 2020). Records are available for individual lakes (or lake sections) and are reported weekly as an integer value ~~from 0 to 10, where 0 represents no ice cover and 10 represents full ice cover.~~

2.1.2 Gridded Lake Surface Forcing Data

The formation, growth, and decay of lake ice cover are fundamentally governed by a surplus or deficit in the energy ~~balance~~ of the ~~ice~~ ~~cover~~, determined by heat exchange with the atmosphere, heat stored in the water, and heat input from inflows (Williams, 1965). Some of the most important atmospheric determinants of ice cover are air temperature, solar radiation, snowfall and wind, with cloud cover liquid precipitatio so playing a role (Williams, 1965). Another significant contributing factor is lake morphometry (area, depth and volume), which determines the lake's energy storage properties (Jeffries and Morris, 2007; Korhonen, 2006). Many of these variables are used in the parameterization of physically-based lake ice models, most commonly air temperature, solar radiation, lake depth and snow depth, due to their strong, well-established influence on ice phenology and thickness (Duguay et al., 2003; Mironov, 2008; Pour et al., 2012). Relative humidity, wind speed, cloud cover and precipitation may also be parameterized, such as in CLIMo (Duguay et al., 2003).

To address seasonality in ice cover and heat storage, we introduced two additional variables, Accumulated Freezing Degree Days (AFDD) and Accumulated Thawing Degree Days (ATDD hese variables have been shown to relate well to ice

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

thickness and heat storage (Murfitt et al., 2018) and have been used in other ice modelling work (Arp et al., 2020). They are defined as the sum of temperature low (for AFDD) or above (for ATDD) zero degrees Celsius during their respective seasonal period (starting August 1st for AFDD, and February 1st for ATDD). These derivatives are designed to provide seasonal context to the model, and in conjunction with air temperature (to capture daily variations), were found to improve the model's timing accuracy of break-up and freeze-up during early prototyping.

Atmospheric forcing variables were drawn from the European Centre for Medium Range Weather Forecasting (ECMWF) Reanalysis v5 (ERA5) (Hersbach et al., 2020), and ERA5-Land (Muñoz-Sabater et al., 2021). ERA5 is a global atmospheric reanalysis dataset comprising decades of data on a large number of atmospheric, land and oceanic climate variables at roughly 31km resolution (Hersbach et al., 2020). To improve spatial resolution further, ERA5-Land provides a replay of the ERA5 Earth surface variables at approximately 9 km grid spacing (Muñoz-Sabater et al., 2021).

Lake depth was taken from the Global Lake DataBase (GLDB) (Toptunova et al., 2019). The database was produced using several data sources, including 14,960 lakes within situ data, indirect depth estimates, global lake cover and digitized bathymetry, and is available in ~1 km resolution. GLDB was developed for use by numerical weather prediction models and has already been adopted by several large global weather centers (e.g. ECMWF in the production of ERA5) and limited-area modelling consortia (e.g. HIRLAM and COSMO), for research and operational purposes (Toptunova et al., 2019).

Where available, the desired atmospheric forcing variables were drawn from the ERA5-Land, otherwise ERA5 was used. AFDD, ATDD and relative humidity are not natively available in ERA5-Land or ERA5 and so were calculated from the available variables. Snowfall is not currently available over lakes within ERA5 and was thereby excluded from this study. Finally, the hourly ERA5 variables were aggregated into daily samples using an average or sum operation. Table 1 lists each of the gridded predictor variables, their respective source(s) and the initial preprocessing steps applied.

2.1.3 FLake Data

FLake is a bulk model capable of predicting the vertical temperature structure and mixing conditions in lakes of various depths on time scales ranging from a few hours to several years (Mironov, 2008; Mironov et al., 2010). The model employs a similar set of predictors as the data-driven approach, including 2-m air temperature, wind speed, relative humidity and lake depth, while also simulating ice thickness and lake state. FLake is well-established within the lake modelling community and has been incorporated into numerical weather prediction systems, including the Integrated Forecasting System. It is also the lake parametrization used for producing ERA5 lake variables (ECMWF, 2023; Muñoz-Sabater et al., 2021). Because of this, FLake predictions of lake ice thickness were drawn directly from ERA5-Land and aggregated to daily averages. This approach ensured that FLake predictions were produced using equivalent forcing values to the LIF-DL, which supports ~~their~~ comparison ~~during evaluation~~.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

2.2 Study Lakes

Five study lakes were used in this analysis, namely Great Bear Lake, Great Slave Lake, Lake Athabasca, Reindeer Lake, and Lake Winnipeg (Figure 1). These represent some of the largest lakes within Canada, and globally, while also exhibiting strong seasonal ice cover patterns, transitioning between no ice and complete ice cover annually. Furthermore, this set of lakes encompasses a range of lake depths, surface areas and latitudinal positions, all of which result in varying durations and timings of ice phenology events. Large lakes were used for this study because their spatial patterns of ice formation and melt are the most pronounced and are resolvable at medium resolutions (4 km of IMS). This then allowed for the spatial accuracy of the LIF-DL ice cover predictions to be evaluated.

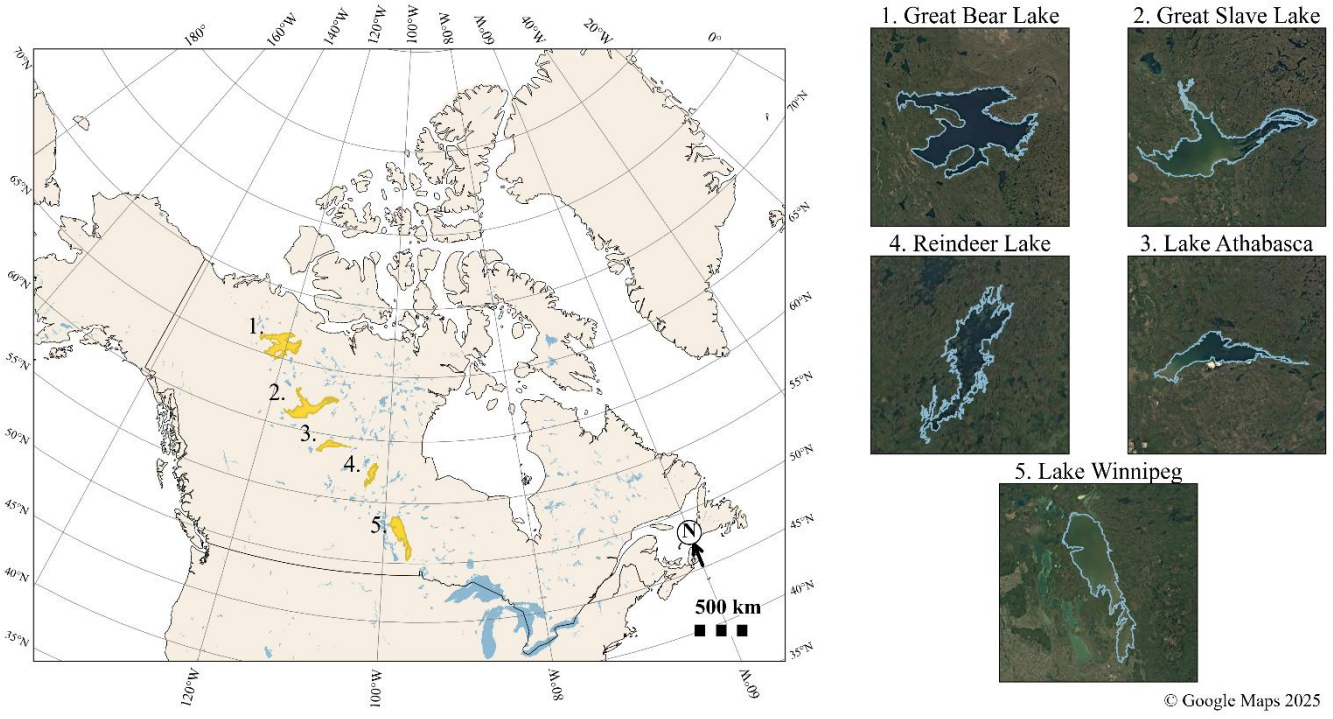

Figure 1. Study lakes and their respective locations within Canada. Right panels show corresponding satellite imagery: © Google Maps 2025.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Table 1. The forcing variables used, including their respective source(s) and any preprocessing steps applied.

Variable Name	Units	Source(s)	Preprocessing Steps
Temperature at 2 m	°C	ERA5-Land: Temperature at 2m	Daily average
Surface solar radiation downwards	Jm^{-2}	ERA5-Land: Surface Solar Radiation Downwards	Daily total
Relative humidity at 2 m	fraction	ERA5-Land: Temperature at 2m and Dewpoint Temperature at 2m	Calculate Relative Humidity using equation 7.98 from IFS Documentation CY48R1 (ECMWF, 2023, p.134–135), then take daily average
Wind speed at 10 m	ms^{-1}	ERA5-Land: U and V components of wind at 10 m	Calculate magnitude from components then take daily average
Total precipitation	m	ERA5-Land: Total Precipitation	Daily total
Total Cloud Cover	~~fraction~~	ERA5: Total Cloud Cover	Daily average
Accumulated freezing degree days (AFDD)	°C	ERA5-Land: Temperature at 2m	Accumulated mean air temperatures below 0°C starting February 1st.
Accumulated thawing degree days (ATDD)	°C	ERA5-Land: Temperature at 2m	Accumulated mean air temperatures above 0°C starting August 1st.
Lake depth	m	Global Lake Database: mean depth	

2.3 Data Preprocessing

We utilized multiple data preprocessing steps to prepare the data for tuning and training the LIF-DL. Bounding boxes were created by roughly centring 128x128 pixels (512 km by 512 km) squares around each of the study lakes in the IMS dataset. Then, lake masks were produced for each of the study areas by manually removing surrounding water bodies from the bounding region, leaving only the target lake. The north arm of Great Slave Lake was removed due to a significant under-reporting of lake depth in the GLDB. The deepest reported lake depth for the North Arm in the GLDB is roughly 40 m, but it is known to be over 200 m in some areas (Rühland et al., 2023). The exclusion of the North Arm in other lake-ice modelling studies is also common due to the limited availability of accurate bathymetry in this section of the lake (Ménard et al., 2002; Rouse et al., 2008).

Forcing variables (Table 1) were harmonized with the IMS grid (EPSG:3411, 4 km) to prevent the introduction of error into ice cover observations. Reprojection was done using a nearest neighbour interpolation, after which the lake masks were applied to remove surrounding water bodies. Figure 2 shows the result of this preprocessing chain as applied to the lake depth data

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

from the GLDB for each study lake. IMS ice cover observations were converted to a one-hot encoding format, using 3 classes: 0: masked, 1: water, 2: ice.

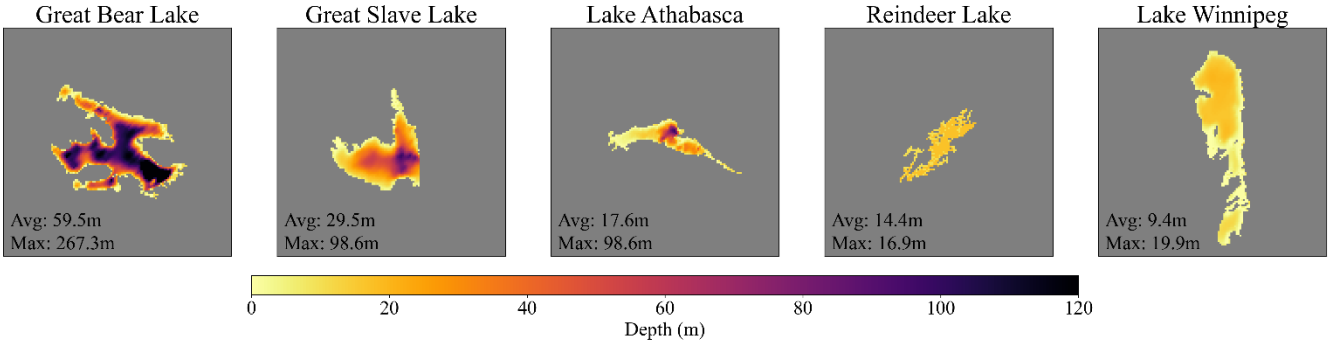

Figure 2. Global Lake DataBase (GLDB) lake depth data reprojected onto the IMS coordinate reference system (EPSG:3411), cropped by their respective 128x128 bounding box, and masked to remove surrounding water bodies.

The resultant harmonized dataset included daily lake ice cover observations and forcing variables from February 25th, 2004, to December 31st, 2021, for each of the 5 study lakes. Of the total data available, the years 2004–2016 were used for tuning and training, which was normalized between 0 and 1 using min-max scaling on a per-variable basis.

The years 2018–2021 were used to create a holdout data set for independent testing of the trained model. The exclusion of 2017 was done to induce a temporal gap between the training and holdout datasets. The forcing variables in the testing set were scaled using the minimum and maximum values derived from the training dataset. To provide additional testing, CIS records and FLake model predictions were used over the testing period. For some of the lakes in this study, the CIS record divided the lake body into two sections. These separate records were combined and averaged to obtain a single ice cover observation for the entire lake. Finally, the scale was shifted from 0–10 to 0–1, representing the fraction of ice cover. FLake predictions sourced from ERA5-Land were preprocessed the same way as the forcing variables. To convert the FLake predictions from their raw ice thickness values (m) to binary ice cover extent maps, a threshold value of 1 mm was applied.

2.4 Lake Ice Forecasting with Deep Learning (LIF-DL)

The proposed LIF-DL model produces ice-cover predictions by considering the combined effect of the predictor variables across both space (lake surfaces) and time. A lead time of seven days (one week) is used, meaning that a single pass through the model produces a one-week forecast. This parameterization was selected because it provides a sufficiently large temporal window to capture ice cover extent changes during freeze-up and break-up, yet remains small enough to facilitate faster training and reduce computational memory requirements. To achieve the objective of forecasting over arbitrary time periods, the model is designed to be deployed autoregressively (time-stepping), reusing its forecasted ice cover as initial conditions, as shown in Figure 3. This approach requires that atmospheric forcing be available over the entire forecasting period and is similar to the time-stepping approaches used in existing lake models (Duguay et al., 2003; Mironov, 2008).

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

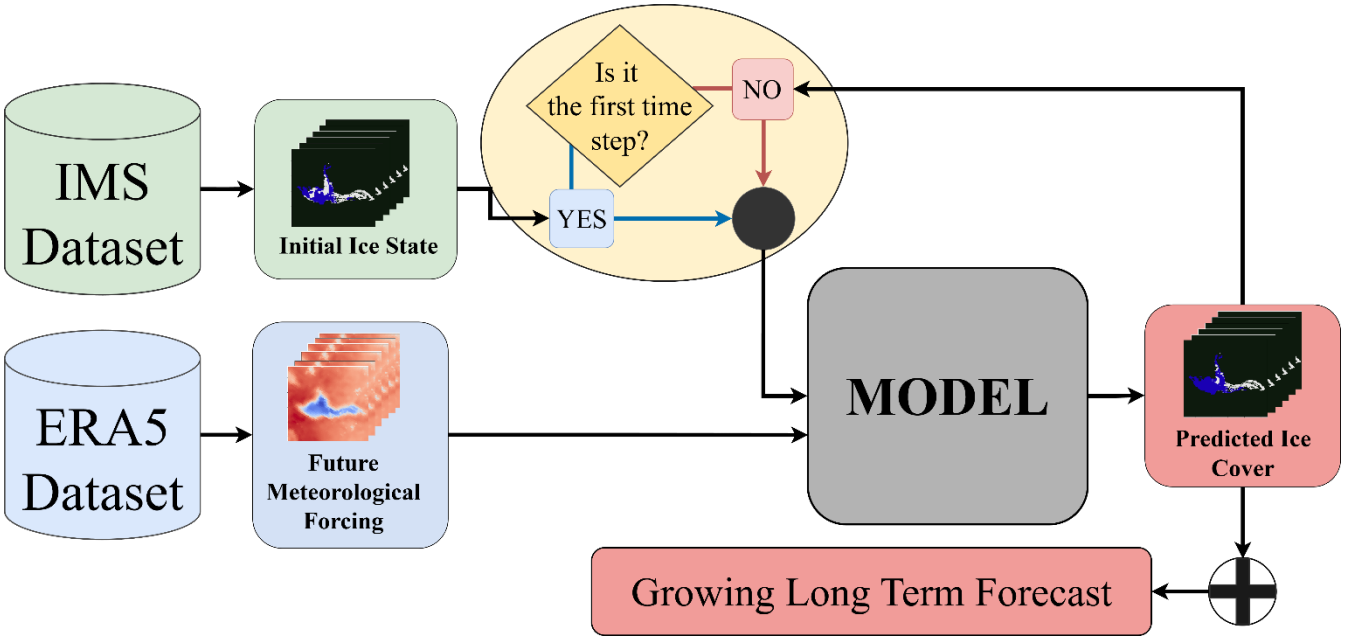

Figure 3. Autoregressive method for producing long-term forecasts using the LIF-DL model. Initial conditions are used to produce the first prediction, after which model predictions are used as input to continue forecasting.

LIF-DL is an encoder-decoder model that compresses the input data while simultaneously extracting relevant information in the encoder step, performs operations within the resultant latent space, and finally applies a decoder step to transform the latent representation into meaningful predictions at the original dimensionality. The encoders and decoders are built using Convolutional Neural Networks (CNNs), which are well adapted for handling spatial data and are widely used for this purpose (Alzubaidi et al., 2021; Haupt et al., 2022).

LIF-DL is also a sequence-to-sequence model, taking sequences of data as input and producing sequences of predictions as output. Within the LIF-DL architecture, the CNN encoders and decoders are not designed to consider the temporal dimension of the data. Temporal relationships are considered within the latent space, where Spatial-Temporal Transformer Networks (STTN) are used. STTNs were initially developed for video inpainting and as such are designed to consider relationships across space and time simultaneously (Zeng et al., 2020). Transformer-based networks utilize an attention mechanism which allows the model to focus on the most relevant parts of the input when making a prediction (Vaswani et al., 2017). Transformers have demonstrated success in spatial-temporal environmental modelling tasks (Lam et al., 2023; Li et al., 2021), and improvements in accuracy and computational cost compared to recurrent neural networks (RNNs) or purely CNN-based approaches (Dosovitskiy et al., 2020; Vaswani et al., 2017).

The STTN is a multi-layer, multi-head spatial-temporal transformer. The transformer is designed to simultaneously attend to all input frames with coherent contents. Specifically, a transformer matches the queries (Q) and keys (K) on spatial patches

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

across different scales in multiple heads; thus, the values (V) of relevant regions can be detected and transformed. Moreover, the transformers can be fully exploited by stacking multiple layers to improve attention results based on updated region features.

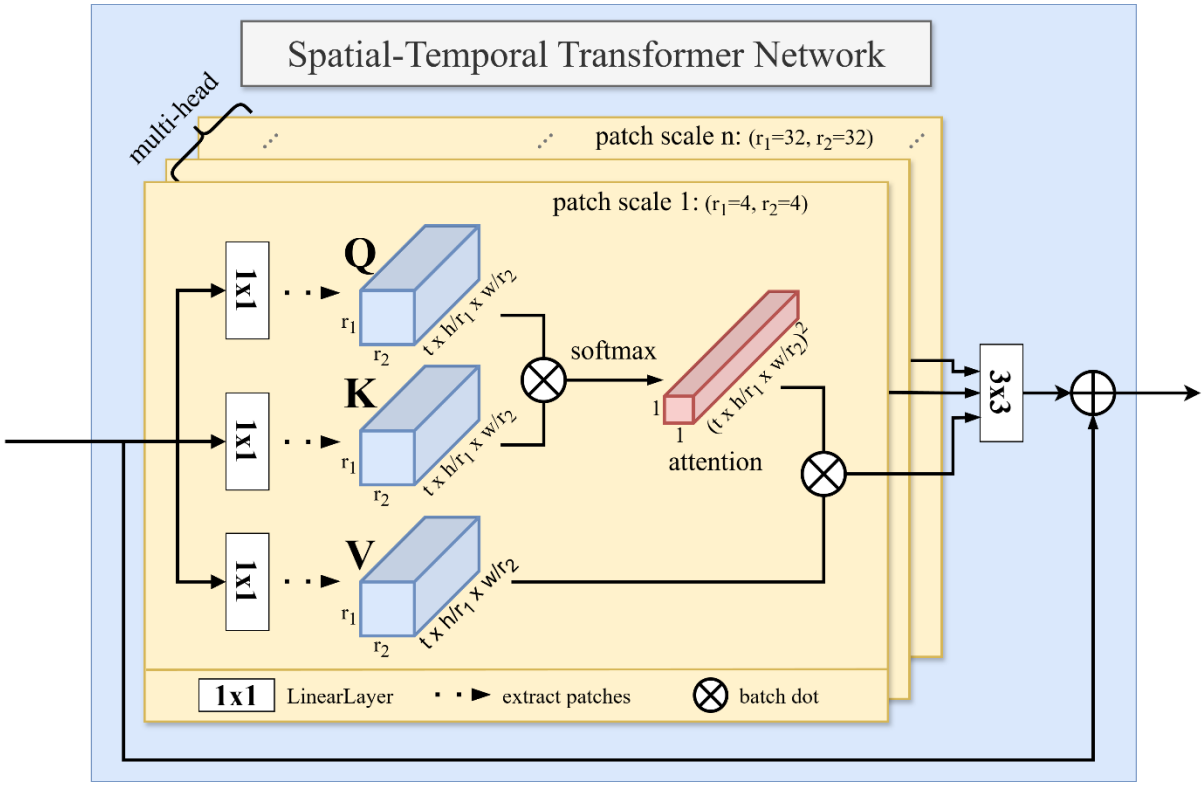

Figure 4. Overview of the Spatial Temporal Transformer Network (STTN) (adapted from Zeng et al., 2020). 1 × 1 and 3 × 3 denote the kernel size of 2D convolutions.

The full model architecture, depicted in Fig. 5, handles two separate input streams, which are combined within the encoded latent space. The first input stream handles the initial ice cover conditions for the previous week, while the second encodes the atmospheric forcing for the forecasting period. The use of two input streams was chosen to separate the inputs with different temporal coverage and to explicitly use the atmospheric forcing to update the ice cover. During initial testing, this layout was also found to perform better than a single stream approach.

Each input is passed to a separate CNN encoder, which reduces the spatial dimensions from 128x128 to 32x32 while augmenting the feature dimension to a configurable hidden dimension size. The encoded forcing data is then used to supply the query to the first STTN block, while the encoded ice cover supplies the key and value. In this way, the forcing data stream is seen to be updating the ice cover data stream via the attention mechanism. After this combination, data are passed through a configurable number of additional STTN blocks, as done in Zeng et al. (2020) and, finally, decoded via CNNs to produce a sequence of ice cover predictions in the original dimension size.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

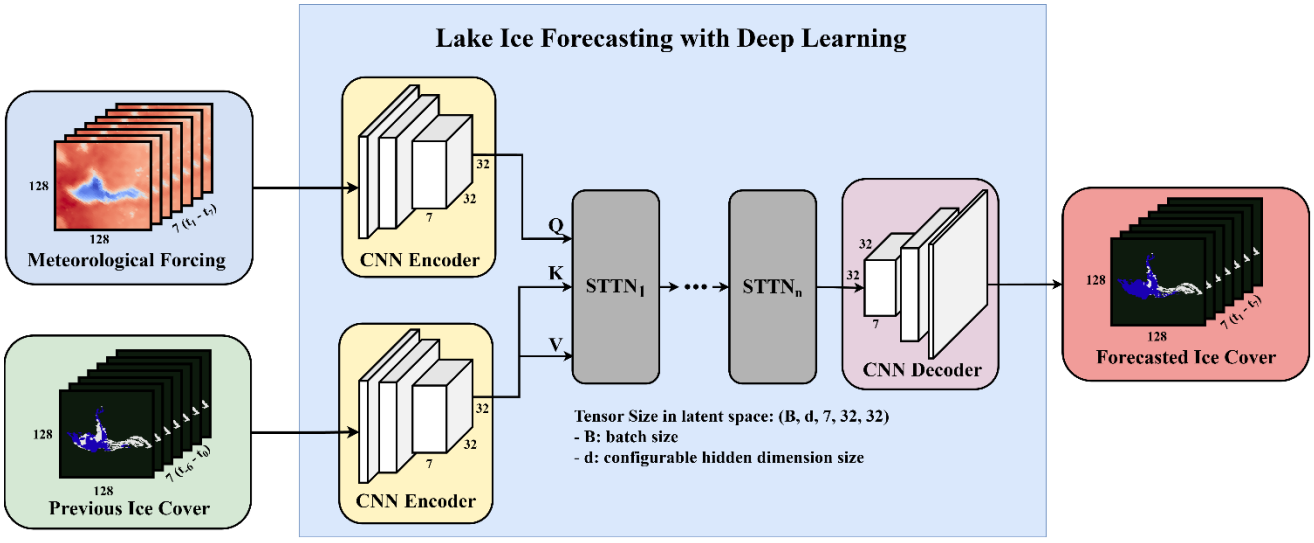

Figure 5. Overview of the Lake Ice Forecasting with Deep Learning (LIF-DL) architecture.

2.5 Model Optimization

Model optimization involved two steps: hyperparameter tuning and training, both of which relied solely on the training data. Each sample was derived from a two-week window, where the first week provided the initial ice cover input, and the second week supplied the atmospheric forcing and corresponding target ice cover. To evaluate generalization and prevent overfitting, 20 % of the training samples were randomly set aside as a validation set. This subset was used during hyperparameter tuning to compare different model configurations and during training to monitor performance on unseen data.

Both tuning and training phases were carried out using PyTorch (Paszke et al., 2019). We used the Adam optimizer, and a custom masked binary cross-entropy loss, which excluded land pixels from to ensure only lake pixels influenced model optimization. avoid larger lakes from dominating the loss, the loss was normalized by the number of lake pixels in each sample. Batches were also balanced across open water, partial ice, and full ice cover cases, based on the average fraction of ice cover in the input. This was done to address the implicit class imbalance within lake ice cover data, which is dominated by periods of no ice cover and full ice cover. Finally, a configurable dropout term was added to the LIF-DL layers as a regularization term to reduce overfitting (Hinton et al., 2012).

Hyperparameter tuning aimed to identify the configuration that minimized validation loss. A total of 200 randomly sampled configurations were evaluated, with tunable parameters including hidden dimension, number of STTN blocks, batch size, learning rate, and dropout rate. Each configuration was trained for up to 100 epochs, with early stopping triggered if the validation loss did not improve by a threshold of 0.001 for five consecutive epochs. The configuration with the lowest validation loss was then selected for full training.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Training utilized a learning rate scheduler which minimized the initial learning rate by a factor of 0.1 whenever validation loss did not improve by a threshold of 0.001 for five epochs. Training was stopped when validation loss did not improve by the same threshold for more than 10 epochs. Model weights were checkpointed whenever a new minimum validation loss was achieved. The final model corresponds to the checkpoint with the best validation performance across all training epochs.

2.6 Evaluation Methods

The trained LIF-DL model is evaluated using data from 2018–2021. These years are outside of the training data range and therefore represented new samples for the LIF-DL. Seasonal variable importance estimates were produced to analyze the relationships learned by the LIF-DL. Next, LIF-DL was deployed autoregressively to predict ice cover over the 4-year period using ERA5 atmospheric forcing. The 4-year ice-cover forecast was evaluated against IMS and CIS observations, and performance was compared to the baseline FLake model. Performance was assessed by investigating the overall spatiotemporal forecast, seasonal spatial patterns, fractional ice cover trends and timing of phenological events. In the case of seasonal assessments, freeze-up included September through December, and break-up included April through July. These periods were selected as they covered the break-up and freeze-up of all lakes, for all years under consideration. Table 2 summarizes the various evaluation methods used and is supplied at the end of this section.

2.6.1 Variable Importance

Variable importances are estimated to assess whether LIF-DL utilizes expected variables when producing forecasts. This step aims to justify the LIF-DL forecasts by demonstrating that the model is not exploiting unseen biases to achieve high accuracy, a common pitfall of data-driven approaches. Variable importance was estimated using the gradients at the model input layer, as calculated by backpropagation. De Sá (2019) showed that this method produces meaningful measurements, and moreover, that the estimates are highly correlated with variable importances calculated by random forests (a well-established variable importance estimation method). The assumption underlying this method is that the more important a feature is, the more the weights connected to the respective input neuron will change during the training of the model (De Sá, 2019). The trained LIF-DL was used to make a prediction on each sample day in the testing set, which were then used to calculate the gradients at the input layer by backpropagating the loss between prediction and target. This set of gradients was then divided into break-up and freeze-up subsets, based on the date ranges specified above. The gradients within each subset were aggregated together and normalized to get the final seasonal variable importance estimates.

2.6.2. Overall Forecast Performance

Overall spatiotemporal forecast accuracy for LIF-DL and FLake was evaluated against IMS observations over the testing period. Model performance was quantified using overall accuracy (OA), F1-score, and Intersection over Union (IoU), all calculated from pixel-level counts of true positives, false positives, false negatives, and true negatives. OA measures the

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

proportion of correctly classified pixels across both ice and water. The F1-score, the harmonic mean of precision and recall, provides a balanced measure of performance when class distributions are imbalanced, such as during periods of very low or very high ice cover. IoU quantifies the spatial overlap between predicted and observed ice cover, with a value of 1 indicating perfect agreement. Metrics were computed for the whole 4-year period as well as separately for freeze-up and break-up to evaluate seasonal performance differences.

2.6.3 Spatial Accuracy of Predictions

To assess the spatial accuracy of the LIF-DL and FLake predictions, maps of freeze-up start (FUS) and break-up start (BUS) were produced. BUS is the first change from ice to water for a given pixel and is reported as day-of-year. Similarly, FUS is the first occurrence of ice for a given pixel. Over each of the four freeze-up and break-up periods, BUS and FUS were derived from the LIF-DL and FLake predictions, as well as the IMS observations.

To isolate the spatial patterns of BUS and FUS timing, the day-of-year maps were first converted to timing anomaly maps, by subtracting their respective mean. Next, the BUS and FUS timing anomaly maps were averaged across the 4-year period to produce a single mean anomaly map for each lake, event (BUS or FUS) and data source.

To enable direct comparison across sources, the resulting mean anomalies were normalized by dividing by the combined 99th percentile of absolute anomalies across all three sources. This normalization retained relative spatial variance while removing differences in absolute timing. It also ensured compatibility with the Structural Similarity Index Measure (SSIM), which assumes inputs are on a −1 to 1 scale. Overall, this procedure removed lake-wide timing offsets between datasets, allowing an unbiased comparison of the spatial patterns in BUS and FUS across sources. Evaluation was then conducted using the SSIM, Kendall Tau-b correlation, and Local Moran's I statistic, to compare the normalized timing anomalies of FUS and BUS between the IMS observations and the LIF-DL and FLake predictions.

SSIM quantifies the structural similarity of two images by considering their luminance (mean), contrast (variance), and structure (local spatial correlation), returning a score between −1 and 1, where 1 indicates perfect structural agreement (Zhou Wang et al., 2004). The implementation was modified to exclude land pixels, ensuring that only spatial patterns of ice cover contributed to the similarity measure. The non-parametric Kendall Tau-b correlation was also calculated to assess whether the spatial ordering of early and late freeze-up or break-up timing was similar between observations and predictions (Kendall, 1945). Finally, to identify and compare regions of spatial clustering, the Anselin Local Moran's I statistic was applied to each normalized mean anomaly map, producing a quadrant classification of local autocorrelation (Anselin, 1995). These quadrants distinguish high–high (HH) and low–low (LL) clusters (areas of similar timing anomalies) from high–low (HL) and low–high (LH) spatial outliers (areas of dissimilar anomalies). In the context of BUS and FUS, HH is the late cluster, LH captures early outliers, HL captures late outliers, and LL is the early cluster. To quantify the agreement in spatial clustering patterns between

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

the different quadrant classification maps, a weighted IoU metric was used, where weights were proportional to the frequency of each cluster type.

2.6.4 Ice Cover Fraction Trend Comparison

IMS observations, FLake and LIF-DL forecasts were all converted from a spatiotemporal timeseries to a one-dimensional fraction of ice cover time series, to allow for comparison to CIS observations. These time series were used to quantify the error between each observation and forecast pair. The metrics used were the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and seasonal Mean Absolute Scaled Error (MASE). MASE is calculated as the MAE scaled by the seasonal baseline forecasting error, where the baseline forecast is the previous year's observation. As a relative metric, seasonal MASE values less than 1.0 indicate superior forecast performance to the naïve method, and values greater than 1.0 indicate weaker performance.

2.6.5 Accuracy of Ice Phenology Timing

Within the IMS observations and model predictions, there may be small unfrozen areas, for example, near a lake's outlet, for a long period during the freeze-up period, or a small amount of accumulated lake ice on the lakeshore during the breakup process. Therefore, Kropáček et al. (2013), proposed using 5 and 95 % of the lake area as thresholds to extract ice phenology instead of the traditional 0 and 100 %, which was also adopted in Cai et al. (2019, 2022). Considering the 0.1 discretization of CIS observations, in this study, freeze-up was deemed to start when the ice cover fraction first reaches 0.1, while freeze-up end was deemed to be when the ice cover fraction first climbs above 0.9. Similarly, the break-up start was deemed to start when the fraction of ice first drops to 0.9 and end when it falls below 0.1. In this way, the continuous values of the spatial-temporal data records are aligned well with the discretized observations of the CIS record.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Table 2. Name, description, possible values and interpretation of the various assessment methods used in this study.

Name	Description	Possible Values	Interpretation
Variable Importance	Estimate of the importance of an input variable on LIF-DL prediction	0-	Higher values represent higher estimated relative importance.
OA (Overall Accuracy)	Proportion of correctly predicted pixels.	0-1	Higher values indicate better overall prediction performance.
F1-Score	Harmonic mean of precision and recall, balancing false positives and false negatives.	0-1	Higher values indicate stronger balance between precision and recall.
IoU (Intersection over Union)	Measures the spatial overlap between predicted and observed ice cover	0-1	Higher values indicate greater spatial agreement between prediction and observation.
SSIM	Measures the structural similarity between predicted and observed ice cover timing maps.	-1-1	Higher values indicate stronger structural correspondence in ice cover timing (e.g. similar FUS/BUS patterns).
Kendall Tau-b	Measures the rank correlation between predicted and observed ice cover timing maps.	-1-1	Higher values indicate stronger spatial correspondence in ice cover timing (e.g. similar FUS/BUS patterns).
Local Moran's I	Measures local spatial autocorrelation of FUS/BUS maps, identifying clusters of similar or dissimilar ice cover timing.	Low-Low, Low-High, High-Low, High-High, Not significant	Low-low and high-high capture clustering of similar values, for early or late timing respectively. Low-high and high-low capture spatial outliers.
MAE (Mean Absolute Error)	Mean absolute difference between predicted and observed fractional ice cover.	0-1	Lower values indicate lower overall prediction error.
RMSE (Root Mean Squared Error)	Square root of the average squared differences between predicted and observed fractional ice cover.	0-1	Lower values indicate lower overall prediction error (penalizes large errors more heavily).
MASE (Mean Absolute Scaled Error)	MAE scaled by a naïve seasonal baseline (last year's observation), to assess relative model performance.	0-infinite	Values < 1 indicate performance better than the seasonal baseline; values > 1 indicate worse performance.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

3. Results and Discussion

3.1 Variable Importance

The variable importance estimates, presented in Figure 6, indicate that the LIF-DL learned key relationships between atmospheric forcing variables and ice cover. Of greatest importance were temperature variables (air temperature, AFDD and ATDD), after which came solar radiation and lake depth. This follows the understanding of how lakes' thermodynamics drive ice formation, growth and decay (Duguay et al., 2003; Williams et al., 2004). Wind speed, precipitation and cloud cover all consistently scored the lowest, which was expected, as these variables exhibit more dynamic and localized effects on ice cover (Brown and Duguay, 2010). Notably, the LIF-DL exhibited variable importances that varied by season.

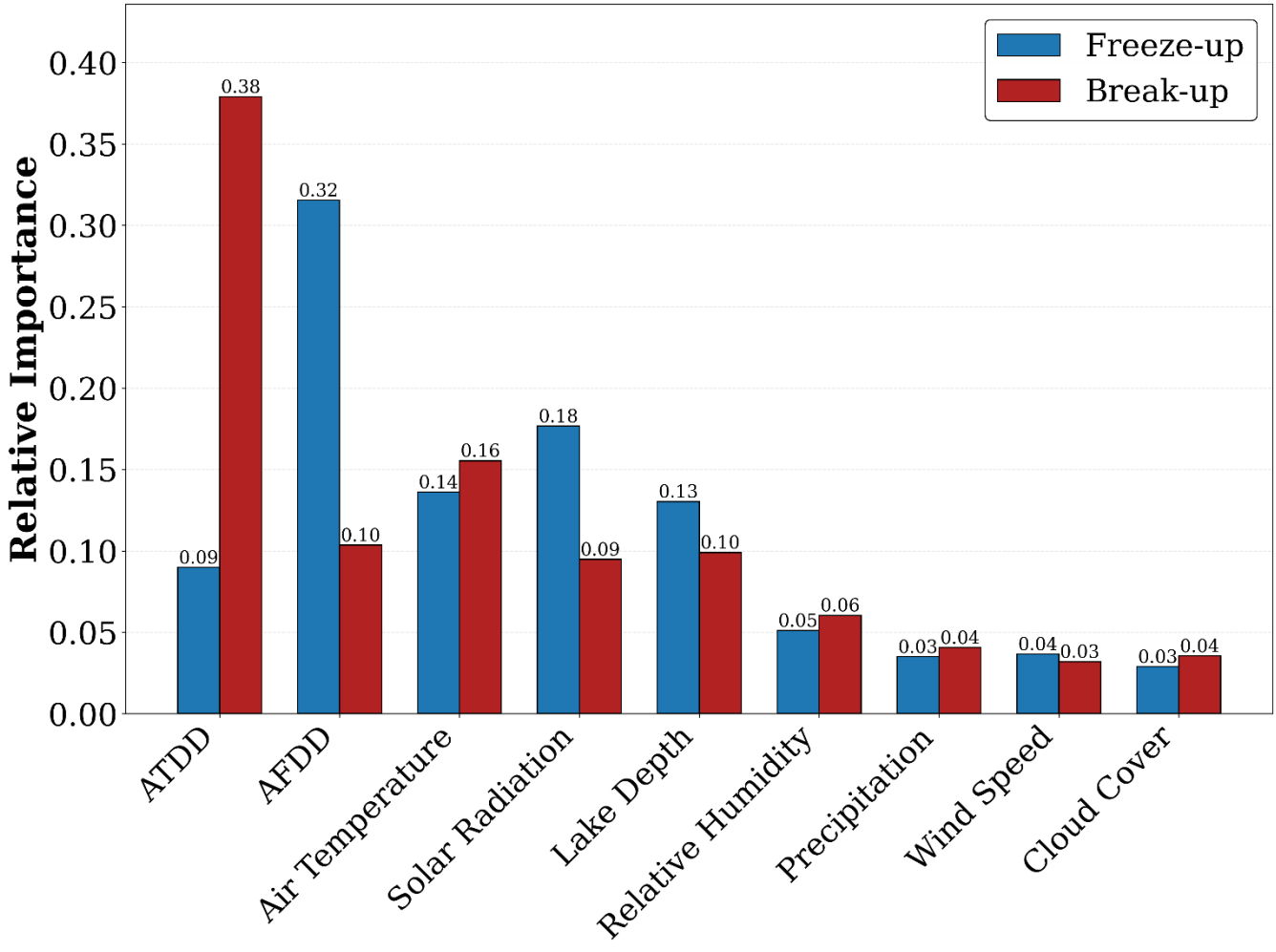

Figure 6. Variable importance estimates for break-up and freeze-up predictions.

The major seasonal differences in variable importance are for AFDD, ATDD and solar radiation. During the freeze-up, AFDD is the most valuable parameter (relative importance of 0.32), followed by solar radiation (0.18). Furthermore, ATDD is

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

relatively unimportant during freeze-up (0.09), compared to the break-up, where ATDD is the most important parameter (0.38). Also, during break-up, Solar Radiation and AFDD have much lower importance (0.09 and 0.1 respectively), while Air Temperature is the second most important variable (0.16). The hypothesized role of air temperature relative to its accumulated derivatives (ATDD and AFDD) is that it captures of short-term events, while the accumulated variables capture seasonal information. Other notable differences are that lake depth showed higher importance for freeze-up than break-up (0.13 vs 0.1), and air temperature was lower for freeze-up (0.13 vs 0.16). These seasonal differences align well with the understanding of lake ice dynamics.

For the freeze-up, the lake surface is initially open water, and solar radiation, air temperature and lake depth are all known to be important determinants of water temperature at this time (Williams, 1965). Combined with AFDD, these variables capture substantial information related to the lake's energy balance, which influences the formation of lake ice cover. Conversely, during break-up, the lake is initially fully ice-covered, limiting mixing of the water column and reflecting much solar radiation due to high surface albedo. This changes when the melt begins and open water patches form, which can absorb the incoming radiation. The fact that the break-up period was determined to be March–July meant that more samples were in a state of full-ice cover, during which time the influence of lake depth and solar radiation is minimized. This may explain why lake depth and solar radiation had lower importance during the break-up period, in favour of air temperature, as compared to the freeze-up samples.

While these variable importance estimates are a positive indication that the LIF-DL learns key relationships, the authors acknowledge that the black-box nature of the deep-learning architecture makes it difficult to conclude with certainty.

3.2 Full Forecast Evaluation

Table 3 presents the comparison of the 4-year spatiotemporal forecasts from LIF-DL and FLake with IMS observations. Overall, the LIF-DL outperforms FLake, with the most pronounced differences in performance during freeze-up, and for Great Bear Lake.

The consistently high OA achieved by both models partly reflects the nature of lake ice cover forecasting. Long periods of the year are dominated by either full ice cover or open water, which are easy to predict, and thereby inflate this metric. However, F1-scores and IoU provide a more balanced view, as they account for both the precision and recall of ice detection and the spatial overlap between predictions and observations, respectively. By these metrics, LIF-DL consistently achieves higher agreement with IMS, particularly during the more challenging freeze-up period, indicating it captures both the timing and spatial extent of ice cover more effectively than FLake.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Freeze-up represented the lowest performance for both models, but also the most significant performance gain over FLake for the LIF-DL. LIF-DL's achieved average freeze-up OA, F1-Score and IoU of 0.93, 0.9 and 0.82, compared to FLake, which achieved 0.88, 0.86 and 0.75 respectively. Stronger IoU scores are a positive indication that the LIF-DL forecasted ice cover with higher spatial agreement to IMS observations than did the FLake model. In contrast, break-up performance was stronger than freeze-up for both models, and the LIF-DL only slightly outperformed the FLake model on break-up overall.

Table 3. Forecasting performance over the 4-year testing period (2018-2021), evaluating model (LIF-DL, FLake) predictions against IMS observations.

		Overall Performance			Break-up Performance			Freeze-up Performance		
Lake	**Model**	**OA**	**F1-score**	**IoU**	**OA**	**F1-score**	**IoU**	**OA**	**F1-score**	**IoU**
Great Bear Lake	**LIF-DL**	0.97	0.98	0.96	0.97	0.98	0.96	0.95	0.93	0.87
	FLake	0.92	0.94	0.89	0.91	0.94	0.89	0.86	0.83	0.71
Great Slave Lake	**LIF-DL**	0.97	0.97	0.94	0.96	0.97	0.94	0.94	0.90	0.83
	FLake	0.94	0.95	0.91	0.95	0.96	0.92	0.88	0.85	0.74
Lake Athabasca	**LIF-DL**	0.96	0.97	0.93	0.96	0.96	0.92	0.93	0.90	0.82
	FLake	0.93	0.94	0.88	0.95	0.95	0.90	0.85	0.82	0.69
Reindeer Lake	**LIF-DL**	0.96	0.97	0.94	0.96	0.97	0.94	0.93	0.90	0.82
	FLake	0.95	0.96	0.92	0.95	0.96	0.92	0.91	0.89	0.80
Lake Winnipeg	**LIF-DL**	0.95	0.95	0.91	0.96	0.95	0.90	0.91	0.86	0.75
	FLake	0.96	0.96	0.92	0.95	0.95	0.90	0.92	0.90	0.82
Average	**LIF-DL**	0.96	0.97	0.94	0.96	0.96	0.93	0.93	0.90	0.82
	FLake	0.94	0.95	0.90	0.94	0.95	0.91	0.88	0.86	0.75

On the per-lake basis, LIF-DL performed best for Great Bear Lake and worst for Lake Winnipeg, while FLake had the strongest performance on Lake Winnipeg, and the weakest for Lake Athabasca. Great Bear Lake represented the greatest performance gain from the LIF-DL over FLake, where LIF-DL performed considerably better for both freeze-up and break-up. Great Bear Lake represented the largest and deepest lake of the study sites, and so this performance disparity may be representative of the enhanced importance of spatial understanding for such lakes. This is supported by the higher freeze-up performance of LIF-DL for Great Slave Lake and Lake Athabasca as well, the next two deepest lakes in the study set. Lake Winnipeg, on the other hand, represented the shallowest lake on average, and was the only site where FLake scored slightly better than the LIF-DL on freeze-up.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

3.3 Spatial Accuracy

Figures 7–10 summarize the spatial accuracy of the LIF-DL and FLake models, as compared to IMS observations across both break-up start (BUS) and freeze-up start (FUS) periods. LIF-DL demonstrates superior spatial accuracy in both BUS and FUS predictions, with higher structural similarity (SSIM), better spatial ordering (Kendall Tau-B), and stronger clustering agreement (IoU on Local Moran's I clusters). It was found that FLake particularly struggles with producing accurate BUS patterns, representing the most significant performance gain from LIF-DL. Furthermore, Lake Winnipeg and Reindeer Lake show the largest performance gaps between models, favouring the LIF-DL. The following section describes these results in more detail, organized by the two seasonal periods, break-up and freeze-up.

3.3.1 Break-up Start (BUS)

Figures 7 and 8 present the spatial accuracies of BUS timing anomalies from the LIF-DL and FLake forecasts compared to IMS observations, capturing both overall structural similarity and agreement between spatial clusters. LIF-DL-derived BUS maps achieved much higher SSIM scores (0.15–0.37) than FLake (-0.04–0.05), and higher Kendall Tau-B correlations (0.56–0.75 versus 0.25–0.56). The near-zero SSIM values for FLake indicate that there was little to no structural similarity in break-up patterns between its forecasts and the IMS observations. The clustering of BUS timing anomalies further supported that LIF-DL better captures the dominant BUS patterns, having achieved IoU agreement with IMS clusters between 0.54–0.66, compared to FLake (0.11–0.30). Visually, the strong structural similarity between LIF-DL and IMS BUS anomaly patterns is evident, particularly around major bays and river inflows.

One of the most dominant break-up patterns observed in the IMS was early timing anomalies concentrated close to river inflows or outflows, which were classified as LL clusters (see Fig. 8). This includes Johnny Hoe River which flows into the McVicar Arm of Great Bear Lake (bottom of lake), the Mackenzie (outflow) and Slave (inflow) rivers on Great Slave Lake (far left and central bottom respectively), the far-left end of Lake Athabasca where multiple inflows and outflows converge, and the Winnipeg River which meets Lake Winnipeg at its southern end. LIF-DL reproduced very similar LL clustering to IMS around these areas, which indicates that it incorporates the influence of rivers into break-up predictions. This is likely due to its data-driven formulation, which allowed it to learn the influence of these features from the observation data explicitly. This is contrasted with FLake forecasts, which could not incorporate the influence of rivers on BUS, as it lacks any such parameterization, and as such, its Local Moran's I maps lack these clusters. The capacity of the deep learning model to capture such specific physical processes without direct parameterization represents a promising opportunity for improving lake models. However, it is also an indication that the LIF-DL is unlikely to generalize well to unseen lakes.

Late BUS timing was positively correlated with lake depth and negatively correlated with distance from shore, as evidenced by the HH clusters for IMS observations. LIF-DL forecasts displayed very similar HH clustering patterns, while FLake again

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

struggled. The dominant BUS pattern of FLake predictions instead showed a stronger correlation with latitude, exhibiting a northward melt progression. This highlights the importance of spatial context—a feature leveraged by the LIF-DL—for simulating realistic break-up patterns across lake surfaces.

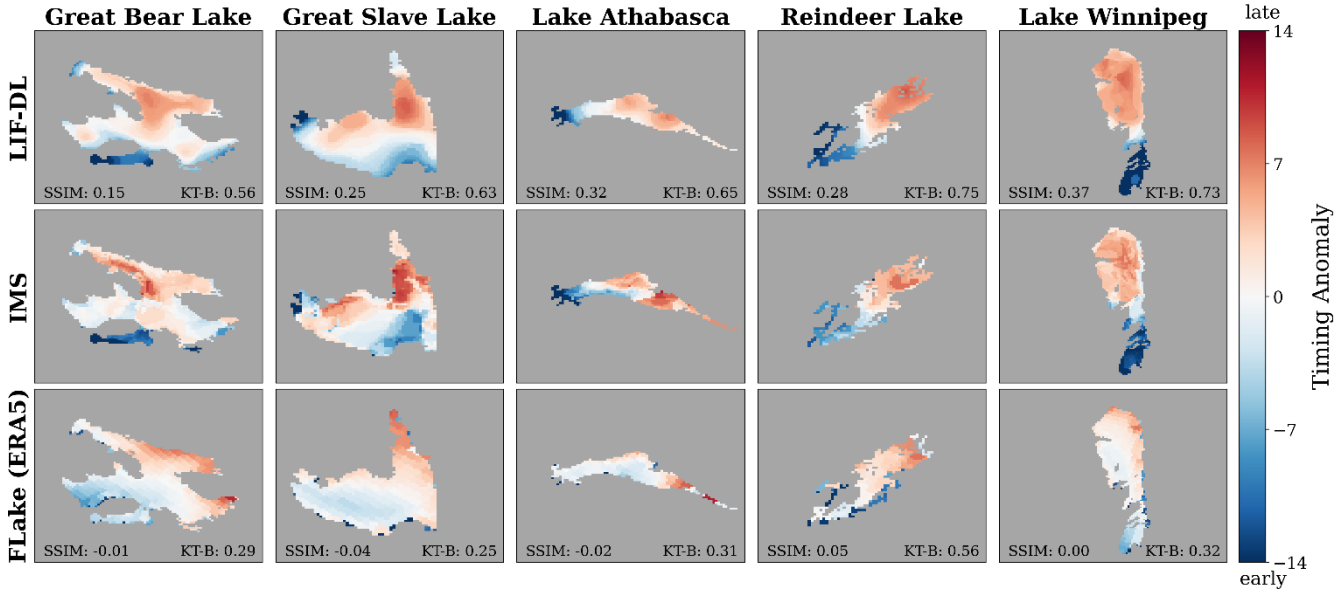

Figure 7. Average (2018-2021) timing anomalies in break-up start (BUS), from IMS observations and LIF-DL and FLake predictions. SSIM and Kendall Tau-B (KT-B) are reported, which were calculated from normalized anomaly maps.

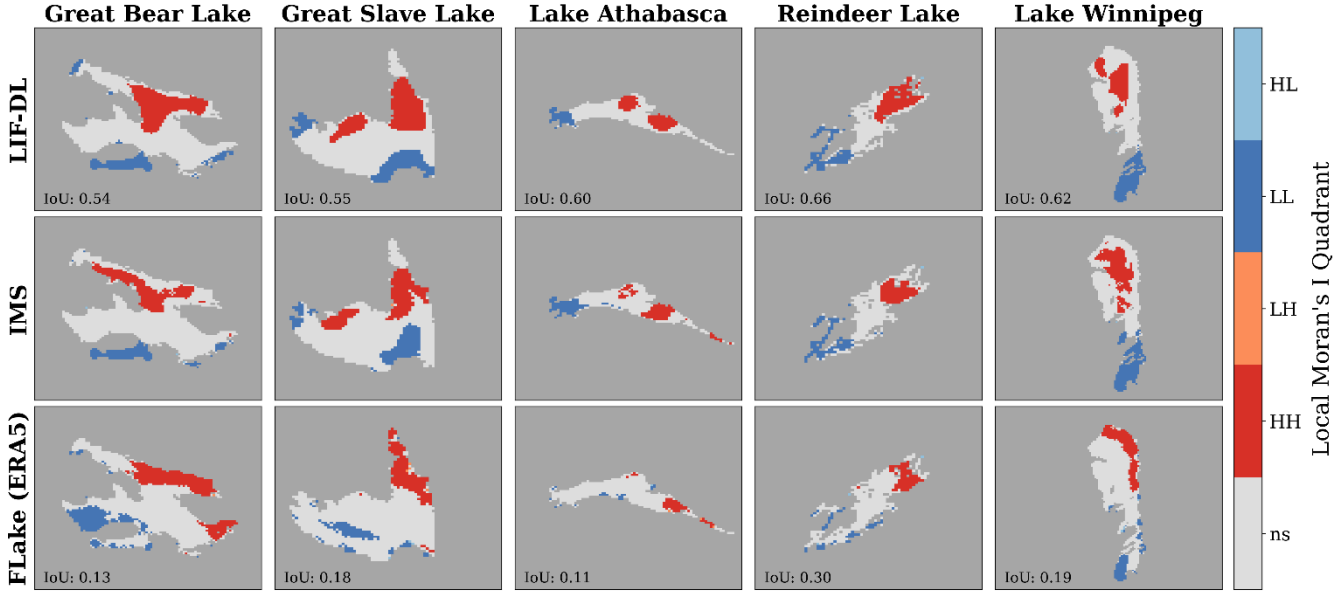

Figure 8. Local Moran's I quadrants from the average (2018-2021) BUS anomalies. ns refers to non-significant cluster. IoU is reported, capturing the agreement between clusters in predictions (LIF-DL, FLake) and observations (IMS).

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

3.3.2 Freeze-up Start (FUS)

Figures 9 and 10 present the spatial accuracies of FUS timing anomalies from the LIF-DL and FLake forecasts compared to IMS observations, showing both overall structural similarity and agreement between spatial clusters, respectively. LIF-DL achieved SSIM scores of 0.01–0.38, which were higher than FLake for each lake (-0.01–0.17), apart from Lake Winnipeg, where both models achieved near-zero SSIM compared to IMS FUS patterns. Similarly, Kendall Tau-B correlations were on average higher for LIF-DL (0.28–0.69) than for FLake (0.00–0.60).

The low SSIM scores of the FLake predictions were mostly due to its significantly higher timing variance, that is, a longer duration of freeze-up, as compared to IMS observations, since SSIM considers pixel variance. Nevertheless, Kendall Tau-B values showed that while the FLake overall structural similarity was weak, its relative ordering (rank correlation) of timing anomalies correlates with IMS observations, at least for Great Bear Lake, Great Slave Lake and Lake Athabasca. This is supported by examining the Local Moran's I clusters of FLake FUS, which also showed good agreement with IMS for those lakes (IoU of 0.37–0.53). Regardless, LIF-DL still outperformed FLake at capturing the dominant spatial clusters, achieving higher IoU scores for every lake except Great Bear Lake. Both LIF-DL and FLake had the weakest FUS pattern agreement for Lake Winnipeg (IoU 0.18 and 0.05 respectively), and the strongest agreements for Great Bear Lake (IoU 0.43 and 0.53) and Great Slave Lake (0.54 and 0.51). This is the near opposite of the performance pattern observed for BUS, where Lake Winnipeg represented the highest performance. Generally, the LIF-DL had more stable spatial accuracy across lakes than did the FLake model.

The dominant FUS pattern observed in the IMS data was that ice cover forms near shore and in sheltered bays, where wind speeds are often weaker, and the lake is typically shallower. From the initial formation of ice cover near-shore, the ice-water boundary progresses inward, expanding into the deeper and more open areas of the lake last. The LIF-DL predictions create a smooth transition from early near shore to late in the deeper and more open areas. FLake also captures this pattern, though its FUS anomalies are more discontinuous between pixels, and the overall variation in timing is much higher. The 'roughness' of the FLake FUS pattern is in part due to resampling from a lower spatial resolution (~9 km in ERA5-Land; Muñoz-Sabater et al., 2021). However, it was found that FLake tends to predict near-shore ice far earlier than observed in IMS, anywhere from 14–30 days too early, which was determined from the residual—IMS vs FLake—map of FUS.

The extreme early bias along shorelines is likely a consequence of the gridded FLake approach being unable to consider the spatial lake context. Large lakes store significant heat and undergo large-scale mixing, maintaining heat energy near shore longer than would be observed for smaller and shallower lakes. In the gridded context, without lake-wise spatial context, near-shore pixels are treated as independent shallow lakes, which explains why FLake predicts a far earlier FUS date along the

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

shore than observed. This again highlights the benefit of modelling the entire lake surface in a one-shot approach, as is done with LIF-DL, and is shown to improve the spatial accuracy of FUS predictions.

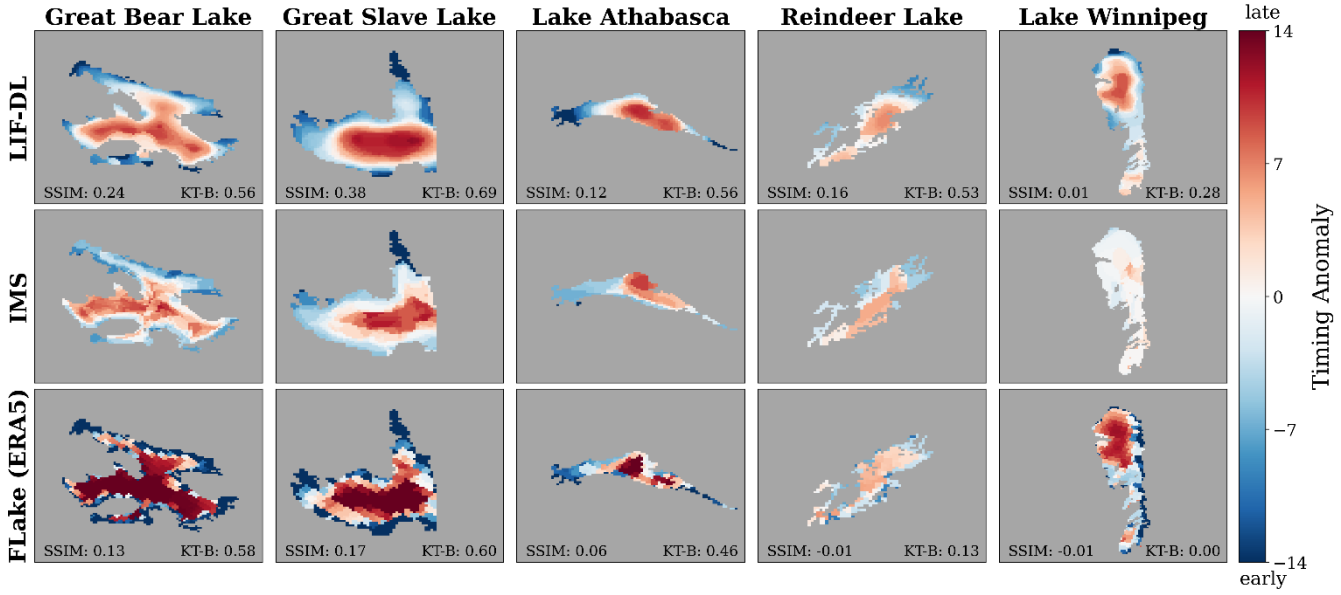

Figure 9. Average (2018-2021) timing anomalies for freeze-up start (FUS), from IMS observations and LIF-DL and FLake predictions. SSIM and Kendall Tau-b (KTB) are reported, which were calculated from normalized anomaly maps.

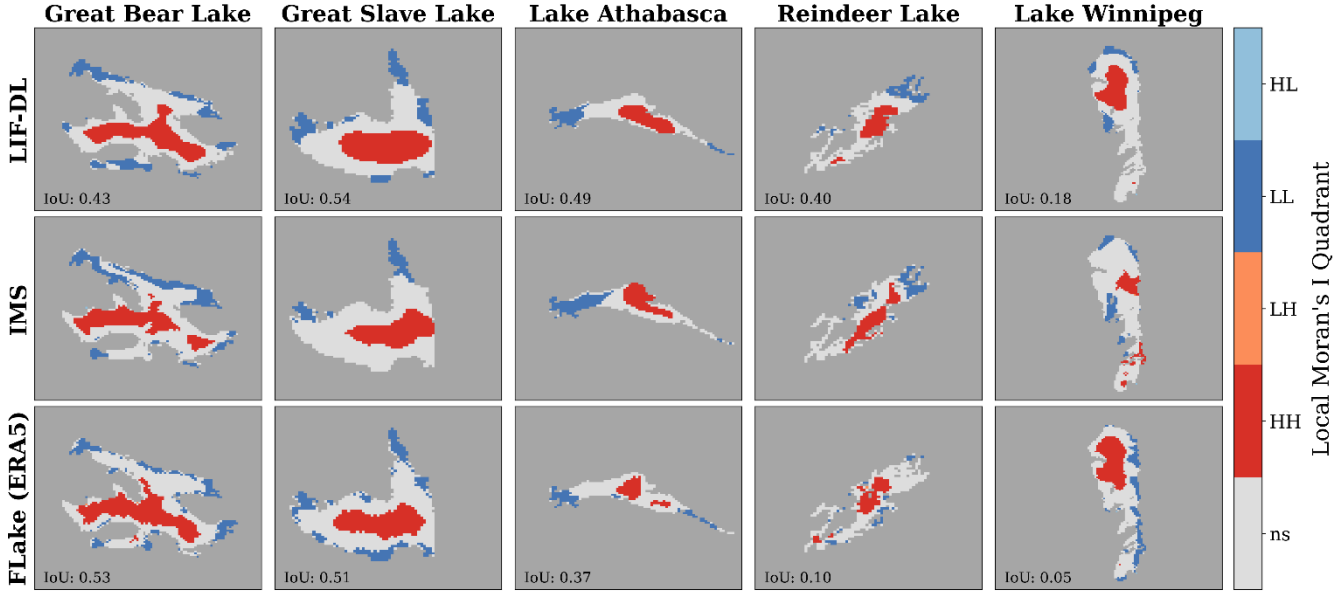

Figure 10. Local Moran's I quadrants from the average (2018-2021) FUS anomalies. ns refers to non-significant cluster. IoU is reported, capturing the agreement between clusters in predictions (LIF-DL, FLake) and observations (IMS).

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

3.4 Ice Cover Fraction Trend Comparison

Figure 11 illustrates the temporal evolution of lake ice fraction for the 4-year period (2018–2021) for each of the study lakes, as observed in IMS and CIS, and as predicted by LIF-DL and FLake. Table 4 reports the errors in the fraction of ice cover between each of the observational and model sources. Overall, the LIF-DL forecast had very high ice cover fraction agreement, achieving an average RMSE of 0.10 against IMS and CIS, outperforming FLake, which had average RMSE's of 0.15 and 0.20 respectively. The LIF-DL forecast meets the GCOS requirement of 10 % uncertainty (World Meteorological Organization (WMO) et al., 2022), being below an MAE of 0.10 (10 % ice cover) for all validation cases. Across the MASE, LIF-DL achieved average values less than 1.00 when compared to IMS (0.60) and CIS (0.90), while FLake had average MASE above 1.00 (1.10 and 1.50, respectively). This means that the LIF-DL outperforms the naïve seasonal baseline, which is another indication that it appropriately leverages atmospheric forcing to produce predictions with meaningful yearly variation. In contrast, FLake generally underperforms the seasonal baseline, especially when validated against CIS, where it consistently underperforms (MASE 1.2–1.8).

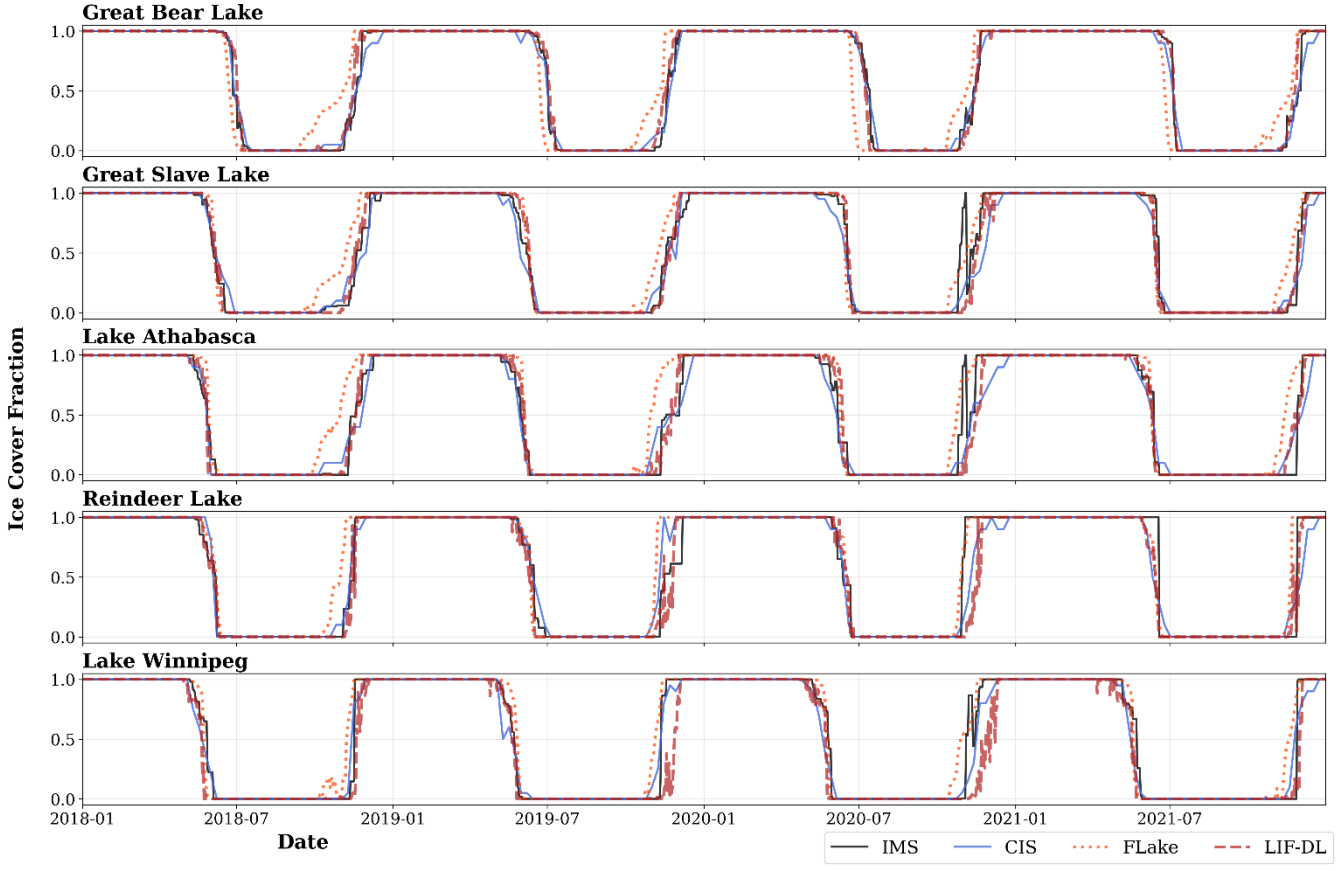

Figure 11. Lake-wide ice cover fraction as observed (IMS, CIS) and predicted (FLake, LIF-DL) over 2018-2021.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Table 4. Fraction of ice cover errors between each observation (IMS, CIS) and model (LIF-DL, FLake), over the entire four-year period (2018-2021). Comparisons to CIS are rounded to the nearest decimal.

Lake	Validation	Model	MAE	RMSE	MASE
Great Bear Lake	IMS	LIF-DL	0.02	0.06	0.41
		FLake	0.07	0.18	1.70
	CIS	LIF-DL	0.00	0.10	0.60
		FLake	0.10	0.20	1.70
Great Slave Lake	IMS	LIF-DL	0.03	0.09	0.51
		FLake	0.05	0.12	0.96
	CIS	LIF-DL	0.00	0.10	0.90
		FLake	0.10	0.10	1.60
Lake Athabasca	IMS	LIF-DL	0.03	0.10	0.54
		FLake	0.06	0.17	1.19
	CIS	LIF-DL	0.00	0.10	0.80
		FLake	0.10	0.20	1.80
Reindeer Lake	IMS	LIF-DL	0.03	0.12	0.47
		FLake	0.04	0.15	0.61
	CIS	LIF-DL	0.00	0.10	0.90
		FLake	0.00	0.10	1.20
Lake Winnipeg	IMS	LIF-DL	0.04	0.16	1.04
		FLake	0.04	0.12	0.91
	CIS	LIF-DL	0.00	0.10	1.10
		FLake	0.00	0.10	1.20
Average	IMS	LIF-DL	0.03	0.10	0.59
		FLake	0.05	0.15	1.07
	CIS	LIF-DL	0.00	0.10	0.90
		FLake	0.10	0.20	1.50

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

The success of LIF-DL in producing strong 4-year forecasts autoregressively is notable, given that its training regime only involved single-week predictions. During the ice-free (value of 0) and ice-on (value of 1) periods, both models demonstrate perfect correspondence with observations, which corroborates why the overall accuracies reported in Table 2 (Sect. 3.2) are high. The forecasted trends in the ice cover fraction for each lake again illustrate how both models struggled more with predicting freeze-up compared to break-up, especially FLake.

The most significant performance differences were observed for Great Bear Lake and Lake Athabasca, where FLake performed particularly poorly, with RMSEs of 0.20 against CIS observations and 0.17–0.18 against IMS. In contrast, LIF-DL performed much better, achieving RMSEs of 0.10 against CIS and 0.06–0.10 against IMS. Additionally, FLake performed particularly poorly on the MASE metric for these sites, scoring between 1.2–1.8 compared to LIF-DL's 0.41–0.80. As can be seen in Fig. 11, FLake tends to begin predicting ice cover much earlier than observed, especially for these two sites, which explains its weaker performance. This is likely due to Flake's near-shore bias towards early freeze-up, which was uncovered during spatial analysis (Sect. 3.3).

Similar to the overall performance results (Sect. 3.2), FLake slightly outperformed the LIF-DL for Lake Winnipeg when validated against IMS, which represented the lowest performance for the LIF-DL forecast. Looking at the Lake Winnipeg ice cover fraction trends (bo row in Fig. 11), freeze-up correspondence for the LIF-DL is high in years 2018 and 2021 but very weak in 2019 and 2020. For the break-up, LIF-DL also has some erroneous 'spikes' leading up to the break-up periods of 2019 and 2021. These errors are the primary contributors to the lower performance of LIF-DL on Lake Winnipeg, compared to other lakes.

Comparison with CIS observations revealed modest disagreement with IMS, particularly during freeze-up periods, where CIS often reports slightly earlier freeze-up start and longer freeze-up durations. Despite these differences, LIF-DL forecasts generally fall within the range of variability observed between CIS and IMS, suggesting that the model's error is comparable to the uncertainty present in existing observational datasets. This supports the reliability of LIF-DL predictions for operational and research use.

3.5 Temporal Accuracy of Ice Phenology Events

Table 5 reports the MAEs and RMSEs achieved for the timing of each lake's phenological events, along with the averages across all sites. LIF-DL error against CIS is slightly higher than against IMS, partly due to the weekly resolution of CIS. Furthermore, errors are comparable to existing observational disagreement (Cai et al., 2019; Murfitt and Brown, 2017), supporting the reliability of LIF-DL forecasts. Overall, LIF-DL outperformed FLake, having lower MAE and RMSE for nearly every category. The LIF-DL achieved far more stable error, with average MAEs between 3 and 10 days, compared to FLake,

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

which had average MAEs of 5 and 22 days. LIF-DL had similar performance on break-up timing to its performance on freeze-up timing, while FLake performed significantly worse on freeze-up than for break-up.

Table 5. Timing errors (MAE, RMSE) of ice phenology events between observations (IMS, CIS) and forecasts (LIF-DL, FLake).

			Break-up Start		Break-up End		Freeze-up Start		Freeze-up End	
Lake	**Validation**	**Model**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Great Bear Lake	**IMS**	**LIF-DL**	3	3	2	3	4	5	3	4
		FLake	6	7	12	13	30	32	10	11
	CIS	**LIF-DL**	3	3	8	8	6	7	10	11
		FLake	9	10	19	19	27	32	17	17
Great Slave Lake	**IMS**	**LIF-DL**	4	4	1	2	6	6	8	9
		FLake	5	6	4	4	24	27	8	10
	CIS	**LIF-DL**	7	10	9	10	9	9	4	4
		FLake	8	10	12	13	15	16	11	11
Lake Athabasca	**IMS**	**LIF-DL**	4	5	4	5	7	8	7	8
		FLake	13	13	3	4	23	25	15	18
	CIS	**LIF-DL**	6	7	9	10	11	14	14	14
		FLake	12	13	10	11	8	8	27	28
Reindeer Lake	**IMS**	**LIF-DL**	8	9	2	3	5	5	9	11
		FLake	7	9	3	4	11	12	12	16
	CIS	**LIF-DL**	11	11	10	12	7	8	4	5
		FLake	7	7	10	12	4	5	16	16
Lake Winnipeg	**IMS**	**LIF-DL**	16	18	6	7	7	8	14	15
		FLake	9	9	4	5	20	21	3	3
	CIS	**LIF-DL**	18	21	9	10	8	10	6	8
		FLake	7	8	6	7	16	21	13	13
Average	**IMS**	**LIF-DL**	7	8	3	4	6	6	8	9
		FLake	8	9	5	6	22	23	10	11
	CIS	**LIF-DL**	9	11	9	10	8	10	8	8
		FLake	8	9	11	13	14	16	17	17

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

The LIF-DL achieved MAEs between 4–11 days for freeze-up start and 3–14 days for freeze-up end, while FLake achieved MAEs from 4–30 days for freeze-up start and 3-28 days for freeze-up end. FLake's freeze-up start timing was particularly poor for Great Bear Lake (27–30 days) and Great Slave Lake (15–24 days), due to its significantly earlier prediction of freeze-up, as previously described.

Conversely, for break-up, LIF-DL and FLake had more even performance, with MAEs between 1–18 days, and 3-19 days respectively. The break-up performance of LIF-DL was weakest for Lake Winnipeg (MAE 16–18 days for break-up start), considerably worse than its average performance (MAE 7-9 days for break-up start), whilst FLake performed near its average for Lake Winnipeg (MAE 7–9 days), outperforming LIF-DL in this case. FLake performed the worst for break-up start on Lake Athabasca (MAE 12–13 days) and break-up end on Great Bear Lake (13–19 days), whereas the LIF-DL performed near its best for these cases (4–6 days and 2–8 days respectively).

4. Conclusions

The objective of this study was to develop a spatial-temporal lake ice cover forecasting model using deep learning and benchmark its performance to assess the potential of such approaches within the field of lake ice modelling. The proposed Lake Ice Forecasting with Deep Learning (LIF-DL) model can forecast ice cover extent over an entire lake surface in a spatially explicit manner. This approach addresses the limitations of traditional one-dimensional models (e.g. FLake), which neglect horizontal spatial interactions critical to ice cover dynamics. s represents a novel approach to lake ice modelling, being the first spatial deep learning-based approach in this field. This study demonstrates that deep learning methods can capture physically meaningful ice cover processes, improving upon existing methods and thereby having the potential to greatly enhance lake ice modelling efforts.

LIF-DL performance was benchmarked against FLake, which is used as the lake parameterization scheme in many weather forecasting models as well as ERA5 reanalysis (Hersbach et al., 2020; Mironov et al., 2010). Across multiple evaluations—overall forecast, spatial patterns, ice cover fraction, and phenology timing—against observational datasets (IMS, CIS), the LIF-DL consistently outperformed FLake.

For ice phenology, LIF-DL predicted freeze-up events on average within 6–8 days, and break-up events within 3–9 days, improving upon FLake, corresponding errors of 10–22 days and 5–11 days. FLake displayed particularly poor performance for freeze-up on Great Bear Lake, where mean absolute errors reached 27–30 days for freeze-up start, while the LIF-DL had errors of 4–6 days. In terms of ice cover fraction, LIF-DL achieved MAEs ≤ 0.1 across all tests, meeting the GCOS 10 % uncertainty threshold (World Meteorological Organization (WMO) et al., 2022). LIF-DL also consistently achieved MASEs

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

less than 1.0, demonstrating it outperformed the naïve seasonal baseline, whereas FLake consistently underperformed (average MASEs 1.1–1.5).

Spatially, LIF-DL forecasts showed far greater agreement with IMS observations. Break-up start (BUS) patterns showed good structural similarities (SSIMs 0.15–0.37) between LIF-DL and IMS, while the FLake prediction had almost no structural similarity (SSIMs -0.04–0.05). Using Local Moran's I analysis, we showed that the dominant early BUS clusters (LL) were positioned near river inlets or outlets in the IMS observations (Fig. 8), and that the LIF-DL forecast captured these clusters well (IoUs 0.54–0.66), while the FLake forecast did not (IoUs 0.11–0.19). In the case of freeze-up start (FUS) the LIF-DL again had more spatially consistent forecasts (SSIMs 0.01–0.38), though with a less significant improvement over FLake (SSIMs -0.01–0.17). The dominant early (LL) clusters for freeze-up were concentrated along near-shore regions, while the dominant late (HH) clusters were concentrated in the main lake basins (Fig. 10). Both models were able to capture these patterns well, with the LIF-DL scoring IoU values between 0.18–0.54 and FLake achieving IoU values between 0.05–0.53.

For break-up, the spatial consistency achieved by LIF-DL represents a major improvement over FLake. LIF-DL accurately captured the influence of river inflows on early break-up and the persistence of ice in deeper lake regions where thicker ice would be expected. FLake, on the other hand, exhibited a simplistic northward progression of break-up timing, failing to reflect these local spatial controls.

For freeze-up, while FLake captured the general spatial pattern reasonably well, it suffered from substantial timing errors, likely because it could not account for lake-wide thermal mixing. This limitation led to premature ice formation near shorelines, especially in large, deep lakes. LIF-DL substantially reduced these errors, as it can consider the spatial interactions that represent the moderating influence of lake-wide mixing on near-shore water temperatures.

The superior performance of LIF-DL across spatial and temporal metrics demonstrates that deep learning can effectively be applied to lake ice modelling. Variable importance analysis indicated that the LIF-DL was most sensitive to physically meaningful drivers of lake ice cover, including accumulated freezing and thawing degree days, air temperature, solar radiation, and lake depth. This agreement suggests that LIF-DL learned meaningful physical relationships, providing confidence to its forecasts. Notably, the model maintained stable performance over a four-year evaluation period, despite being trained solely on one-week forecasts, indicating robustness under long-range forecasting and varying atmospheric conditions. Moreover, LIF-DL's timing errors for freeze-up and break-up events are comparable to existing observational disagreement (Cai et al., 2019; Murfitt and Brown, 2017), and other existing modelling approaches (Brown and Duguay, 2011; Pour et al., 2012). Together, these findings highlight LIF-DL's strong potential for operational application in lake ice cover modelling.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Nevertheless, several limitations remain. First, the interpretability of the LIF-DL remains limited; while variable importance estimates provide some insight, it does not fully reveal the model's learned relationships. Second, the generalizability to unseen lakes was not evaluated. The model's ability to capture river-influenced break-up patterns without explicit river parameterization suggests it may have memorized these lake-specific features. Third, while IMS has been extensively evaluated and represents the predominant gap-free ice cover dataset, it still contains errors and biases that may propagate into the LIF-DL, and its 4 km resolution limits its applicability to smaller lakes or for capturing finer scale ice processes.

Future work will seek to address these limitations and further improve model performance. Promising directions include integrating a physics-based ice component to create a hybrid model, improving interpretability and generalizability, and enabling gridded simulations of ice thickness to be incorporated. Additional enhancements could involve parameterizing snow cover and incorporating emerging high-resolution datasets, such as the ~1 km ESA CCI Lakes lake ice cover product (ESA CCI Lakes, 2025), to improve spatial fidelity.

Building on its demonstrated accuracy and robustness, LIF-DL has significant implications for both operational forecasting and scientific research. Accurate, spatially-resolved lake ice forecasts can enhance numerical weather prediction (NWP) by improving lake-atmosphere coupling, particularly for large lakes where ice cover is highly heterogeneous during freeze-up and break-up periods. Operational applications extend to shipping, navigation, and ice roads, where knowing the timing and location of freeze-up and break-up is critical for routing. Moreover, LIF-DL could be applied to forecast ice cover under future climate change scenarios, supporting adaptation strategies for communities and industries dependent on seasonal ice. Similarly, it could be applied for hindcasting, leveraging long-term reanalysis products such as ERA5, providing valuable insights into historical trends and the spatial heterogeneity of ice in response to global warming.

In conclusion, spatiotemporal deep learning represents a substantial advance in lake ice modelling. LIF-DL outperforms the traditional physics-based model FLake, particularly in spatial accuracy and freeze-up timing, demonstrating the value of integrating spatial temporal data-driven approaches. Future developments, including hybrid modelling frameworks, higher-resolution datasets, and expanded geographic coverage, offer the potential to scale these methods regionally, nationally, and globally. The continued integration of observational data, physically informed modelling, and machine learning will be essential for improving both our scientific understanding and practical forecasting of lake ice, supporting operational decision-making and long-term climate assessments alike.

Code and data availability

The associated model code and scripts are available on https://github.com/H2OSam/lif-dl. The exact version of the code used to produce the results and figures is archived under https://doi.org/10.5281/zenodo.17575279 (Johnston, 2025a). Pre-processed lake datasets and the trained model data are archived under https://doi.org/10.5281/zenodo.17543536 (Johnston, 2025b).

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

IMS data was retrieved through the National Snow and Ice Data Center: https://doi.org/10.7265/N52R3PMC (U.S. National Ice Center, 2004). The Copernicus Climate Data Store (CDS) can be used to download the ERA5 data under https://doi.org/10.24381/cds.adbb2d47 (Copernicus Climate Change Service, 2018) and ERA5-Land data under https://doi.org/10.24381/cds.e2161bac (Copernicus Climate Change Service, 2019). GLDB dataset can be downloaded here: http://www.flake.igb-berlin.de/old/ep-data.shtml (Toptunova et al., 2019). The Canadian Ice Service records were accessed by directly contacting the service (https://www.canada.ca/en/environment-climate-change/services/ice-forecasts-observations/about-ice-service.html) (Environment and Climate Change Canada, 2023).

Author contribution

CD conceptualized the study, and SJ developed the methodology. CD and SJ performed data curation, and then SJ developed the software and conducted the investigation, validation and visualization under the supervision of CD and JM. SJ prepared the original manuscript draft, which was reviewed and edited with the help of CD and JM. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada and the Digital Research Alliance of Canada.

Financial Support

This research has been supported by the Natural Sciences and Engineering Research Council of Canada (grant no. RGPIN-2023-05752 to C. Duguay).

References

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions, J Big Data, 8, 53, https://doi.org/10.1186/s40537-021-00444-8, 2021.

Anselin, L.: Local Indicators of Spatial Association—LISA, Geographical Analysis, 27, 93–115, https://doi.org/10.1111/j.1538-4632.1995.tb00338.x, 1995.

Arp, C. D., Cherry, J. E., Brown, D. R. N., Bondurant, A. C., and Endres, K. L.: Observation-derived ice growth curves show patterns and trends in maximum ice thickness and safe travel duration of Alaskan lakes and rivers, The Cryosphere, 14, 3595–3609, https://doi.org/10.5194/tc-14-3595-2020, 2020.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Balsamo, G., Salgado, R., Dutra, E., Boussetta, S., Stockdale, T., and Potes, M.: On the contribution of lakes in predicting near-surface temperature in a global weather forecasting model, Tellus A: Dynamic Meteorology and Oceanography, 64, 15829, https://doi.org/10.3402/tellusa.v64i0.15829, 2012.

Brown, L. C. and Duguay, C. R.: The response and role of ice cover in lake-climate interactions, Progress in Physical Geography: Earth and Environment, 34, 671–704, https://doi.org/10.1177/0309133310375653, 2010.

Brown, L. C. and Duguay, C. R.: The fate of lake ice in the North American Arctic, The Cryosphere, 5, 869–892, https://doi.org/10.5194/tc-5-869-2011, 2011.

Brown, L. C. and Duguay, C. R.: Modelling Lake Ice Phenology with an Examination of Satellite-Detected Subgrid Cell Variability, Advances in Meteorology, 2012, 1–19, https://doi.org/10.1155/2012/529064, 2012.

Cai, Y., Ke, C., Li, X., Zhang, G., Duan, Z., and Lee, H.: Variations of Lake Ice Phenology on the Tibetan Plateau From 2001 to 2017 Based on MODIS Data, JGR Atmospheres, 124, 825–843, https://doi.org/10.1029/2018JD028993, 2019.

Cai, Y., Duguay, C. R., and Ke, C.-Q.: A 41-year (1979–2019) passive-microwave-derived lake ice phenology data record of the Northern Hemisphere, Earth Syst. Sci. Data, 14, 3329–3347, https://doi.org/10.5194/essd-14-3329-2022, 2022.

Carrea, L., Crétaux, J.-F., Liu, X., Wu, Y., Calmettes, B., Duguay, C. R., Merchant, C. J., Selmes, N., Simis, S. G. H., Warren, M., Yesou, H., Müller, D., Jiang, D., Embury, O., Bergé-Nguyen, M., and Albergel, C.: Satellite-derived multivariate world-wide lake physical variable timeseries for climate studies, Sci Data, 10, 30, https://doi.org/10.1038/s41597-022-01889-z, 2023.

Copernicus Climate Change Service: ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/CDS.ADBB2D47, 2018.

Copernicus Climate Change Service: ERA5-Land hourly data from 1950 to present, Copernicus Climate Change Service (C3S) Climate Data Store (CDS) [data set], https://doi.org/10.24381/CDS.E2161BAC, 2019.

Dauginis, A. A. and Brown, L. C.: Recent changes in pan-Arctic sea ice, lake ice, and snow-on/off timing, The Cryosphere, 15, 4781–4805, https://doi.org/10.5194/tc-15-4781-2021, 2021.

De Sá, C. R.: Variance-Based Feature Importance in Neural Networks, in: Discovery Science, vol. 11828, edited by: Kralj Novak, P., Šmuc, T., and Džeroski, S., Springer International Publishing, Cham, 306–315, https://doi.org/10.1007/978-3-030-33778-0_24, 2019.

Derksen, C., Smith, S. L., Sharp, M., Brown, L., Howell, S., Copland, L., Mueller, D. R., Gauthier, Y., Fletcher, C. G., Tivy, A., Bernier, M., Bourgeois, J., Brown, R., Burn, C. R., Duguay, C., Kushner, P., Langlois, A., Lewkowicz, A. G., Royer, A., and Walker, A.: Variability and change in the Canadian cryosphere, Climatic Change, 115, 59–88, https://doi.org/10.1007/s10584-012-0470-0, 2012.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, https://doi.org/10.48550/ARXIV.2010.11929, 2020.

Duguay, C. R., Flato, G. M., Jeffries, M. O., Ménard, P., Morris, K., and Rouse, W. R.: Ice-cover variability on shallow lakes at high latitudes: model simulations and observations, Hydrol. Process., 17, 3465–3483, https://doi.org/10.1002/hyp.1394, 2003.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Duguay, C. R., Prowse, T. D., Bonsal, B. R., Brown, R. D., Lacroix, M. P., and Ménard, P.: Recent trends in Canadian lake ice cover, Hydrological Processes, 20, 781–801, https://doi.org/10.1002/hyp.6131, 2006.

ECMWF: in: IFS Documentation CY48R1 - Part IV: Physical Processes, ECMWF, 134–135, 2023.

Environment and Climate Change Canada: Canadian Ice Service: Ice Coverage Records, 2023.

Haupt, S. E., Gagne, D. J., Hsieh, W. W., Krasnopolsky, V., McGovern, A., Marzban, C., Moninger, W., Lakshmanan, V., Tissot, P., and Williams, J. K.: The History and Practice of AI in the Environmental Sciences, Bulletin of the American Meteorological Society, 103, E1351–E1370, https://doi.org/10.1175/BAMS-D-20-0234.1, 2022.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Q.J.R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R.: Improving neural networks by preventing co-adaptation of feature detectors, https://doi.org/10.48550/ARXIV.1207.0580, 2012.

Jeffries, M. O. and Morris, K.: Some aspects of ice phenology on ponds in central Alaska, USA, Ann. Glaciol., 46, 397–403, https://doi.org/10.3189/172756407782871576, 2007.

Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., and Kumar, V.: Physics-Guided Machine Learning for Scientific Discovery: An Application in Simulating Lake Temperature Profiles, ACM/IMS Trans. Data Sci., 2, 1–26, https://doi.org/10.1145/3447814, 2021.

Johnston, S.: H2OSam/lif-dl: LIF-DL v1.0, Zenodo [code], https://doi.org/10.5281/zenodo.17575279, 2025a.

Johnston, S.: Lake Ice Forecasting with Deep Learning - Archived Data, Zenodo [data set], https://doi.org/10.5281/zenodo.17543536, 2025b.

Kendall, M. G.: The Treatment of Ties in Ranking Problems, Biometrika, 33, 239–251, https://doi.org/10.1093/biomet/33.3.239, 1945.

Kheyrollah Pour, H., Attiah, G., Thompson, J., Scott, K. A., and Duguay, C.: Climogrid V1.0: A Spatially Distributed Thermodynamic Lake Ice Model, https://doi.org/10.2139/ssrn.4979507, 2024.

Korhonen, J.: Long-term changes in lake ice cover in Finland*, Hydrology Research, 37, 347–363, https://doi.org/10.2166/nh.2006.019, 2006.

Kropáček, J., Maussion, F., Chen, F., Hoerz, S., and Hochschild, V.: Analysis of ice phenology of lakes on the Tibetan Plateau from MODIS data, The Cryosphere, 7, 287–301, https://doi.org/10.5194/tc-7-287-2013, 2013.

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., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Li, Y., Lang, J., Ji, L., Zhong, J., Wang, Z., Guo, Y., and He, S.: Weather Forecasting Using Ensemble of Spatial-Temporal Attention Network and Multi-Layer Perceptron, Asia-Pacific J Atmos Sci, 57, 533–546, https://doi.org/10.1007/s13143-020-00212-3, 2021.

Ménard, P., Duguay, C. R., Flato, G. M., and Rouse, W. R.: Simulation of ice phenology on Great Slave Lake, Northwest Territories, Canada, Hydrol. Process., 16, 3691–3706, https://doi.org/10.1002/hyp.1230, 2002.

Mironov, D., Ritter, B., Schulz, J.-P., Buchhold, M., Lange, M., and Machulskaya, E.: Parameterisation of sea and lake ice in numerical weather prediction models of the German Weather Service, Tellus A: Dynamic Meteorology and Oceanography, 64, 17330, https://doi.org/10.3402/tellusa.v64i0.17330, 2012.

Mironov, D. V.: COSMO Technical Report No. 11: Parameterization of Lakes in Numerical Weather Prediction. Description of a Lake Model, https://doi.org/10.5676/DWD_PUB/NWV/COSMO-TR_11, 2008.

Mironov, D. V., Heise, E., Kourzeneva, E., Ritter, B., Schneider, N., and Terzhevik, A.: Implementation of the lake parameterisation scheme FLake into the numericam weather prediction model COSMO, Boreal Environment Research, 15, 218–230, 2010.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth Syst. Sci. Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, 2021.

Murfitt, J. and Brown, L. C.: Lake ice and temperature trends for Ontario and Manitoba: 2001 to 2014, Hydrological Processes, 31, 3596–3609, https://doi.org/10.1002/hyp.11295, 2017.

Murfitt, J. and Duguay, C. R.: 50 years of lake ice research from active microwave remote sensing: Progress and prospects, Remote Sensing of Environment, 264, 112616, https://doi.org/10.1016/j.rse.2021.112616, 2021.

Murfitt, J. C., Brown, L. C., and Howell, S. E. L.: Estimating lake ice thickness in Central Ontario, PLoS ONE, 13, e0208519, https://doi.org/10.1371/journal.pone.0208519, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, https://doi.org/10.48550/ARXIV.1912.01703, 2019.

Pour, H. K., Duguay, C. R., Martynov, A., and Brown, L. C.: Simulation of surface temperature and ice cover of large northern lakes with 1-D models: a comparison with MODIS satellite data and *in situ* measurements, Tellus A: Dynamic Meteorology and Oceanography, 64, 17614, https://doi.org/10.3402/tellusa.v64i0.17614, 2012.

Rouse, W. R., Oswald, C. J., Binyamin, J., Spence, C., Schertzer, W. M., Blanken, P. D., Bussières, N., and Duguay, C. R.: The Role of Northern Lakes in a Regional Energy Balance, Journal of Hydrometeorology, 6, 291–305, https://doi.org/10.1175/JHM421.1, 2005.

Rouse, W. R., Blanken, P. D., Bussières, N., Walker, A. E., Oswald, C. J., Schertzer, W. M., and Spence, C.: An Investigation of the Thermal and Energy Balance Regimes of Great Slave and Great Bear Lakes, Journal of Hydrometeorology, 9, 1318–1333, https://doi.org/10.1175/2008JHM977.1, 2008.

https://doi.org/10.5194/egusphere-2025-5576
Preprint. Discussion started: 24 February 2026

Rühland, K. M., Evans, M., and Smol, J. P.: Arctic warming drives striking twenty-first century ecosystem shifts in Great Slave Lake (Subarctic Canada), North America's deepest lake, Proc. R. Soc. B., 290, 20231252, https://doi.org/10.1098/rspb.2023.1252, 2023.

Tom, M., Aguilar, R., Imhof, P., Leinss, S., Baltsavias, E., and Schindler, K.: Lake Ice Detection from Sentinel-1 SAR with Deep Learning, https://doi.org/10.48550/arXiv.2002.07040, 6 May 2020.

Toptunova, O., Choulga, M., and Kurzeneva, E.: Status and progress in global lake database developments, Adv. Sci. Res., 16, 57–61, https://doi.org/10.5194/asr-16-57-2019, 2019.

U.S. National Ice Center: IMS Daily Northern Hemisphere Snow and Ice Analysis at 1 km, 4 km, and 24 km Resolutions, Version 1, https://doi.org/10.7265/N52R3PMC, 2004.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, in: Advances in Neural Information Processing Systems, 5999–6009, 2017.

Williams, G., Layman, K. L., and Stefan, H. G.: Dependence of lake ice covers on climatic, geographic and bathymetric variables, Cold Regions Science and Technology, 40, 145–164, https://doi.org/10.1016/j.coldregions.2004.06.010, 2004.

Williams, G. P.: Correlating Freeze-Up and Break-Up with Weather Conditions, Can. Geotech. J., 2, 313–326, https://doi.org/10.1139/t65-047, 1965.

World Meteorological Organization (WMO), United Nations Environment Programme (UNEP), International Science Council (ISC), and Intergovernmental Oceanographic Commission of the United Nations Educational, Scientific and Cultural Organization (IOC-UNESCO): The 2022 Global Climate Observing System (GCOS) Implementation Plan (GCOS-244), 2022.

Wu, Y., Duguay, C. R., and Xu, L.: Assessment of machine learning classifiers for global lake ice cover mapping from MODIS TOA reflectance data, Remote Sensing of Environment, 253, 112206, https://doi.org/10.1016/j.rse.2020.112206, 2021.

Zeng, Y., Fu, J., and Chao, H.: Learning Joint Spatial-Temporal Transformations for Video Inpainting, https://doi.org/10.48550/ARXIV.2007.10247, 2020.

Zhou Wang, Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity, IEEE Trans. on Image Process., 13, 600–612, https://doi.org/10.1109/tip.2003.819861, 2004.