Thank you to both reviewers for providing valuable feedback and comments on our manuscript. Our detailed responses to your comments, suggestions and concerns are detailed below. Please note that in the revised manuscript, all tracked changes are included for your reference.

Reviewer 1

The manuscript addresses an important modelling problem concerning the cryosphere with a novel spatial machine learning approach. Generally, the manuscript is well written and well organised, and it is particularly strong in its evaluation, the interpretation of results, and the generation of insights. The scientific context is sufficiently laid out and limitations are carefully addressed throughout the manuscript. Data and code are provided to a detailed extent. Nonetheless, there are two major concerns, primarily regarding the clarity in introducing the machine learning model, and the framing of this work as a forecasting system:

1. **Formal definition of model, inputs and outputs**: Section 2.4 and the following, which introduce the model and its evaluation, are lacking clarity. No notation was used to formally and mathematically introduce the inputs and outputs of the model. This makes it impossible to grasp the modelling pipeline and should be revised. I suggest introducing variable names and spatio-temporal indices. A formula should formalise the input to output mapping. Without this formal introduction, understanding the sequence-to-sequence nature of the model, and the spatio-temporal attributes of the chosen architecture, are challenging to comprehend even for an attentive reader. Figure 3 to Figure 5 should be revised as well, and should contain notation but also connect the various model components.

We have made corresponding improvements to Section 2.4 following your comment.

Figures 3-5 were revised and combined (now Figure 3, Line 220), based on yours + Reviewer 2's comments.

We have added variable names and formal input/output equations for introducing the model, with spatio-temporal indices (Lines 215-275):

"The proposed LIF-DL model (Fig. 3) produces daily one-week-ahead ice-cover predictions by considering the combined effect of the predictor variables across both space (lake surfaces) and time. Let $Y_t \in \mathbb{R}^{H \times W}$ *represent the observed lake ice cover for a given day t, where* $H = 128$ *and* $W = 128$ *denote the spatial dimensions. LIF-DL is designed as a sequence-to-sequence model operating on a fixed prediction horizon (lead time) of seven days.*

A single pass through the model uses a look-back window of one week of initial conditions to predict the subsequent week:

$$\hat{Y}_{t+1:t+7} = f(Y_{t-6:t}, X_{t+1:t+7})$$

where $\hat{Y}_{t+1:t+7} \in \mathbb{R}^{1 \times H \times W \times 7}$ *is the 7-days-ahead prediction of ice cover,* $Y_{t-6:t}$ *is the preceding week of ice cover, and* $X_{t+1:t+7} \in \mathbb{R}^{C \times H \times W \times 7}$ *represents the 7-day sequence of* C *atmospheric forcing variables over the target prediction period. This 7-day horizon was selected because it provides a sufficiently large temporal window to capture dynamic 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 prediction beyond 7 days, the model is deployed autoregressively, reusing its own predicted ice cover ($\hat{Y}$*) as the initial ice cover conditions (*Y*) in a time-stepping loop. This approach requires that*

atmospheric forcing variables (X) 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). Within this study, ERA5 reanalysis data are utilized as the core atmospheric forcing during model training and validation. Because reanalysis data is unavailable for future horizons, the model's operational viability remains untested, and future work will be required to evaluate how potential error propagation from forecasted meteorological forcing variables impacts model performance.

As illustrated in Figure 3, LIF-DL utilizes a dual-branch encoder architecture to handle different input streams. The first branch handles the initial ice cover conditions for the previous week ($Y_{t-6:t}$), while the second takes the atmospheric forcing for the prediction window ($X_{t+1:t+7}$). Each branch uses a Convolutional Neural Network (CNN) to encode their respective inputs. CNN encoders are well-adapted for extracting spatial features and are widely used for this purpose (Alzubaidi et al., 2021; Haupt et al., 2022). Each encoder reduces the spatial dimensions of the input tensors from 128×128 *to* 32×32*, while projecting the channel dimension to a configurable hidden feature size* d*, resulting in latent space tensors of size* $\mathbb{R}^{d \times 7 \times 32 \times 32}$*. 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 architecture was also found to perform better than a single stream approach.*

While the CNN encoders compress spatial data, they do not resolve any temporal dynamics. Temporal relationships are considered after the encoding step, where latent representations are passed into Spatial-Temporal Transformer Networks (STTN) (Zeng et al., 2020).

STTNs were initially developed for video inpainting and as such are designed to consider relationships across space and time simultaneously. 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 latent features from the two branches are mapped to the standard Transformer Attention mechanism components—Query (Q), Key (K), and Value (V)—as follows:

- *The encoded atmospheric forcing data supplies the Query (Q)*
- *The encoded initial ice cover data supplies the Key (K) and Value (V)*

Through this multi-head attention mechanism, the forcing data stream dynamically updates the ice cover data stream by matching spatial patches across different scales. To improve feature extraction and allow the model to learn deeply nested spatiotemporal relationships, we stack a configurable number (n) of STTN blocks sequentially. After the first STTN block is applied, the two branches are effectively merged, and all subsequent STTN blocks take a single input, which is transformed into the Q, K, *and* V.

Finally, the latent representation output from the final STTN block is passed to a CNN decoder. The decoder reverses the encoder's spatial compression, transforming the latent features from a dimensionality of 32×32 *back to the original spatial resolution of* 128×128*, yielding the final target sequence prediction* $\hat{Y}_{t+1:t+7}$*."*

2. **Forecasting given *future* reanalysis data**: The second major concern is that this work framed the proposed task as forecasting. However, it uses future ERA data, which in real-life would not be available beforehand, to forecast future LIC. This set-up does not emulate a realistic forecasting scenario and therefore it remains questionable whether it can claim to evaluate the ability of LIF-DL to forecast LIC. Actual forecasts may only rely on forecasted (rather than reanalysis/proxy ground truth) atmospheric forcings. Such forecasts contain more uncertainty themselves than ERA data. Furthermore, due to (1.) and missing descriptions, it is not necessarily clear that future forcings are the critical input.

This is a very fair critique. We acknowledge that testing the model with true forecasting data was never done in this study but is a plan for future explorations of the proposed model.

Given this limitation, we have re-framed the manuscript into a "prediction" based approach, rather than "forecasting". Additionally, we have expressly mentioned this caveat within Section 2.4 to make this more clear for the reader.

One issue remaining is that much work already uploaded (e.g. model code, data…) uses the model name Lake Ice Forecasting with Deep Learning (LIF-DL). Considering the substantial effort which would be needed to rename this across all assets and plans for further testing using forecast data to expand the algorithm, we have retained the original name.

These two concerns should be addressed by a revised manuscript.

Below I present more minor line-wise comments by section:

Abstract:

Line 6: Lakes only cover a significant proportion of Northern high-latitude landscape, but not the Southern high-latitude landscape. Insert "Northern" to make this distinction clear.

Northern was added

Line 8: Mentioning lake ice thickness this early on alongside LIC may lead readers to assume that lake ice thickness is also modelled in this paper. For the abstract I recommend focussing on the key variable modelled in this paper.

Mention of lake ice thickness was removed

Line 9: The word "lakes" should not be capitalised here.

Lakes as an ECV is defined by the WMO and capitalized in official documentation, and we have adopted this formatting in the manuscript.

Line 9: Double "prediction": Potentially change the first "prediction" to "forecasting".

Change applied – first "prediction" was changed to "forecasting" (Line 8)

Line 15: Changing "lake conditions" to "lake phase" or "lake state" could help to avoid confusion about what conditions are modelled.

Updated to say, "lake state" (Line 14):

"…to capture relationships between atmospheric forcings, lake depth and lake state (frozen or open water)."

Line 15: I recommend changing the order to naming inputs first and outputs second, as such: "[…] to capture relationships between atmospheric forcings, lake depth, and lake phase (frozen or open water)."

Change applied – see above reply (Line 14)

Line 16: I suggest sticking to one order of naming the five lakes throughout the paper (e.g. the order used in Figure 1).

Change applied (Line 15) – "…Great Slave Lake, Great Bear Lake, Lake Athabasca, Reindeer Lake and Lake Winnipeg."

Line 21: Referring to "one-week-ahead" forecasts would be clearer here. Maybe also specify that the model makes daily LIC predictions at 4 km spatial resolution, and that forecasts are 1 to 7 days ahead. Referring to the task as a segmentation task would also add more clarity early on.

Changes applied:

Line 12 – "… a novel data-driven segmentation model for predicting LIC…"

Line 18 – "LIF-DL was trained using 2004–2016 data to predict daily, 4-km resolution ice cover extent one week ahead. It was then deployed autoregressively to model ice cover during the 2018–2022 holdout period."

The abstract exceeds typical word count limits and should be shortened.

We have shortened the abstract based on your feedback along with Reviewer 2s (Lines 6-35)

Introduction:

Line 51: Bracket "(freeze-up/break-up)" is not necessary here and harms reading flow. This is already explained in the latter part of the same sentence.

"freeze-up/break-up" was removed (Line 47)

Line 52: Replace "stimulating" e.g. with "leading to".

Changed to "leading to" (line 48)

Line 55: Potentially relate to similar trends observed in sea ice to provide wider scientific context.

We have added additional information to provide more context on the broader trends in the cryosphere:

Line 45-47:

"Global trends in the cryosphere reveal the impacts of climate change. For example, according to the 2025 Arctic Report card, June snow cover in the Arctic has reduced by half since the 1960s and the lowest maximum sea ice extent was detected by satellite methods in 47 years (Meier et al., 2025; Mudryk et al., 2025)."

Line 63: "Lakes" does not need to be capitalised.

Lakes as an ECV is defined by the WMO and capitalized in official documentation, and we have adopted this formatting in the manuscript.

Line 73: What composition variable is CLIMo predicting?

Removed composition.

Line 73: I suggest replacing "more wholly" with something like "more comprehensively".

Mention of being more "comprehensive/wholly" was removed with the changes made based on your next comment. See Lines 69-75

Line 74: Specify what aspect of the model is two-layer, and explain what type of model FLake is.

This has been clarified (Line 69-75)

"FLake is one-dimensional (1D) bulk lake model that simulates the vertical temperature structure and mixing conditions of the water column (Mironov, 2008). The model utilizes a two-layer representation of the water column, which is divided into an upper mixed layer and lower thermocline. It further includes a thermodynamic module for simulating the formation and melting of ice."

Line 85: I suggest referring to this as the "point-wise gridded application of one-dimensional lake models" to conform with the wider literature. "Multiple points" is not specific enough.

Change applied: (Line 79) – "...do a point-wise gridded application, ..."

Line 90: No comma needed: "Data-driven deep learning approaches […]."

Change applied

(Line 91: Include additional citations such as perhaps

- Rolnick, David, et al. "Tackling climate change with machine learning." A*CM Computing Surveys (CSUR)* 55.2 (2022): 1-96.
- Reichstein, Markus, et al. "Deep learning and process understanding for data-driven Earth system science." *Nature* 566.7743 (2019): 195-204.)

Change applied (Line 89)

Additionally, revised the following sentence to focus on deep learning within ice cover modelling (Line 91-93), adding additional references:

"In the case of freshwater ice, deep learning models have proven successful for predicting phenology timing and ice conditions on rivers (De Coste et al., 2022; Liu et al., 2022), as well as modelling ice cover concentrations across the Laurentian Great Lakes (Abdelhady and Troy, 2025)."

- *Abdelhady, H. U. and Troy, C. D.: A deep learning approach for modeling and hindcasting Lake Michigan ice cover, Journal of Hydrology, 649, 132445, https://doi.org/10.1016/j.jhydrol.2024.132445, 2025.*
- *De Coste, M., Li, Z., and Dibike, Y.: Machine-learning approach for predicting the occurrence and timing of mid-winter ice breakups on canadian rivers, Environmental Modelling & Software, 152, 105402, https://doi.org/10.1016/j.envsoft.2022.105402, 2022.*
- *Liu, L., Davedu, S., Fujisaki-Manome, A., Hu, H., Jablonowski, C., and Chu, P. Y.: Machine Learning Model-Based Ice Cover Forecasting for a Vital Waterway in Large Lakes, JMSE, 10, 1022, https://doi.org/10.3390/jmse10081022, 2022.*

Line 92: The positioning of the subsentence ", such as Spatial-Temporal Transformer Networks […]", is not ideal, as it may rather suggest that STTNs are datasets.

Changed positioning accordingly. The sentence now reads (Line 89):
"This trend is further supported by recent advancements like the Spatial-Temporal Transformer Network (STTN; Zeng et al., 2020)—originally developed for video inpainting—which effectively leverages complex spatial-temporal data"

Line 96: Towards reads oddly.

The wording of this sentence has been adjusted (Line 94):

"Building on these successes, this study proposes a novel spatial-temporal deep learning framework to enhance the accuracy and scale of lake ice cover prediction."

Line 99: I believe this should say "develop a deep learning model" rather than "develop a model using deep learning"?

The wording was adjusted and this line now reads, "…develop a deep learning-based spatial-temporal lake ice cover model…" (Line 97)

Line 101: Be more specific about the adaptation of the pre-existing model: Was STTN extended?

Adjusted wording to say, "leveraged Spatial-Temporal Transformer Networks", since the internal mechanisms of this component were not extended in the study. (Line 99)

Line 102: The text only just mentioned that the STTN was developed for video inpainting.

We removed mention of video inpainting in this sentence, since it is mentioned earlier.

Data Sources:

Line 112: North Pole should be capitalised.

Change applied (Line 110)

Line 115: Improve "applied persistence". Is temporal extrapolation used to fill data gaps? Or say "lake phase was assumed to remain unchanged (or stationary in time) when no data was available".

Wording was modified here for clarity (Line 112-114)

"At times when insufficient sensor data is available, the IMS assumes lake state remains stationary in time, 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)."

Line 120: Add a sentence to explain how these two "ground truth" datasets differ and foreshadow which one is used in this study. Also mention the spatial resolution of CIS.

We have added information in the IMS paragraph to clarify its use case (Line 105):
"Gridded ice cover observations were sourced from ... providing both target data during model training as well as holdout data for model evaluation."

We clarified the spatial resolution of CIS data (time-series), as well as its use case in its paragraph (Line 118):

"The Canadian Ice Service (CIS) weekly ice cover product was used 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 consist of time-series data, where ice cover for an entire lake (or specific lake region) is aggregated into a single weekly integer ranging from 0 (no ice cover) to 10 (full ice cover)."

Line 135: Potentially say "two additional temporally aggregated variables".

Change applied – Line 136

Line 138: From what I understand this is the sum of days with temperatures below/above 0, not the sum of temperatures. Table 1 also misspecifies this. Is this calculated for each calendar year or for 365 days following 1 August? Maybe add a sentence to convey the intent here ("freezing days since the last summer are accumulated…")

The ATDD and AFDD refer to the daily sum of temperatures and are reported in degrees Celsius. We acknowledge that the name of these variables (Accumulated X Degree Days) is confusing, but this is the common name + definition within existing literature.

On the 365 days point, it is calculated on every 1-year period given the starting dates specified. We have added to the paragraph to make this clearer (Line 139):

"They are defined as the sum of air temperatures below (for AFDD) or above (for ATDD) zero degrees Celsius during their respective year-long period (starting February 1st for AFDD, and August 1st for ATDD)."

Line 164: Replace "also" with "additionally" to make clear that LIF-DL does not predict these.

Change applied (Line 165)

Study Lakes:

Like 180: Figure 1: Order of the lake plots: Tile 3 is usually expected on the left (reading direction).

Change applied (Line 180)

Line 181: Table 1: Make the text in Table 1 left-bounded for the ease of readability (particularly the leftmost column). There also is an issue with the relative humidity row. Fix AFDD and ATDD description: Some places suggest this is the number of days while others suggest this is a temperature.

Change applied – ATDD and AFDD are accumulated temperatures. We modified their preprocessing steps in the table to improve clarity (Line 182):

"AFDD: Accumulated (sum) of air temperatures below 0^{o}C, starting from February 1st."

"ATDD: Accumulated (sum) of air temperatures above 0^{o}C, starting from August 1st."

Data preprocessing:

Line 195: Explain why nearest neighbour interpolation/regridding was chosen over e.g. bilinear interpolation.

We initially wanted to avoid smoothing the forcing variables when upsampling. We acknowledge that bilinear may have been better suited here and will consider it in future.

Change applied to Line 195:
"Reprojection was done using a nearest neighbour interpolation to avoid introducing smoothed forcing values, after which the lake masks were applied to remove surrounding water bodies"

Line 197: Why do we need one-hot encoding when "masked" is not part of the prediction task? Inference can just be run for test regions.

Thank you for the comment. Yes, it would be logically equivalent to convert the IMS into simple binary classification maps. This would not impact results (BCE on 2 classes in either case). To avoid confusion, we have removed the mention of one-hot here.

Line 208: "To provide additional testing, CIS records and FLake model predictions were used over the testing period." How else were these used?

We updated the wording here to make it clear why they were used (CIS is additional observation data for ice cover fraction, FLake is a modelling baseline for benchmarking) (Line 209):

"FLake model predictions were used as a prediction baseline to benchmark the performance of LIF-DL over the 2018-2021 testing period. These were sourced from ERA5-Land and preprocessed in 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. Finally, CIS records were used as additional observations for evaluating ice cover fraction predictions from both models."

Line 209: "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." This is not clear to me. Why did various records exist for the same areas?

Within the CIS, SAR swaths are typically used to derive ice-cover fractions. For very large lakes, a single swath may be insufficient to cover the entire lake surface. The CIS separates these lakes into two sections based on SAR swath coverage. For the purposes of comparison in this study, one value was needed for the entire lake, so an average of the weekly fractions was used.

We have moved this explanation into the initial introduction of CIS data (Line 121)

LIF-DL:

Significant changes have been made to this section given your general comment up top, along with the specific comments below. We have included the entire revised section at the top of this document, under your relevant general comment. In response to the specific comments below we have attached small snippets of the relevant changes.

Line 215: Point to Figure 5 for the complete model visualisation.

Figure 3, 4 and 5 were replaced with a single Figure 3. The section begins by pointing to the combined figure (Line 215): "The proposed LIF-DL model (Fig. 3)…"

Line 217: Specify that it produces a daily one-week-ahead forecast (or a 1- to 7-day-ahead forecast).

Change applied – we used "7-day-ahead prediction" mostly throughout this section.

Line 217: "Parametrization" means something different in the context of mac hine learning. I suggest saying "forecast horizon" or "supervised learning set-up" to avoid confusion.

Change applied – modified to "This 7-day horizon was selected…" (Line 229)

Line 221: See general comment on this point. Why would we assume to have access to ERA data for the future? This is reanalysis data, not forecast data.

We now explicitly acknowledge this limitation within the text (Line 236):

"Within this study, ERA5 reanalysis data are utilized as the core atmospheric forcing during model training and validation. Because reanalysis data is unavailable for future periods, the model's operational forecasting viability remains untested, and future work will be required to evaluate how potential error propagation from forecasted meteorological forcing variables impacts model performance."

Line 225: This schematic lacks clarity: The "first time step" gate is not very logical and inputs and outputs should have variable names, and formally introduced temporal indexing.

We have merged Figures 3-5 together and removed the logic gate in the process. Now the figure contains a simple arrow to indicate that under autoregressive prediction output predictions are recycled as input ice cover. (Line 220, Figure 3)

Line 231: At this point it is not clear at all that this is a sequence-to-sequence set-up. Formal notation (in addition to a slight hint in Figure 5) in the main text must be used to comprehensively introduce inputs, outputs, their dimensionalities and time indices. This is a main weakness of the manuscript.

Formal notation has been added, defining the inputs and outputs (Line 225):

"A single pass through the model uses a look-back window of one week of initial conditions to predict the subsequent week:

$$\hat{Y}_{t+1:t+7} = f(Y_{t-6:t}, X_{t+1:t+7})$$

where $\hat{Y}_{t+1:t+7} \in \mathbb{R}^{1\times H\times W\times 7}$ is the 7-days-ahead prediction of ice cover, $Y_{t-6:t}$ is the preceding week of ice cover, and $X_{t+1:t+7} \in \mathbb{R}^{C\times H\times W\times 7}$ represents the 7-day sequence of C atmospheric forcing variables over the target prediction period. This 7-day horizon was selected because it provides a sufficiently large temporal window to capture dynamic ice cover extent changes during freeze-up and break-up yet remains small enough to facilitate faster training and reduce computational memory requirements."

Line 235: Figure 3, 4 and 5 are not very well connected in the text or visualisations. "MODEL" in Figure 3 should be replaced with the LIF-DL, and the three inputs into the STTN blocks (Q,K, V) from Figure 5 should match the inputs shown in Figure 4 for more coherence and clarity. Visually integrating the dual-branch encoder-STTN-decoder set-up in Figure 3 would be beneficial. Figure 4 is not needed if the reader can refer to existing literature and there are no novel aspects presented.

Figures 3-5 have been merged into a single Figure (Fig. 3) following this and Reviewer 2's comments (line 220)

Line 245: The pre-defined Transformer architecture already utilises "multiple layers". What is meant here? Stacking multiple layers of Transformers or using the original Transformer architecture?

To make this clearer, we have updated the associated sentence in the revised section (Line 267):
"To improve feature extraction and allow the model to learn deeply nested spatiotemporal relationships, we stack a configurable number (n) of STTN blocks sequentially."

Line 250: You may want to refer to this as a dual-branch architecture.

Change applied (Line 241) – "…LIF-DL utilizes a dual-branch encoder architecture to handle…"

Line 252: Relating to overall comment: It is not clear at this point in the text that future atmospheric forcing data is assumed to be available.

We have changed this to make it clear that reanalysis is the critical forcing input, and the associated limitation of this assumption for true forecasting (Line 236)

Line 264: The Figure caption is not sufficient and variables names and indexing needs to be used.

Figure 5 was improved by combining it with Figure 3&4, as well as improving the caption. (Line 220)

Model Optimisation:

Line 265: Maybe change to "hyperparameter tuning and parameter/model training" for parallelism and clarity.

Changed to "hyperparameter tuning and model training…" (Line 276)

Line 271: This is an unusual description since the model also is built in Pytoch.

We removed mention of PyTorch from this section.

Line 272: Calling this a "custom loss" if one filtering operation is applied is a bit of an overstatement.

Changed to "a masked binary cross-entropy loss…" (Line 281)

Evaluation Methods:

Line 297: Table 2 may be moved to the Appendix.

Change applied – Moved to become Table A1 (Line 690)

Line 312: From line 293 I expect a comparison with both IMS and CIS.

Overall Forecast Performance compared pixel-wise metrics, which meant it could not be used for comparison to CIS, which reports only a lake ice cover fraction.

To avoid confusion, we modified Section 2.6 to make it clear the use case of CIS vs IMS (Line 306)

"Specifically, the spatially explicit (pixel-to-pixel) assessments—including overall spatiotemporal performance, and seasonal spatial patterns—were evaluated exclusively against the gridded IMS product. Conversely, the lake-wide aggregated assessments—covering fractional ice-cover similarity and phenological event timing—were evaluated against both IMS and the non-gridded, time-series-based CIS observations."

Line 327: Mention if this is done per lake.

Changed to clarify it was done per lake (Line 341):

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

Line 352: It is not clear enough what a one-dimensional fraction of ice cover time series is. This is why notation (e.g. tensor notation) is necessary.

Removed the term "one-dimensional" to avoid confusion and added tensor notation for clarity between spatio-temporal products and ice-cover fraction timeseries. (Line 366):

"IMS observations, FLake and LIF-DL predictions were each converted from a spatiotemporal timeseries ($\mathbb{R}^{H\times W\times t}$) to a fraction of ice cover time series ($\mathbb{R}^{t}$), to allow for comparison to CIS observations."

Results and Discussion:

Line 374: Maybe change to "the thermodynamics of lakes drive […]".

Change applied (Line 386)

Line 379: Figure 6: In the top right corner or figure caption repeat the definitions of the Freeze-up and Break-up seasons and make clear that these are "variable importance estimates for predictions during freeze-up and break-up seasons".

Based on comments from Reviewer 2 – Figure 6 (now Figure 4) was updated to show importance versus time (Line 393). Freeze-up and Break-up periods are indicated within the new version of the figure.

Line 385: Air temperature does not need to be capitalised.

We removed capitalisation of the forcing variables in this paragraph. (Line 395)

Line 420: Table 3: Consider displaying an additional digit to make differences a bit more clear. The superior performance should be highlighted in some way (applies to all results tables).

We had originally included additional digits in our tables but found it to be visually 'cluttered' and did not provide sufficient additional information to warrant keeping them.

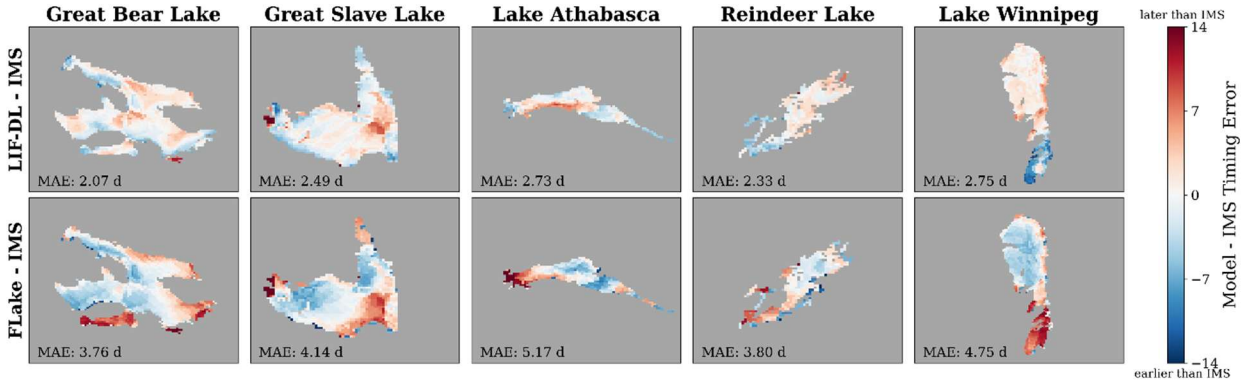
We have bolded the superior performance within each result table.

Line 467: Figure 7: Add (proposed) and (ground truth) labels for clarity. Maybe add a visualisation of the errors in the Appendix.

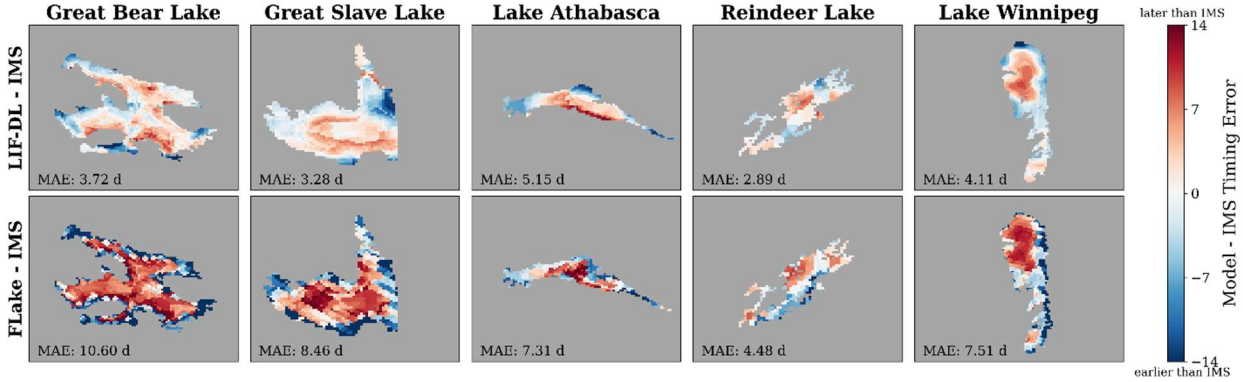
We have updated the figure labels to read: LIF-DL (proposed), IMS (observation), and FLake (baseline), for the spatial comparison figures (Figure 5-8).

We feel that 'observation' is more appropriate than saying ground truth for IMS.

Below are figure drafts of the error between timing anomalies of each model (LIF-DL, FLake) against IMS.



Break-up ^



Freeze-up ^

We decided to not include these in the appendix because interpretation in what the errors represent can be challenging, and do not capture spatial accuracies which was a primary focus of this section.

Specifically, we took the average timing anomalies per-pixel, per-lake, per-product, and this is what was plotted in Figure 5 and Figure 7. By averaging per-lake and per-product, the 'relative spatial timing distribution' is calculated across the lake surface – i.e. a representation of the freeze-up or break-up patterns within that specific product, for each specific lake.

When taking the error (difference) between two products, the result would be raw timing anomaly errors, i.e. how much their respective anomalies differ at each pixel. This is not the same as taking the actual error in timing for each event. And we feel there would be much confusion there for a reader, who would likely interpret as the former.

Line 526: Figure 11: FLake and LIF-DL are visually hard to discern (orange and red dotted/dashed lines are too similar). Maybe only chose zoomed in view on a selection of FUS/BUS segments.

We updated Figure 11 (now Figure 9) so that FLake uses a purple colour, which is in higher contrast to the red of LIF-DL (Line 569). We decided to leave the 4-year window because it was relevant to the discussion surrounding performance across long time horizons.

I acknowledge the significant work contained in this manuscript and encourage the authors to address the two major and additional minor weaknesses. The FUS and BUS perspectives in the evaluation as well as the variable importance analysis already provide significant scientific insight. For future work I would also suggest considering "teacher forcing" as a training strategies for auto-regressive roll-outs, and a short discussion of the importance of lake ice cover to indigenous communities. Another research avenue would be to train a unified model across the full region.

A sentence in the introduction was revised to highlight the importance of lake ice cover to indigenous communities, and an additional reference added (Line 50):

"Shortening ice cover seasons also present substantial socio-economic challenges for northern municipalities and Indigenous communities (Derksen et al., 2012; Knoll et al., 2019). Stable lake ice is critical not only for winter transportation (ice roads), but also for accessing traditional subsistence hunting and fishing grounds and preserving cultural identity"

- *Knoll, L. B., Sharma, S., Denfeld, B. A., Flaim, G., Hori, Y., Magnuson, J. J., Straile, D., and Weyhenmeyer, G. A.: Consequences of lake and river ice loss on cultural ecosystem services, Limnol Oceanogr Letters, 4, 119–131, https://doi.org/10.1002/lol2.10116, 2019.*

Thank you for your detailed and constructive feedback! We will take your suggestions into account for future work. We plan to scale the approach regionally and will investigate teacher forcing in future studies.

Reviewer 2

General Comment:

This manuscript develops a deep learning-based lake-ice model to improve the performance of simplified lake models in numerical weather prediction and climate simulations. The authors employ atmospheric forcing data (ERA5), lake geometry, and several accumulated variables as model inputs, with lake-ice coverage as the target output. A spatial-temporal transformer network (STTN) is used to establish the relationships between these variables and train the model.

Overall, this work represents a valuable attempt to develop machine learning-based lake models, and the manuscript is generally well written. This paper deserved to be published in GMD. Nevertheless, several deficiencies remain to be addressed before publication, particularly in the abstract and conclusion sections. Improvements are also needed in the description of figures and results, as well as in the overall structure and organization of the manuscript.

My detailed and specific questions and comment are provided in the attached PDF file. Please feel free to contact me if you encounter any issues with the PDF document.

Below, we have provided individual responses to some of your most significant comments from the marked-up pdf provided. Smaller tweaks are not responded to directly, but relevant changes can be seen in the tracked changes of the revised manuscript.

Abstract:

Line 6-11: I think this part should be shortened to 1-2 lines in the abstract. the abstract should focus on your main results and breakthroughs!

We have edited the abstract to make it more concise.

Line 30-32: I didn't see any evidence from you results. from my point of view. your deep-learning codes are likely based on statistical methods, rather than PINN (some method includes the dynamic-thermodynamic method). In the analysis, you didn't show any proof about the physical process. I think this statement is not accurate, this is just one of your assumption.
You are correct that out methods do not rely on PINN or similar physics-guided approaches. Our statement here was not meant to claim that the model understands true physical processes, rather we were trying to express that it is sensitive to the input variables with proportional importance to what we would expect given the literature. For example, air temperature had higher model sensitivity than cloud cover, which follows academic understanding of the thermodynamic drivers of ice cover.

We have adjusted the wording to make this more clear. "Variable importance analysis indicated that the model's sensitivity closely aligned with physically meaningful drivers, including air temperature, accumulated degree days, solar radiation, and lake depth." (Line 28)

Line 36-42: Please think about shorten this part. I don't think there should be such a long outlook here. Usually, it appears in the last part of Conclusion.

We have edited the outlook to make it more concise. (Line 33)

Lines 79-86: I suggest to clarify this part more.

From my point of view, the weather and climate has much coarser resolution than the size of lake. For instance, ERA5 resolutions is 0.1 *0.1. The lakes are just several grids by several grids.

in this sense, there are several potential draw backs:

1) coarse resolution may result in large errors;

2) this small lakes may potential lead to model instability.

therefore, I think the weather forecast team adopted this kind of parameterized Flake. it is fast and probably not cause model instability.

If the model resolution is fine enough, I think the ocean-sea ice module can be directly applied in these regions, as in the Arctic and Antarctic region

Thank you for your comment. We have worked to clarify this section by addressing the limitation with more specificity, namely that, while at coarse resolutions 1D models may perform well enough, they are not suited for finer spatial resolutions, where lake ice cover heterogeneity becomes a greater source of model instability. Line 80:

"At coarse resolutions (e.g 30x30 km), this approach works well. However, at finer scales (e.g. less than 10 km), 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)."

Line 123-124: please clarify here, you have changed it unit from 0-10 to 0-1, since you used ice fraction as your metrics.

We have added a line clarifying this (Line 123): "Finally, we rescaled these observations into ice cover fractions (0-1), for comparison to the other data sources."

Line 125: maybe use "exchanges", the changes of ice are definitely not "balance"

We have improved the clarity of this statement (Line 126): "…are fundamentally governed by a surplus or deficit in the heat energy exchanges with the atmosphere, heat stored in the water…"

Line 128: for me, I will probably use long wave-radiation. If you use cloud cover, you probably mess up the correlations. using cloud cover are directly correlated with long/shortwave radiation.more cloud cover reduces the solar radiation but increase long-wave radiation.

Thank you for your suggestions. We originally used cloud cover due to its inclusion in the 1-D lake ice models such as CLIMo.

Considering the variable importances we calculated, this variable was clearly uninformative for the model, so we expect this could be improved with alternative variables (e.g. shortwave radiation) in future work.

Line 135: Not sure whether this way or this explanation will lead to misleading for the future studies.

From my point of view, this kind of variables are related to the memory of the system. for instance, if ATDD is large, the lake is likely accumulate more heat, then, the sea ice is likely appears later. what if you use the air/water temperature or SIC for all the ATTD days as the input? it likely contains the same information. But the difference is that you only extract the accumulated temperature as the indicatior/ predicator.

I suggest to clarify all these things in the future study if you really want to apply this model to the weather and climate models.

Similar to the point above, ATDD and AFDD were selected based on their previous use in lake ice cover modelling work. We have clarified this paragraph based on your comments as well as Reviewer 1's (Line 136):
"To address seasonality in ice cover and heat storage, we introduced two additional temporally aggregated variables, Accumulated Freezing Degree Days (AFDD) and Accumulated Thawing Degree Days (ATDD). These variables have been shown to relate well to ice 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 air temperatures below (for AFDD) or above (for ATDD) zero degrees Celsius during their respective year-long period (starting February 1st for AFDD, and August 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."

Thank you for your comment and we will consider the additional variables you suggested in future versions of the model.

Line 138: please clarify temperature is "air" or "water" temperature

We have clarified this as air temperature (Line 139)

from the definitation "Days", I think AFFD, ATDD should be days. But from here, I think it is degrees? please clarify.

We have clarified the definitions of AFDD and ATDD: "They are defined as the sum of air temperatures below (for AFDD) or above (for ATDD) zero degrees Celsius during their respective year-long period (starting February 1st for AFDD, and August 1st for ATDD)" (Line 139)

Line 158: why? is it because you need to make all data to satellite obsertation frequency? or because of high computational requirement? for the method, I don't see the requirement.

This is a good point; one could technically leverage hourly forcing to predict daily ice-cover using our framework.

We opted to keep everything on the daily frequency for simplicity given this was our first attempt at developing such a model, and because it required much lower computational resources. It would certainly be interesting to see how the hourly forcing could enhance predictions in future work.

Line 166-167: Do you mean that you also take ice thickness variables in to your LIF-DLmodel ? if this is the case, you should say this in P155 rather than here. otherwise, it is confusion.

AS I understand, Flake is embeded in ERA5-land, the variables are avaliable from ERA5 web. so, it should belong to ERA5 datasets, rather than FLake.

The FLake data is available from ERA5 since it is embedded in ERA5-Land, as you pointed out.

Here, we were trying to distinguish between the forcing variables (which were used in LIF-DL) and the baseline model FLake (which was not incorporated into the LIF-DL framework).

By separating FLake into its own section, we could make it clear that the ice-thickness from ERA5 was not used as input to the LIF-DL model. Ice predictions from FLake were only used as a comparative baseline for our model.

Line 181 (Table 1): this defination is weird.

1)I have February 1-3 above 0 degree, then from 4-7 below 0 degree, then it has not effects on Sea ice process? is it accumulated values or mean values?

2) what is the differences between this one and 2-m airtemperature. I believe this value are extracted from 2-m air temperature alone, why youhave to reuse this variables twice?

Following our definitions from Section 2.1.2 – the ATDD and AFDD are the accumulated (sum) of the mean air temperature variable over their respective time-period, with the constraint that ATDD only accumulates temperatures above 0, while AFDD only accumulates temperatures below 0.

The goal was to capture "lake memory" over longer time horizons, such as a particularly warm summer (large ATDD) or particularly cold winter (large AFDD), which impact the timing of freeze-up and break-up.

We have adjusted the wording in the Table 1 slightly to clarify this (Line 183):

"AFDD: Accumulated (sum) of air temperatures below 0°C, starting from February 1st."

"ATDD: Accumulated (sum) of air temperatures above 0°C, starting from August 1st."

Line 187: what if you leave the land part. I guess the land bathymetry may also strongly impact the lake condition. for instance, high mountain may result in strong/weak wind, cold air intrusion. not sure why you mask out the land

Within the GLDB, land elevation is not represented, so all land pixels have a "lake-depth" of ~0 m. It could be very interesting though to integrate elevation data with bathymetry in future for the reasons you mentioned.

Line 194: Do you mean, you interpolated your data on 4km grid?

Yes – we adjusted the wording to say interpolated (Line 194)

Line 248 (Figure 4): Personally, I prefer to merge Figure 3 &4 to Figure 5, which is more informative and clean.

We have made significant adjustments to Section 2.4, following major comments from Reviewer 1. Please see the new figure (Figure 3, Line 220).

Line 272: not sure why excluded land piexls. the atmosphere conditions include weather impacts from the upstream. for instance, the cold fronts may happen couple days ago on the land pixes and propagate the lake region days later. In this case, include land pixels play a role.

The LIF-DL model is only tasked with predicting lake ice cover over the lake surface, and therefore, there are no outputs for land pixels. This is why the loss function masks land pixels.

Regarding your comment about forcing variables, these are not masked, and so the influence of atmospheric variables surrounding the lake body can be considered by the model for producing those lake-pixel predictions.

Line 350: May be this is called total ice area? trend is not an proper word here,

We have removed the word trend from this section. (Line 365)

Line 371 (Section 3.1) : is it possible draw a figure to show dependence of variable dependence against time, which is likely more informative.

Change applied – great suggestion! (Figure 4, Line 393)

The methods were updated to reflect this change:

"To investigate the importance of the variables relative to time, the gradients were summed and normalized by month, resulting in monthly variable importance estimates." (Line 323)

And the results section on variable importance was also updated to interpret the new figure (Line 394 - 439)

Line 400: you also say it is a black-box, it hard to judge whether or not is physical model.

While LIF-DL is a black-box model, it utilizes input variables with well-established physical links to lake ice dynamics. The objective of this section was to demonstrate that the model's data-driven feature importances align with physical intuition (e.g., the primary role of air temperature over precipitation). However, we acknowledge that feature importance metrics do not provide full transparency into the model's internal representations. To address this limitation, future work will focus on developing physics-informed hybrid models.

We expanded this comment in the revised section (Line 433-438):

"Since LIF-DL is purely data-driven, its learned relationships were not directly guided by physical priors. The fact that the model independently learned to proportionally prioritize atmospheric variables in close accordance with established scientific literature is a strong validation of its architecture. However, the authors acknowledge that the black-box nature of the LIF-DL makes it difficult to conclude these relationships with absolute certainty. Future iterations of this work will look to explicitly incorporate physical processes into the LIF-DL framework to ground its understanding of the system, prevent mathematical artifacts, and improve model interpretability."

Section 3.3.1

the explanation and description are not good here. I have no ideas which figures I should focus on. throughout the following 3 paragraphy, only one statement (l450) id point to Fig8.

I cannot see where the related rivers are.

From Figure 8 &9, I feel like F-lake ERA5 is better.

and Section 3.3.2:

The same problem as in 3.3.1.

Think about the following structure "The ... is slightly smaller in .. than in Flake (Figure 9)

Otherwise, I can hardly follow.

From Figure 9 & 10 I can see your method is slighter better than FLAe-era5 .

We have made significant updates to these sections to improve the readability and better explain the plots.

Changes were made so that Figures appeared closer to their first mention, and specific interpretations of each figure were separated to improve the readability.

Regarding your specific confusion with model performance – Figures 8 and 10 (now Figures 5 & 7, Lines 489 & 530) display the specific timing anomalies within each data source (modelled or IMS), as opposed to a global average across all sources. Therefore, high performance is indicated by similar patterns of freeze-up or break-up between a model and the observations (IMS), rather than by absolute timing anomaly being small.

The relevant performance metrics (SSIM and KT-B) are provided in the Figures, which are calculated from comparing the timing anomalies of each model prediction against IMS anomalies. These demonstrate the superiority of the LIF-DL for break-up start (higher SSIM and KT-B scores), and are reported in the text. These methods are explained in Section 2.6.3.

To make this clearer and avoid confusion, wording was update in the figure interpretation paragraphs.

Line 528 (Table 4): shoud highlight the best one with bold font. Also, please properly cite the table in the text.

The best performing model was bolded in this and subsequent tables.

Line 538: (Figure 11 and table 4). here is an example of how to cite your figure and table

We have added bolded font to the best models and added references to figures/tables for clarity.

Line 547: put an a,b,c,d in front of the lake name.

Thank you for the suggestion, but we feel that the lake name is sufficiently clear for referring to each specific subplot within this Figure.

Line 550: have any idea why the "spikes" occur. usually, deep learning give smoother results.

This is a good question, however, further analysis is still needed to understand the full reasoning.

Our initial thinking was that it is related to lake depth and the small number of study sites used. It appears to be worst for Lake Winnipeg and Reindeer Lake, both of which are shallower than the other study sites. So it could be related to the LIF-DL struggling to capture ice dynamics for this particular case. Another possible reason is generalizability to unseen data. It's possible that the 2019 and 2020 years (which exhibited the greatest noise in LIF-DL predictions) represented more anomalous forcing conditions, and that the model could not accurately adapt to these based on its training data, leading to chaotic behaviour.

Line 566 (table 5): highlight the best one

Highlights have been made

Line 582 (Conclusions):

conclusion part is not perfect.

this part should not be just a compilation of the numbers and the metrics.

If you want to apply the LIF-FL model to forecast model or climate mode, 1) you should clarify what these model needed from your model. what is the major challenges.

then you developed the a machine-learning model and tested it.

then, what challenges yu have over-come and what remains.

From your conclusion, I feel like you have a LIF-DL model ready for forecast and climate model. But I think it is not really at that stage. I suggest to conclude the metric part again. for instance, we use BUS to measure the ... error. we note that ...

don't jjust list the metric and numbers. it make no sense.

Also. the conclusion should be objective. I see in the results part, LIF-DL is not always the best. it also show some weird results like the spikes. so it the time to disscuss on this problem here,

We have made significant improvements to the conclusions based on your feedback. Overall, we have shortened the discussion about results and added more detail on limitations and future work. It has also been made more concise to concentrate its focus. Please refer to the revised manuscript.

Line 586: What is the meaning of one-dimensional model and ice cover dynamics?

From my understanding, one dimension mean only time? it has no information on lat-lon. probably, in ERA5-, the model give one value of fractionv(as in figure 11). However, Figure 8-9 gives lon-lat error distributions. I don't knew what on-dimensional model here.

When referring to one-dimensional models, we specifically mean models which cannot leverage spatial interactions, and can only make predictions for a single point. In the case of FLake, gridded predictions are achieved by applying the model independently at every grid cell in ERA5-Land, hence the "spatial" predictions observed in Figure 11 you mention.

2) ice dynamic, is it also include thermodynamic? As far as I know, ice dynamic usually means ice drift, convergence/divergence. but in you study, all these processes are not mentioned at all. please clarify.

When saying ice dynamics, we meant the spatial patterns in ice formation / melt across the lakes surface. The physical timing/location of ice cover across a lake is greatly influence by that particular lakes morphometry, which cannot be captured by 1D models that are applied across grid cells independently.

Line 589: physically meaningful: I think there are no proof that you ML model is physical meaningful. it is still based on statistical . Flake includes explicit physical process through simplified.

We have modified this so say:

"Variable importance analysis indicated that LIF-DL is most sensitive to forcing variables like seasonally accumulated air temperatures, solar radiation and lake depth, which follows expectations from literature (Williams et al., 2004; Williams, 1965)." (Line 651)

Line 625: this paragraphy should be move to backward and merge with the last paragraphy.

Conclusions were significantly changed to render it more concise and accurate. Please see the revised manuscript, Lines 635 – 689.