1) Abstract should be most informative. The abstract discusses ice jam and prediction necessity in the first half. The reader should also be able to obtain information from the data, modeling process, and validation metrics.

We explained and added that: "The hydro-meteorological variables (e.g., temperature, precipitation, and snow depth) along with the corresponding jam or no-jam events are used as the inputs to the models. We hold out 10% of the data for testing. And we applied 100 re-shuffling and splitting iterations with 80 % of the remaining data for training and 20% for validation. The results show that the CNN-LSTM model yields the best results with the validation and test F1 scores of 0.82 and 0.92, respectively."

2) Acronyms (e.g., CNN, LSTM, CN-LSTM, F1) should be defined in the first use in the abstract.

We explained them in the abstract, except for F1 score that is not an acronym.

3) Provide a reference for empirical and statistical prediction methods (threshold methods, multi-regression models, logistic regression models, and discriminant function analysis).

We added :"(Barnes-Svarney and Montz, 1985; Mahabir et al., 2006; Massie et al., 2002; White, 2003; White and Daly, 2002, January; Zhao et al., 2012)".

4) Introduce "involved hydro-meteorological variables" in line 24.

We added: "involved hydro-meteorological variables (e.g., temperature, precipitation, snow depth, and solar radiation)".

And later explained all the variables in detail in "Data and study area" section.

5) Why did you choose deep learning over other machine learning models for ice jam prediction? Is that all due to an automatic feature selection? Clarification is needed.

We added that: "There are some machine learning methods available for TSC such as K nearest neighbour and support vector machine. However, the focus of this research is on the deep learning models that have greatly impacted sequence classification problems and they can also be used for multivariate TSC with good performance. Deep learning methods are able to consider two-dimensionality in multivariate time-series and their deeper architecture could further improve the classification especially for complex problems, which is why their results are more accurate and robust than other methods (Wu et al., 2018a, April). However, they are more time consuming and difficult to interpret."

And also later we added in the text that: "Deep learning models perform well for the problems with complex-nonlinear dependencies, time dependencies, and multivariate inputs."

Also we developed machine learning models and compared their performance with developed deep learning models in the text. The validation and test F1 scores of Machine learning models are presented in Tables 10 and 12, showing that deep learning performed much better than machine learning in ice-jam prediction.

Table 10. F1 scores of validation for SVM, DT, and KNN and MLP models with 100 random train-validation splits.

Models	F1 score		
	mean	max	min

SVM	0.76	0.82	0.69
DT	0.74	0.80	0.67
KNN	0.75	0.84	0.68
MLP	0.75	0.83	0.68

Table 12. Test F1 scores for SVM, DT, and KNN and MLP models.

Models	F1 score
SVM	0.75
DT	0.71
KNN	0.78
MLP	0.70

6) There is a lot of focus on time series predictions in the literature review while you should be more specific about ice jam prediction. The literature should at least include data-driven models for predicting ice jams.

The ice-jam prediction is considered as a multivariate time-series classification problem. The focus of our previous paper was on the literature review on ice jam prediction methods. As we mentioned in the previous question, we developed machine learning methods and compared the results with the results of the deep learning models, as you suggested.

7) In lines 97-98: the authors state "Deep learning methods are promising to address the requirements of ice jam predictions." Is there any research to use deep learning for ice jam prediction? If so, what is the contribution to the current research?

No there is no research for that. You are right so we changed the phrases to address your comment:" Although deep learning methods seem to be promising to address the requirements of ice-jam predictions, none of these methods yet have been explored for ice jam prediction."

8) Although there are several deep learning methods, why did you select CNN, LSTM, and a combined CNN-LSTM?

we added that "Through our comprehensive literature review, we noticed that CNN (e.g., Brunel et al., 2019; Cui et al., 2016; Devineau et al., 2018, June; Kashiparekh, 2019, July; Nosratabadi et al., 2020;Yan et al., 2020; Yang et al., 2015, June; Yi et al., 2017; Zheng et al., 2016), LSTM (e.g., Fischer and Krauss, 2018; Lipton et al., 2015; Nosratabadi et al., 2020; Torres et al., 2021), and a combined CNN-LSTM (e.g., Karim et al., 2o17; Livieris et al., 2020; Ordóñez and Roggen, 2016; Sainath et al., 2015, April; Xingjian et al., 2015) have been widely used for TSC.

9) The authors consider 0 value for "NaN" precipitation values. 0 value means there is no precipitation. However, there might be precipitation and there is no measurement. Since the modeling is based on time series, it is better to impute missing values instead of considering 0 values.

You are right. We removed that phrase from the text as although there are some NaN records in the database for precipitation, we did not use them in our study.

10) The text contains several typos.

We tried to edit and remove them.

11) Provide a reference for the statement "The most popular deep neural networks for TSC are MLP, CNNs, and LSTM."

We added: "(Brownlee, 2018; and Torres et al., 2021)".

12) The sub-heading "Input data and study area" is not appropriate. The reader may think you mean input variables of the model. Sub-heading "Data and study area" should be suitable.

We agree. We changed that.

13) Detailed information, at least as a statistical description, should be provided for the data used in the study.

We added a Table for that.

Table 1. Statistics of hydro-meteorological variables used in the models.

Statistics	Liquid P (mm)	Tmin (°C)	Tmax (°C)	Net radiation (W m-2)	ATDD (°C)	AFDD (°C)	Snowdepth (cm)
min	0.00	-40.00	-25.97	-67.77	0.00	-2109.33	0.00
max	50.87	12.05	27.48	222.69	280.82	-35.41	121.86
average	1.04	-9.41	0.98	59.75	8.83	-898.48	15.99
median	0.00	-7.73	1.68	59.41	1.27	-890.74	11.50

14) Provide a reference "As a benchmark, a CNN model with the parameters and layers similar to previous studies is developed."

We added: "(e.g., Ordóñez and Roggen, 2016)."

15) Section 2.5.1. Overcome overfitting is too long. It needs to be shortened.

We removed all subsections and tried to make this part shorter. We also moved some parts from Results to "Overcome overfitting" section. We moved some explanations of methods that are used to the Appendix. There are now 14 lines in "Overcome overfitting" section in the text.

16) It is not clear how you optimize the structure of the model. Did you use any hyperparameter tuning method e.g. GridSearch, random search, Bayesian optimization? Or only a trial and errors approach?

We were aware of other methods for hyperparameter tuning but we selected a manual trial and error method, as grid search experiments suffer from poor coverage in dimensions (Bergstra and Bengio, 2012) and manual experiments are much easier and more interpretable in investigating the effect of one hyperparameter of interest.

17) The training chapter "2.5 Training" has focused only on the general statements of the modeling. The authors should give more information about the modeling process, optimum values for the parameters, loos function, etc.

We restructured that part. We added Table 3 for optimum values of parameters.

We moved subsections "Activation function", "Learning rate", "Padding", "Activation functions in CN layers", "Dense layer", "Network optimization", and "Update expression" to Appendix.

Table 2. Common values and selected values for different parameters of the models.

Parameter	Common values	Selected value
Mini-batch size	16, 32, 64	16
Number of convolution filters	32, 64, 128	128
Filter size	3, 5, 7	(5,1) and (5,3)
Number of LSTM units	32, 64, 128	128
Number of dense layer units	16, 32, 128, 256	32
Momentum in SGD	0.5, 0.99, 0.9	0.9

18) Provided Information in "3.1 Hyperparameters optimization" is not a result! It is a model development and should be moved to the materials and method chapter as a "model development" sub-chapter. The structure of the manuscript needs to be revised.

We moved that to materials and method section.

19) I suggest developing a traditional machine learning model such as MLP, random forest, … to show the efficiency of the developed deep learning models.

As We explained above, I developed SVM, KNN, decision tree, and multilayer perceptron models as the common machine learning models for TSC and compared the results with deep learning models.

20) Abbreviations should be checked throughout the text. Different abbreviations (e.g. CN-LSTM, CNN-LSTM, CNLSTM) has been used.

Done. We selected CNN-LSTM.

21) Discussion should be improved significantly. The authors have focused only on the structure of the models. The results should also be connected to the physical characteristics of the ice jam. How does the model can help manage the water resources? Compare the results with the literature, not a simple report.

A) In terms of relation to the physical system, we added a section "Discussion on the interpretability of deep learning models" and later explained that we will cover this issue in our future work:

"Even though the developed deep learning models performed pretty well in predicting ice jams in Quebec, the interpretability of the results with respect to the physical processes of the ice jam is still essential. It is because although deep learning models have achieved superior performance in various tasks, these really complicated models with a large number of parameters might exhibit unexpected behaviours (Samek et al., 2017 & Zhang et al., 2021). This is because the real-world environment is still much more complex. Furthermore, the models may learn some spurious correlations in the data and make correct predictions with the 'wrong' reason (Samek and Müller, 2019). Hence, interpretability is especially important in some real-world applications like flood and ice-jam predictions where an error may cause catastrophic results. Also, interpretability can be used to extract novel domain knowledge and hidden laws of nature in the research fields with limited domain knowledge (Alipanahi et al., 2015) like ice-jam prediction.

However, the nested non-linear structure and the "black box" nature of deep neural networks make interpretability of their underlying mechanisms and their decisions a significant challenge (Montavon et al., 2018, Zhang et al., 2021 and Wojtas and Chen, 2020). That is why, interpretability of deep neural networks still remains a young and emerging field of research. Nevertheless, there are various methods available to facilitate understanding of decisions made by a deep learning model such as feature importance ranking, sensitivity analysis, layer-wise relevance propagation, and the global surrogate model. However, the interpretability of developed deep learning models for ice-jam prediction is beyond the scope of this study and it will be investigated in our future works."

B) In terms of comparison of the results with the literature, the present work is the first application of the comprehensive ice-jam database for Quebec. To the knowledge of the authors previous studies for ice-jam prediction in Quebec consider only specific locations in a specific river for just a year or two years, hence there is no literature available to compare our ice jam predictions with.

C) How does the model can help manage the water resources?

To address your question, we added a section to Discussion "Model transferability":

"The transferability of a model between river basins is highly desirable but has not yet been achieved because most river ice-jam models are site specific (Mahabir et al., 2007). The developed models in this study can be used to predict future ice jams some days before the event not only for Quebec but also for eastern parts of Ontario and western New Brunswick. For other locations, the developed models can be transferred via retraining and a small amount of fine-tuning using labeled instances, rather than building from scratch. It is because the logic in the model may be transferable to another site with small modifications. To transfer a model from one river basin to another, historic records of ice jams and equivalent hydro-meteorological variables (e.g., precipitation, temperature, and snow depth) as inputs to the model must be available at each site."

22) Conclusion needs to be revised significantly. You should only give the main findings of the research here. Providing reference is not recommended in the conclusion section.

We removed the references and re-wrote the Conclusion:

"**Conclusion**

The main finding from this project is that all the developed deep models performed pretty well and performed much better than the developed machine learning models for ice-jam prediction in Quebec. The comparison of results show that the CNN-LSTM model is superior to the CNN-only and LSTM-only networks in both validation and testing accuracy, though the LSTM and CNN models demonstrate quite good performance.

To our best knowledge, this study is the first study introducing these deep learning models to the problem of ice-jam prediction. The developed models are promising to be used to predict future ice jams in Quebec and in other river basins in Canada with re-training and a small amount of fine-tuning.

The developed models do not apply to freeze-up jams that occur in early winter and are based on different processes than breakup jams. We studied only breakup ice jams as usually they result in flooding and are more dangerous than freeze-up jams. Furthermore, there is a lack of data availability for freeze-up ice jams in Quebec and only 89 records of freeze-up jams are available which is too small.

The main limitation of this study is data availability as recorded ice jams are small which causes deep learning models to easily overfit to small number of data. Another limitation of the presented work is the lack of interpretability of the results with respect to the physical characteristics of the ice jam. This is a topic of future research and our next step is to explore that.

The hydro-meteorological variables are not the only drivers of ice-jam formation. The geomorphological indicators that control the formation of ice jams include the river slope, sinuosity, a barrier such as an island or a bridge, narrowing of the channel, and confluence of rivers. In the future, a geospatial model using deep learning will be developed to examine the impacts of these geospatial parameters on ice-jam formation."

23) Provide the limitations of the research along with the recommendations for the future studies.

We added that to the Conclusion (above).

24) The mean error of training in LSTM is much higher than validation (middle plot in figure 16) which might be a sign of overfitting.

We added some reasons for that: "There are some reasons explaining that. Regularization reduces the validation loss at the expense of increasing training loss. The regularization techniques such as noise layers are only applied during training, but not during validation resulting in more smooth and usually better functions in validation. There is no noise layer in CNN-LSTM model that may cause a lower training error than validation error. However, other regularization methods such as L2 regularization are used in all the models, including the CNN-LSTM model.

Furthermore, the other issue is that batch normalization uses the mean and variance of each batch in training, whereas, in validation, it uses the mean and variance of the whole training dataset. Plus, training loss is averaged over each epoch, while validation loss is calculated after each epoch once the current training epoch is completed. Hence, the training loss includes error calculations with fewer updates."

References

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831-838.

Barnes-Svarney, P. L., & Montz, B. E. (1985). An ice jam prediction model as a tool in floodplain management. Water Resources Research, 21(2), 256-260.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Brunel, A., Pasquet, J., PASQUET, J., Rodriguez, N., Comby, F., Fouchez, D., & Chaumont, M. (2019). A CNN adapted to time series for the classification of Supernovae. Electronic Imaging, 2019(14), 90-1.

Cui, Z., Chen, W., & Chen, Y. (2016). Multi-scale convolutional neural networks for time series classification. arXiv preprint arXiv:1603.06995.

Devineau, G., Xi, W., Moutarde, F., & Yang, J. (2018, June). Convolutional neural networks for multivariate time series classification using both inter-and intra-channel parallel convolutions. In Reconnaissance des Formes, Image, Apprentissage et Perception (RFIAP'2018).

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654-669.

Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2017). LSTM fully convolutional networks for time series classification. *IEEE access*, *6*, 1662-1669.

Kashiparekh, K., Narwariya, J., Malhotra, P., Vig, L., & Shroff, G. (2019, July). ConvTimeNet: A pre-trained deep convolutional neural network for time series classification. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. Neural computing and applications, 32(23), 17351-17360.

Mahabir, C., Hicks, F. E., & Fayek, A. R. (2007). Transferability of a neuro-fuzzy river ice jam flood forecasting model. Cold Regions Science and Technology, 48(3), 188-201.

Mahabir, C., Hicks, F., & Fayek, A. R. (2006). Neuro-fuzzy river ice breakup forecasting system. Cold regions science and technology, 46(2), 100-112.

Massie, D.D., White, K.D., Daly, S.F., 2002. Application of neural networks to predict ice jam occurrence. Cold Reg. Sci. Technol. 35 (2), 115–122.

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1-15.

Nosratabadi, S., Mosavi, A., Duan, P., Ghamisi, P., Filip, F., Band, S. S., ... & Gandomi, A. H. (2020). Data science in economics: comprehensive review of advanced machine learning and deep learning methods. Mathematics, 8(10), 1799.

Ordóñez, F. J., & Roggen, D. (2016). Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. Sensors, 16(1), 115.

Sainath, T. N., Vinyals, O., Senior, A., & Sak, H. (2015, April). Convolutional, long short-term memory, fully connected deep neural networks. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4580-4584). IEEE.

Samek, W., & Müller, K. R. (2019). Towards explainable artificial intelligence. In Explainable AI: interpreting, explaining and visualizing deep learning (pp. 5-22). Springer, Cham.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Torres, J. F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., & Troncoso, A. (2021). Deep Learning for Time Series Forecasting: A Survey. Big Data, 9(1), 3-21.

White, K. D. (2003). Review of prediction methods for breakup ice jams. Canadian Journal of Civil Engineering, 30(1), 89-100.

White, K. D., & Daly, S. F. (2002, January). Predicting ice jams with discriminant function analysis. In ASME 2002 21st International Conference on Offshore Mechanics and Arctic Engineering (pp. 683-690). American Society of Mechanical Engineers.

Wojtas, M., & Chen, K. (2020). Feature importance ranking for deep learning. arXiv preprint arXiv:2010.08973.

Wu, J., Yao, L., & Liu, B. (2018a, April). An overview on feature-based classification algorithms for multivariate time series. In 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 32-38). IEEE.

Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems (pp. 802-810).

Yan, J., Mu, L., Wang, L., Ranjan, R., & Zomaya, A. Y. (2020). Temporal convolutional networks for the advance prediction of ENSO. Scientific reports, 10(1), 1-15.

Yang, J., Nguyen, M. N., San, P. P., Li, X. L., & Krishnaswamy, S. (2015, June). Deep convolutional neural networks on multichannel time series for human activity recognition. In Twenty-fourth international joint conference on artificial intelligence.

Yi, S., Ju, J., Yoon, M. K., & Choi, J. (2017). Grouped convolutional neural networks for multivariate time series. arXiv preprint arXiv:1703.09938.

Zhang, Y., Tiňo, P., Leonardis, A., & Tang, K. (2021). A survey on neural network interpretability. IEEE Transactions on Emerging Topics in Computational Intelligence.

Zhao, L., Hicks, F. E., & Fayek, A. R. (2012). Applicability of multilayer feed-forward neural networks to model the onset of river breakup. Cold Regions Science and Technology, 70, 32-42.