Response to Referee 1

General comments:

The idea of this study is to develop a regional streamflow model using a convolutional long short-term memory artificial neural network, which is the merger of two distinct deep learning (DL) techniques. This and several other innovations presented in the paper are quite impressive, and the overall performance of the model seems good. The revised paper is also in much better shape than the original submission, with considerably more detail given, much better figures, and some significant new analysis to shore up some weak points in the original study, such as including a linear benchmark model for comparison and additional sensitivity analyses demonstrating physically reasonable responses to perturbations in temperature fields, which (to some degree) ties into broader goals like explainable machine learning.

We thank the referee for their detailed review and are pleased they find the revised manuscript to be a substantial improvement.

Unfortunately, the quality of the writing and explanations remains somewhat inadequate. The general impression one receives when reading this manuscript (which may or may not be true, but it is the impression one gets from the writing) is that the authors have some background with areas of geophysical science adjacent to watershed hydrology, but not watershed hydrology itself, and certainly not any aspect of operational hydrology or streamflow modeling. Similarly, treatments of machine learning in the manuscript seem to suggest familiarity with a very narrow range of sophisticated techniques but not a great awareness of the overall field of machine learning and, in particular, prior work on its application to streamflow modeling. Sadly, this may only serve to reinforce negative impressions among the water resource community as a whole about the general usefulness and credibility of machine learning – impressions that have been crippling in some important ways to the advancement of the field of hydrology.

We have addressed the detailed comments below to improve the quality of writing and explanations. Additionally, we have had Dr. William Hsieh review the manuscript with the goal of identifying shortcomings related to the quality of writing and explanations, and have implemented his suggestions for improvement.

Additionally, the primary technical innovation presented here – taking the long short-term memory (LSTM) neural network from time series analysis, which has recently seen several high-profile research applications to streamflow modeling, and adding in a convolutional neural network (CNN) from image analysis – involves an incremental advance over the existing LSTM approach, yet the existing LSTM approach is never implemented here. As a result, the paper cannot provide information on how much of an advantage, if any, the addition of a CNN architecture provides. Perhaps this is not strictly needed for publication, but it is an obvious limitation of the study that may compromise the adoption of this novel streamflow modeling technique by others.

Overall, this manuscript contains many clever and potentially powerful ideas but seems to be poorly executed, and it feels like an appropriate recommendation is for publication pending major revisions.

We are glad the referee agrees that this study is built around many clever and potentially powerful ideas. We address all comments below, with new text in the manuscript identified in **bold blue text**.

Detailed comments:

Line 10, delete "the region of"

Deleted.

Lines 31-32, not clear what the authors mean by a spatially distributed DL model. Use of the concepts of lumped, semi-distributed, and fully distributed hydrologic models has been pretty much exclusive to process-based models and it's not clear how it can be extended to ML-based hydrologic models. Reading the rest of the paper, I can make an educated guess as to what the authors trying to get at here, but the reader should not have to do this and it's still not clear that the terms really apply as such to machine learning models in hydrology. Are they referring to multiple forecast points (corresponding to gages)? If so, that's captured in the concept of a regional hydrology model, which is not necessarily the same thing as a fully distributed hydrology model (as many fully distributed models make predictions at only a single location for example). Are they referring to fully spatially distributed (e.g. gridded) inputs? If so, that was successfully tackled decades ago in ML-based streamflow modeling by Hsieh et al. (2003). Moreover, the regional DL models with many input data and output prediction locations introduced by Kratzert et al. (2018, 2019a, 2019b) feel like they may be just as spatially distributed as the DL model introduced in this submission. The authors need to be much clearer and more specific on what they mean here and consider whether the confusion created by this mixing-and-matching of terminology is really beneficial to their ultimate purpose and the clarity and credibility of this submission. I suspect this is another case (the problem was widespread in the original submission) of slightly misusing standard hydrologic nomenclature. That said, see comment below re: line 137, where the manuscript handles all this much better.

We have edited this section as follows:

Line 30: "These recent DL-based studies have emphasized the development of lumped hydrological models **with inputs that are aggregated to the basin-level**. However, **fewer DL-based studies have explored the use of spatially discretized forcing and geophysical data** (Gauch and Lin, 2020). In contrast, traditional process-based approaches have made

substantial progress towards distributed hydrological models **which are driven by spatially discretized inputs** (Freeze and Harlan, 1969; Marsh et al., 2020; Pomeroy et al., 2007)."

Furthermore, throughout the manuscript we reserve "distributed" to describe distributed process-based models, and refer to DL-based models forced by spatially discretized inputs as a "**DL analogue of a distributed hydrological model**" or a "**DL model that is driven by spatially discretized forcing data**".

Line 41: replace "total April-August streamflow" with "seasonal water supply", which was the point of the exercise and is a major, mainstream task in water resource forecasting and management.

Changed to "**seasonal water supply**".

Lines 45-48: this basic description of ML in hydrology is clunky and imprecise. It could be easily read to imply the authors think that a Bayesian neural network is not an ANN, or that they think ANNs aren't non-deep (in truth, traditional feedforward-error backpropagation ANNs of the sort being referred to here may by deep or non-deep depending on how many hidden layers they have), etc. Mistakes like this up-front in the introductory section may immediately draw the paper's basic credibility into immediate question, no matter how innovative and correct the actual technical work subsequently presented in the paper may be.

We have rephrased this passage to more precisely introduce machine learning applications in Western Canada:

Previously: "In addition to ANNs, which have received particular attention in hydrology (Maier et al., 2010; Maier and Dandy, 2000), numerous types of non-deep machine learning applications have also been developed for hydrometeorological analyses, and in particular, many have been developed for applications in Western Canada.

Updated, Line 46: "**In particular, numerous types of machine learning applications have been developed for hydrometeorological analyses and applications in Western Canada.**"

Line 53: here the authors are implying that ANNs are non-deep, whereas that may or may not be the case (see preceding comment). This error is just sloppy writing and is totally avoidable. Again, the overall impression one gets from these passages is that the authors are not very familiar or comfortable with the field of ML in general, which is not a helpful image to present to the reader.

We edit this sentence to remove this implication:

Previously: "While ANNs and other non-deep machine learning architectures have a long history and continue to find useful applications in hydrology…"

Updated, Line 53: “While **such** machine learning architectures have a long history and continue to find useful applications in hydrology…”

Lines 56-57, comment about advantages of deep learning relative to "labour-intensive manual feature extraction often required for non-deep models" - essentially true but also substantially exaggerated, which again undermines the credibility of the manuscript. Automated predictor selection and feature creation techniques have been used in statistical modeling for decades and have appeared in non-deep machine learning too. A recent example is Fleming et al. (2021b).

The clause “in contrast to labour-intensive manual feature extraction often required for non-deep models” is removed.

Lines 70-75: good description, but might want to consider mentioning here that Kratzert et al. (2019a) additionally used spatially heterogeneous physical basin characteristics as predictors in regional LSTM models. I believe this may be mentioned later in the manuscript but ought to be briefly pointed out here.

We have edited the passage as follows:

Line 70: “LSTM models trained on many basins have been shown to outperform standard hydrological models for prediction at ungauged basins, **and the inclusion of physical basin characteristics as predictors further improved the LSTM model performance**, demonstrating the potential for LSTM models to be used as regional hydrological models (Kratzert et al., 2019a). However, while addressing the need to learn complex sequential information, the LSTM approach does not explicitly **learn from spatially discretized information**, and as such has been primarily used for lumped hydrological modelling.”

Line 80: beach state classification in coastal geomorphology is another example; see Ellenson et al. (2020).

Line 81: “… **and beach state classification (Ellenson et al., 2020).**”

The introductory section’s discussion of explainability in machine learning is inadequate and under-referenced, especially from the viewpoint of socially relevant hydrologic model applications, i.e., things such as actual flood and water supply forecasting at government agencies and the like. At a minimum, on line 100, after the sentence ending with “making”, add the following: “Practical methods are beginning to appear that allow users to easily identify and geophysically interpret, in detail, spatiotemporal patterns or input-output relationships identified by, respectively, new unsupervised learning (e.g., Fleming et al., 2021a) and supervised learning (e.g., Fleming et al., 2021b) algorithms designed for applied operational hydrological modelling environments where interpretability is key. However, there is still much work to be done on developing new and better ways to further the goal of

explainable machine learning for hydrology, in both deep and non-deep contexts and both operations and research settings." Both of these cited manuscripts describe new but non-deep ML methods that are far more focused on, and successful at, providing extensive and complete geophysical interpretations than the method introduced in this submission or other deep learning work in hydrology so far.

The suggested text has been added.

Line 107: add "or practical" after "research" and before "questions"

The suggested text has been added.

Line 137: yes, nicely done! Compared to lines 31-32 (see comment above), this is a much better description of what's being meant by a "distributed" model in the context of DL in this paper, though it's still not clear that using the term to described the CNN-LSTM application is particularly helpful.

We thank the referee for the positive feedback. We are more careful through the text to refer to the CNN-LSTM model not as being distributed, but rather as being a DL model that uses spatially discretized forcing data as input.

Point 2 on line 140, the authors should modify the text slightly to be explicit that they're referring to the spatially distributed input data

This point has been edited as:

Line 145: "2) investigate if the model has learned to focus on the areas **of the spatially distributed input data** that are within or near the watersheds where streamflow is being predicted"

Line 153: this region is never referred to by anyone strong familiar with it as "the south-central domain of western Canada" and this description doesn't even make geographic sense (what's "central" about it?). Ironically, statements like this are likely to undermine the credibility of the work specifically with hydrologists working in the paper's study area. Maybe try just calling it what it is: "southwestern Canada" and/or "the southern portions of the western Canadian provinces of British Columbia and Alberta".

Line 161: Rephrased as "**southwestern Canada**".

Line 187, after "acceptable" insert "for the purposes of this study"

Line 193: We have added "**for the purposes of this study**".

Lines 205 and 207, "maximized" is not quite the right word here; the term suggests optimization

Line 213: We have changed "maximized" to "**highest**", consistent with Eaton and Moore (2010).

Line 210, is "uniquely" the right word here? As described here, this is not unique to this region, as similar processes happen in other parts of the Canadian Prairies and presumably elsewhere as well. It's the only part of the study region with that characteristic, though, which maybe is what the authors are trying to say?

The word "uniquely" has been removed.

Line 224: the standard statistical nomenclature is "unit variance" not "unity variance"

This change has been made throughout the text.

Line 331: insert "(from which CNN technologies primarily originated)" after "image processing"

This text has been added.

Line 331: I'm not confident hydroclimatologists would view this type of work as "hydro-climatic modelling"

We have edited this sentence:

Line 346: "To ensure consistency between terminology in both image processing **(from which CNN technologies primarily originated)** and **this study**…"

Line 332: after "the three weather predictors", should "at all grid cells" be added to further clarify? I think that's what's meant here, but it needs to be entirely clear given the analogies being drawn here between video processing and spatiotemporal climate fields

Line 349: We have added "**at all grid cells**" for clarity.

Lines 370-375: this is mostly sound logic, except perhaps for the PDO, which is generally thought to typically remain in one state for a couple decades or so between regime shifts

The reference to the PDO has been removed.

Section 4.3.1: this is the first place in the manuscript that ensemble modeling is introduced, and strangely, no effort is made to explain here or anywhere else in the paper how that ensemble is formed. Between this treatment (or lack thereof) in the manuscript on the one hand and their rebuttal letter on the other hand, it's not clear the authors are aware that

there is more than one way to create an ensemble of ML models, and they do need to provide a brief explanation of how they did it (one sentence will suffice).

We clarify how we generate this ensemble:

Line 416: "1. Bulk training: **a CNN-LSTM model is initialized with random weights and is then** trained on all 226 stream gauge stations in the region."

Removed: "In total, we train an ensemble of 10 bulk models and further fine-tune each one, yielding an ensemble of 10 bulk models and an ensemble of 10 fine-tuned models per cluster."

Line 427: Replaced with: "**We initialize 10 bulk models with 10 different sets of random weights. Each bulk model is trained and then fine-tuned on each cluster of stream gauge stations, creating 10 fine-tuned CNN-LSTM models per each of the six clusters of stream gauge stations. We use this ensemble of 10 bulk models and 10 fine-tuned models (per cluster) for our analysis**."

Line 460: "Gaussian distribution" not just "gaussian" which is informal slang

This edit has been made where needed.

Lines 505-507: temperature degree-day models go a lot further back than this; provide more several more references, including references specific to the use of this type of snow sub-models within standard watershed hydrology models

We have updated this section as follows:

Line 529: "While the assumption is a simplification of processes dictated by the surface energy balance, the use of positive temperatures as successful indicators for the warming and melting of snow is a common assumption of positive-degree-day models in simulating snow and glacier melt **and was first used by Finsterwalder and Schunk (1887). Such positive-degree-day models have since been widely applied for modelling snow and glacier melt across multiple spatial scales (e.g. Hoinkesand and Steinacker, 1975; Braithwaite, 1995; Hock, 2003; Radic et al., 2014), and have been used in watershed hydrology models such as the UBC watershed model (Quick and Pipes, 1977) and the HBV-model (Bergström, 1976).**"

Figure 6: it is interesting that the overall test-phase NSE reported here for the Englishman River is substantially lower than that for the ensemble non-deep ANN for this same river described by Fleming et al. (2015). This may suggest that the advantage of simultaneous regional modeling across a large domain by the CNN-LSTM network introduced in this paper is accompanied by the disadvantage of weaker performance on a single given river of particular interest, which is often what water resource professionals are primarily concerned with – a particular river with socially destructive flooding events, for example, or a tributary to a reservoir that requires inflow predictions. This result might be traceable, at least in part,

to the benefits of making specific choices, on the basis of general expertise in physical hydrology and familiarity with the particular watershed in question, that can be easily made when modeling a single watershed but are more cumbersome in a regional model. For example, the Englishman River-specific ensemble non-deep ANNs of Fleming et al. (2015) included snow pillow and antecedent streamflow inputs as inputs. That does not imply that the submitted study has done anything wrong – on the contrary, the result likely reflects an expected trade-off between scale and detail in a modeling system. The paper should explicitly acknowledge this point using the Englishman River, and the comparison to previous AI work in that basin, as what appears to be a clear example. For these reasons, the Fleming et al. (2015) study should also obviously be added to Table 2, giving an additional point of comparison for the Englishman River, which is already included in that table but only for a very old study presenting a lower-performing process-based hydrologic model.

This is a good point raised by the referee, and we now explicitly discuss this comparison between the results for the Englishman River as in Fleming et al. (2015) and in our study:

Line 806: "**Prior studies have modelled daily streamflow at the Englishman River near Parksville (08HB002), one of the locations in our study; for example, Fleming et al. (2015) use an ensemble of ANNs to forecast streamflow and achieve NSE values in the range of 0.7 – 0.8, while Lima et al. (2016) use nonlinear extreme learning machines and achieve NSE > 0.8. These examples outperform the NSE value of 0.59 achieved by our CNN-LSTM. Their success could be in part due to the inclusion of more locally-specific input data (e.g. Fleming et al. (2015) include snow pillow and antecedent streamflow data, while Lima et al. (2016) include predictors such as sea level pressure, wind speed, and humidity, among others), a decision which can be more easily implemented for modelling at a single stream gauge station as compared to a regional scale model. These examples highlight what may be a trade-off between scale and detail in the modelling approach, where the advantage of simultaneous streamflow modelling at multiple stream gauge stations across a region as done by the CNN-LSTM network may be met by the disadvantage of weaker performance on one particular river of interest.**"

It is not obvious to us that Fleming et al. (2015), which uses an ANN, should be included in Table 2, which compares only with process-based models. As such we reserve this discussion for the main text.

Drop the term "heat map" from the manuscript. It's a standard term in graphics production, but in the context of a manuscript dealing with various geophysical quantities including temperature, it's unnecessarily ambiguous.

We have removed the term "heat map" and refer to this quantity throughout the manuscript as a "sensitivity map" or "S(x,y)" as defined in Equation 12.

Lines 624-626: can the authors offer a specific hypothesis or two why the eastern and northeastern clusters show such a strong sensitivity to coastal conditions? Could it perhaps

reflect some meteorological setup, e.g., jet stream position, storm tracks, etc.? It's a very prominent feature of the results.

We add:

Line 654: "**Another possible explanation is that there could be temporal patterns of sensitivity. For example, the eastern and north-eastern regions may be sensitive to coastal conditions when storms travel from west to east. Alternatively, the sensitivity maps may be most sensitive to coastal conditions during winter, when the model could be tracking above-freezing temperatures. Future work should investigate these links further to evaluate their meaning and implications for CNN-LSTM performance.**"

We agree that this is a prominent feature of the results and it is the subject of our future work.

A major point of the article is that the resulting CNN-LSTM neural network provides results that are physically explainable in the sense that perturbations to driving fields yield the streamflow responses one would expect on the basis of physical hydrologic knowledge. That's great, but it should be made very clear that this is not necessarily a unique attribute of deep learning – that has not been at all demonstrated here, and much the same might be expected from non-deep machine learning or even statistical models in the same application, provided they are built correctly.

We agree with the referee on this point and we do not claim that the physically explainable perturbation responses are necessarily unique to deep learning models. We add:

Line 152: "**We explore several ways that perturbations to the input temperature and precipitation fields result in streamflow responses that are expected on the basis of physical hydrologic knowledge. While this is not necessarily a unique property of DL and may be found when using non-deep machine learning or other empirical models applied to the same task, our findings are encouraging given the recent use of DL for streamflow prediction tasks.**"

Line 777: this is a grossly inadequate explanation

We have provided the following clarification:

Line 820: "**The CNN-LSTM is designed to receive an input structured as a weather video, while in comparison, ANNs are designed to receive an input structured as a single vector. The input neurons in the ANN correspond to each variable at each grid point and each day in a single weather video, meaning that there are 420,480 input neurons. For example, the input to predict flow on September 30, 2011 is daily maximum temperature, minimum temperature, and precipitation from September 30, 2010 through September 29, 2011, at each grid point in the study region. For the CNN-LSTM, these data are structured as a weather video with shape $365 \times 12 \times 32 \times 3$ (e.g. day $\times$ latitude $\times$ longitude $\times$ variable), but for the ANN, these data are structured as a vector with length 420,480.**"

Line 800: "seasonal-scale input time series" – really? Decades-long time series with a daily sampling interval were used in this study, were they not? So, what are the authors trying to express here?

We have clarified this section to reflect that a single year of temperature and precipitation is used to predict the next day of streamflow, from which there could be insufficient information to know the state of depressional storage (wetting or drying) and thus the correct streamflow response:

Line 845: "Storage in ponds can vary on both seasonal and decadal timescales (Hayashi et al., 2016; Shaw et al., 2012), **but only a single year of daily temperature and precipitation is used to predict the next day of streamflow.** It could be that the CNN-LSTM model **cannot accurately predict the streamflow response in eastern basins because one year of temperature and precipitation is insufficient information to know the state of depressional storage (e.g. seasonal and decadal fluctuations in wetting or drying).**"

Lines 813-816: this supposition is inconsistent with the basics of physical hydrology. Interactions between streamflow and geology (aquifers, soil moisture storage, etc) directly and nonlinearly affect the temporal dynamics of streamflow responses to forcing meteorology. The passage is also inconsistent with the work of Kratzert et al. (2019a), who demonstrated that including static catchment characteristics as predictors in a LSTM streamflow model substantially improves performance, and Kratzert et al. (2019b), who demonstrated that a new variant they developed of the LSTM can extract features corresponding to static basin characteristics that capture geological and other watershed properties.

This passage has been removed.

Line 880, "a single year of temperature and precipitation alone" – elsewhere the manuscript states that the data were over 1980-2015 (or 1979-2015, the paper is inconsistent on that point, e.g., between the abstract and conclusions). Even after subsetting the data into training and testing datasets, that still leaves several years, so where does the "single year" comment come from?

The "single year" referred to the input time series for the model (e.g. one year of daily temperature and precipitation is input to the model to predict one day of streamflow). In this case, we have removed "a single year" to comment only on processes which cannot be learned from temperature and precipitation:

Line 915: "The poor performance in the Prairie region may be due to the importance of processes which are underrepresented or not represented in the training data, such as processes occurring over longer than annual timescales, or at smaller spatial scales, or which are not able to be described from temperature and precipitation alone."

Temperature and precipitation data begin in 1979 while streamflow data begin in 1980 (since the prior year of temperature and precipitation is used to predict the next day of streamflow). We have ensured consistency on this point in the text (e.g. we refer to streamflow predictions between 1980 – 2015, and only refer to temperature and precipitation data in 1979). We also note specifically:

Line 383: "Since 365 days of previous temperature and precipitation are used to predict streamflow, and since the ERA5 data begin on December 1, 1979, the first day of streamflow predicted is January 1, 1980."

We have edited the conclusion so that it only refers to the date range for which streamflow predictions are made (1980 – 2015), so that it is consistent with the abstract:

Prior: "We focused on using a relatively simple deep learning model, with the input data represented by temperature and precipitation reanalysis for the period 1979 – 2015 given on relatively coarse spatial resolution (0.75° x 0.75°)."

Now, on line 904: "We focused on using a relatively simple deep learning model, with the input data represented by temperature and precipitation reanalysis given on relatively coarse spatial resolution (0.75° x 0.75°). **The deep learning model is used to predict daily streamflow between 1980 – 2015 at 226 stream gauge stations.**"

I believe the terms "validation" and "testing" are used inconsistently across the manuscript.

We have reviewed the use of "validation" and "testing" and have not found their use to be inconsistent. One potential source of confusion may be that the process-based models (e.g. in Table 2) report their performance on their "validation period", while we report our CNN-LSTM performance on our "testing period". We have further clarified this difference:

Line 797: "**We note a difference in terminology between the process-based model results and our CNN-LSTM results. Both evaluate models on 'unseen data' that were not used to determine the model parameters; however, the process-based models refer to this dataset as 'validation data' while we refer to this dataset as 'testing data'.**"

Moreover, for the benefit of readers less familiar with machine learning, the manuscript should clearly explain the difference between training, validation, and testing datasets in the context of the CNN-LSTM network used here.

We also clarify the difference between the training, validation, and testing datasets in the context of the CNN-LSTM model:

Line 378: "**We divide our data into three subsets referred to as training, validation, and testing datasets, as is common practice in DL model development (e.g. Goodfellow et al.,**

2016). The training data are used to iteratively update the model parameters such that the error between the model's predictions and known observations is reduced across the training set; the validation data are used to determine when to stop updating the model parameters to prevent the model from overfitting to the training data; and the testing data are used to evaluate the final model's performance."

References:

Ellenson A, Simmons JA, Wilson GW, Hesser TJ, Splinter KD. 2020. Beach state recognition using Argus imagery and convolutional neural networks. Remote Sensing, 12, 3953.

Fleming SW, Bourdin DR, Campbell D, Stull RB, Gardner T. 2015. Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. Journal of the American Water Resources Association, 51, 502-512.

Fleming SW, Garen DC, Goodbody AG, McCarthy CS, Landers LC. 2021b. Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: a challenging test of explainable, automated, ensemble artificial intelligence. Journal of Hydrology, 602, 126782.

Fleming SW, Vesselinov VV, Goodbody AG. 2021a. Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach. Journal of Hydrology, 597, 126327.

Hsieh WW, Yuval, Li J, Shabbar A, Smith S. 2003. Seasonal prediction with error estimation of Columbia River streamflow in British Columbia. Journal of Water Resource Planning and Management, 129, 146-149.

Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M. 2018. Rainfall-runoff modelling using long short-term memory (LSTM) networks. Hydrology and Earth System Sciences, 22, 6005-6022.

Kratzert F, Klotz D, Herrnegger M, Sampson AK, Hochreiter S, Nearing GS. 2019a. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resources Research, 55, 11344-11354.

Kratzert F, Klotz D, Shalev G, Klambauer G, Hochreiter S, Nearing G. 2019b. Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23, 5089-5110.

References:

Bergström, S.: Development and Application of a Conceptual Runoff Model for Scandinavian Catchments, 1976.

Braithwaite, R. J.: Positive degree-day factors for ablation on the Greenland ice sheet studied by energy-balance modelling, 41, 153–160, https://doi.org/10.1017/S0022143000017846, 1995.

Eaton, B. and Moore, R. D.: Regional Hydrology, in: Compendium of forest hydrology and geomorphology in British Columbia, edited by: Pike, R. G., Redding, T. E., Moore, R. D., Winkler, R. D., and Bladon, K. D., B.C. Ministry of Forests and Range, Victoria, British Columbia, 85–110, 2010.

Ellenson, A. N., Simmons, J. A., Wilson, G. W., Hesser, T. J., and Splinter, K. D.: Beach State Recognition Using Argus Imagery and Convolutional Neural Networks, 12, 3953, https://doi.org/10.3390/rs12233953, 2020.

Finsterwalder, S. and Schunk, H.: Der suldenferner, 18, 72–89, 1887.

Fleming, S. W., Bourdin, D. R., Campbell, D., Stull, R. B., and Gardner, T.: Development and Operational Testing of a Super-Ensemble Artificial Intelligence Flood-Forecast Model for a Pacific Northwest River, 51, 502–512, https://doi.org/https://doi.org/10.1111/jawr.12259, 2015.

Freeze, R. A. and Harlan, R. L.: Blueprint for a physically-based, digitally-simulated hydrologic response model, https://doi.org/10.1016/0022-1694(69)90020-1, 1969.

Gauch, M. and Lin, J.: A Data Scientist's Guide to Streamflow Prediction, 2020

Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, 2016.

Hayashi, M., van der Kamp, G., and Rosenberry, D. O.: Hydrology of Prairie Wetlands: Understanding the Integrated Surface-Water and Groundwater Processes, 36, 237–254, https://doi.org/10.1007/s13157-016-0797-9, 2016.

Hock, R.: Temperature index melt modelling in mountain areas, 282, 104–115, https://doi.org/https://doi.org/10.1016/S0022-1694(03)00257-9, 2003.

Hoinkesand, H. and Steinacker, R.: Hydrometeorological implications of the mass balance of Hintereisferner, 1952-53 to 1968-69, 1975.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019a.

Maier, H. R. and Dandy, G. C.: Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications, 15, 101–124, https://doi.org/https://doi.org/10.1016/S1364-8152(99)00007-9, 2000.

Maier, H. R., Jain, A., Dandy, G. C., and Sudheer, K. P.: Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions, 25, 891–909, https://doi.org/https://doi.org/10.1016/j.envsoft.2010.02.003, 2010.

Marsh, C. B., Pomeroy, J. W., and Wheater, H. S.: The Canadian Hydrological Model (CHM) v1.0: a multi-scale, multi-extent, variable-complexity hydrological model -- design and overview, 13, 225–247, https://doi.org/10.5194/gmd-13-225-2020, 2020.

Pomeroy, J. W., Gray, D. M., Brown, T., Hedstrom, N. R., Quinton, W. L., Granger, R. J., and Carey, S. K.: The cold regions hydrological model: A platform for basing process representation and model structure on physical evidence, in: Hydrological Processes, 2650–2667, https://doi.org/10.1002/hyp.6787, 2007.

Quick, M. C. and Pipes, A.: U.B.C. WATERSHED MODEL / Le modèle du bassin versant U.C.B, 22, 153–161, 1977.

Radic, V., Bliss, A., Beedlow, A. C., Hock, R., Miles, E., and Cogley, J. G.: Regional and global projections of twenty-first century glacier mass changes in response to climate scenarios from global climate models, 42, 37–58, https://doi.org/http://dx.doi.org/10.1007/s00382-013-1719-7, 2014.

Shaw, D. A., Vanderkamp, G., Conly, F. M., Pietroniro, A., and Martz, L.: The Fill–Spill Hydrology of Prairie Wetland Complexes during Drought and Deluge, 26, 3147–3156, https://doi.org/https://doi.org/10.1002/hyp.8390, 2012.

Response to Referee 2

I thank the authors for carefully considering my comments and providing detailed and appropriate responses. The revised manuscript is a major improvement in terms of presentation, and highlighting key findings and novelty. I have a few minor points that the authors need to clarify.

We are glad that the referee considers the revised manuscript to be a major improvement. We thank the referee for their points, which we address below (new text in **blue and bold**).

I am not convinced with the argument that ERA5 data has global coverage and appropriate for this study because the DL methods could be transferred to other regions. Just because ERA5 driven DL model performed well in this region does not mean that it is appropriate for other regions. The methods seem to be transferable irrespective of the input data, and 'best' available input data should be used in each study region. This needs to be clarified.

We have updated the following text:

Line 271: "ERA5 reanalysis is globally available from 1979 through the present (once complete it will be available from 1950 onwards) and has been shown to compare well against other reanalysis products (Hersbach et al., 2020). **ERA5 reanalysis was preceded by the ERA-Interim reanalysis, which has been evaluated for use across British Columbia. It was found that daily minimum and daily maximum temperatures are well represented across British Columbia (Odon et al., 2018). Additionally, daily precipitation was found to be well represented, with the caveat that extreme precipitation is less successfully represented (Odon et al., 2019). ERA5 reanalysis better represents precipitation as compared to ERA-Interim reanalysis at the global scale (Hersbach et al., 2020).** Importantly, precipitation from ERA5 has been found to typically outperform ERA-Interim reanalysis in the northern Great Plains region, which experiences a similar climate to the Prairie region in our study area (Xu et al., 2019). **For these reasons we consider the ERA5 reanalysis to be suitable for our study.**"

I am doubtful that the "day when the 30-day running mean of modelled streamflow rises to be halfway between the winter minimum flow (Qmin) and spring maximum flow (Qmax)" is a robust estimation of freshet timing. This is similar to "centre of volume" not being a robust measure of snowmelt timing (Whitfield 2013). Given that this is not a main focus of this paper, I suggested qualifying it as an 'indicator of freshet timing', with a caveat that this may not reflect actual freshet timing.

We qualify this as an "indicator of freshet timing" as suggested:

Line 552: "For each cluster and temperature perturbation, we define **an indicator of freshet timing** ($t_{freshet}$) as the day when the 30-day running mean of modelled streamflow rises to be halfway between the winter minimum flow (Q_{min}) and spring maximum flow (Q_{max})"

Throughout the text we now refer to $t_{freshet}$ as an “indicator of freshet timing” or “freshet timing indicator” rather than “freshet timing”.

Whitfield, P. H., 2013: Is ‘Centre of Volume’ a robust indicator of changes in snowmelt timing? Hydrol. Process., 27, 2691–2698, https://doi.org/10.1002/hyp.9817.