General comments:

This is an intriguing study that combines two distinct deep-learning technologies (the convolutional neural network, CNN, and long short-term memory neural network, LSTM) to create a new method for regional daily streamflow prediction that integrates complex spatiotemporal structures and dependencies. The method is applied to streamflow data from the southern portion of Canada's two westernmost provinces, which is a geophysically complex and interesting region. Some effort is also made to address physical interpretation and meaningfulness of the technique. It is a promising study with widely relevant results that has strong potential for publication in a top-tier hydrology journal like HESS.

We thank the referee for their comment and are glad they agree the results are widely relevant and the study is promising.

That said, the submission as it currently stands appears to have some substantial issues that need to be addressed before it can be considered for publication. The overall feel of how the manuscript is written is one of technical naivete and oversimplification, undermining the credibility of the study. For example, the text of the paper and possibly some of the analytical steps suggest a superficial understanding of the physical hydrology of western Canadian rivers and their associated datasets; and overall, the literature review around machine learning and its hydrologic applications is wholly inadequate and does not provide the reader with accurate and meaningful context to the study. Additionally, several basic elements one normally expects of a machine learning paper today seem to be missing, like clear descriptions of training vs. testing vs. validation data subsets, or the use of informative benchmark models to evaluate the new model against.

We have addressed these concerns in the revised manuscript. Our responses to each individual comment are given further below.

The study is also not reproducible based on the limited information provided in the paper.

We have now improved the information on the data used and added more details in the methods. We also added Table S1 in the Supplementary Information which contains station names, numbers, latitude, longitude, and RHBN status for all stations used in this study. For easier use, we also include these data as 'station_table.csv' on Github.

We note that the first submission included all code used on Github, including detailed steps how to download, access, and structure all data required. The file 'main_publish.ipynb' goes step by step to reproduce the figures used in the paper. In addition, we have now added a notebook 'mini.ipynb' which does not require readers to download any data themselves. Instead, we provide enough preprocessed climate/flow data to create 1 year of input/target and all trained bulk/fine models. From there, users can cluster the stream gauge stations, generate model predictions, evaluate model performance, make sensitivity heat maps, and perturb temperature

and measure the models' responses; essentially, all results from the paper, but with a single year of data instead of the >3 decades of data used in the full 'main_publish.ipynb'.

My recommendation is to accept the paper for publication in HESS pending major revisions. I hope the detailed comments provided below, as well as the references section that follows those detailed comments, will be helpful to the authors as they revise their manuscript.

We appreciate the detailed comments to help improve the manuscript. We have gone through and responded to the individual comments below.

Detailed comments:

*** Line 30: Should also cite Hsu et al. (1995) here, as to my knowledge it was the first peer-reviewed journal paper to present the use of machine learning for rainfall-runoff modeling. (Full literature citations are provided below.)**

We have now included this citation.

*** Lines 34-37: This feels like an overstatement/misstatement of both the limitations of conventional machine learning and the advantages of deep learning in a hydrologic prediction context. For one thing, a basic result in AI, dating back to the late 1980s or so, is that non-deep ANNs (in particular, multilayer perceptrons having a single hidden layer) are theoretically capable of learning any continuous relationship. Another issue: contrary to what is implied in the passage, non-deep ANNs are not the only kind of non-deep machine learning – there are several other major classes (random forests, support vector machines, and so forth). There also continues to be intense research in non-deep ML to create new kinds of AI, including news kinds of neural networks, having certain useful characteristics that have been successfully applied to river prediction; online sequential learning is an obvious example (e.g., Lima et al., 2015, 2016, 2017). Indeed, new kinds of non-deep machine learning algorithms are being developed specifically for hydrometeorological analysis and prediction tasks (e.g., Cannon, 2010, 2011, 2018; Fleming et al., 2015, 2019, 2021). On the other hand, deep learning applications in hydrology are currently in vogue and seem to be very promising in certain circumstances, but the body of work on the subject – particularly around streamflow prediction – remains exceedingly small, and the ultimate suitability of deep learning to this task, including capabilities and limitations, remains unclear at this point. A more mature way of looking at deep learning in hydrologic prediction is that work to date suggests it is a promising research direction that could potentially offer an alternative or complementary approach to non-deep machine learning for certain tasks.**

We have edited this section (and much of the introduction). New text is in **blue** if paragraphs have text from both the first submission and the edited manuscript, while new text is in black if the responses are entirely new text:

Line 38: "Early applications of machine learning in hydrology date back to the 1990s, with artificial neural network (ANN) models used for rainfall-runoff modelling (e.g. **Hsu et al., 1995; Maier and Dandy, 1996; Zealand et al., 1999) and a range of other hydrometeorological analysis such as flood forecasting (Fleming et al., 2015), improving gridded snow-water equivalent data products (Snauffer et al., 2018), and predicting total April-August streamflow (Hsieh et al., 2003).**"

Line 45: "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. For example: Bayesian neural networks, support vector regression, and Gaussian processes have been used for streamflow prediction at a single basin (Rasouli et al., 2012); quantile regression neural networks have been used for precipitation downscaling in British Columbia (Cannon, 2011) and estimation of rainfall intensity-duration-frequency curves across Canada (Cannon, 2018); online sequential extreme learning machines have been used for streamflow prediction in two basins (Lima et al., 2016, 2017); and random forest models have been used to identify temperature controls on maximum snow-water equivalence in Western North America (Shrestha et al., 2021). While ANNs and other non-deep machine learning architectures have a long history and continue to find useful applications in hydrology, DL has more recently become a promising area of investigation due to several key characteristics (Shen, 2018): DL models can automatically extract abstract features from large, raw datasets (Bengio et al., 2013), in contrast to labour-intensive manual feature extraction often required for non-deep models; and the existence of DL model architectures which are explicitly designed to learn complex spatial and/or temporal information, in particular convolutional neural networks (LeCun et al., 1990) and long short-term memory neural networks (Hochreiter and Schmidhuber, 1997)."

Line 156: "Deep learning in hydrology has shown promise for streamflow prediction tasks, but knowledge gaps exist surrounding the development of architectures which explicitly incorporate both space and time, the interpretation of model learning, and the limitations of such modelling approaches."

*** Lines 49-50: The use of point observations (of weather, presumably) does not necessarily imply that a model is spatially lumped. It is very common in process-based hydrologic modeling, including semi-distributed and fully distributed models, to spatially interpolate measurements from point data sources. In fact, some process-based models even integrate that spatial interpolation step into the software platform, along with adjustments for adiabatic lapse rates, etc., etc.**

The term "point observations" has been removed; while we meant that the LSTM approach had been used as a lumped hydrological model with point-observations as input, we agree with the referee that this was not necessarily clear.

*** Lines 67-70: Explainability is an issue for all machine learning models, not just deep learning models; it feels like this passage is conflating ML generally with DL specifically. For a recent example of a new non-deep ML technique specifically introduced to improve interpretability of a practical hydrologic prediction model, see Fleming et al. (2021), which also provides a much better explanation of exactly why geophysical explainability is a key requirement for practical applications of machine learning in hydrologic prediction.**

We add the following:

Line 106: "Fleming et al. (2021) discuss the importance of model interpretability in the context of operational hydrological forecasting where model predictions may be used for potentially high-stakes decision making. The end user may need to communicate why models make a certain prediction in order to answer clients' questions or to satisfy legal requirements. We may begin to build trust in a model's ability to forecast in the near-term by evaluating model performance on a testing dataset that is separate in time from the training and validation datasets. This approach, however, does not offer much insight into the physical relationships that the models are relying on for decision making. Additionally, without an understanding of what models have learned, it is challenging to trust a DL model for predictions in periods or places where observational datasets do not exist (e.g. for reconstructing missing historical streamflow, for predicting streamflow at ungauged basins, or for long-term forecasting of streamflow under climate change scenarios). By interpreting what a DL model has learned, we can better understand where and when a DL model can be trusted and the tasks for which it can be applied."

*** Lines 78-79: the authors are not using the terms white-box and (in particular) black-box in the way they are usually used. Most working in hydrology, in particular, would regard any physically explainable ML as being white-box in some sense. The term "black-box" is normally reserved for machine learning algorithms that do not offer any physical interpretability, which is to say, most of them.**

These lines have been rephrased:

Line 129: "In contrast to the above approaches which interpret the model through explicit use of the model parameters, alternative methods exist which do not use internal network states for interpretation."

*** Lines 85-86: would be useful to note the similarities and differences between recurrent and LSTM neural networks here for a general readership. The text seems to be haphazardly switching between the two, which are related but not identical; LSTM is essentially a specific and advanced form of recurrent ANN. This applies to the title of the paper too; why "recurrent" instead of "long short-term memory"?**

The original intent behind using "convolutional-recurrent" phrasing rather than "convolutional long short-term memory" phrasing was to keep the wording more succinct, which came at the cost of precision (since LSTM is a type of recurrent network). Upon reflection we have decided

to change this to "convolutional long short-term memory" to be more precise, both in the title and throughout the text. We also include:

Line 60: "Long short-term memory (LSTM) neural networks are designed to learn sequential relationships on a range of scales (Hochreiter and Schmidhuber, 1997). LSTMs are a type of recurrent neural network (RNN). Traditional RNNs include a feedback loop between the network output and input in order to learn temporal dependency within the data (Rumelhart et al., 1985); however, they struggle to learn long-term dependencies greater than around 10 time steps (Bengio et al., 1994). LSTMs overcome this limitation through the inclusion of an internal memory state or cell state which can store information, and learning is achieved by including internal gates through which information can flow and interact with the cell state."

*** Lines 104-105 are a bit off as well. There seems to be an implication here that more complex models are better models, and that in contrast this study is aiming for parsimonious models. That's an odd way of looking at the desirability of different modeling approaches and structures. Most modelers view a parsimonious model as being fundamentally better, holding all else equal, i.e., so-called Accom's razor.**

We did not mean to imply or convey that more complex models are necessarily better models; rather, we were considering that it may be possible to achieve better model performance in this instance through increased complexity (e.g. more layers / convolutional filters / LSTM units / etc). Here we simply meant that our goal is not to necessarily achieve better performance by optimizing all hyperparameters and architecture (e.g. we use our few-layer model which works well rather than aiming for a deeper model with more parameters which may work a bit better). This is now a point that we make later, and as such, we remove the sentence that was originally at lines 104-105. The point we now make later is:

Line 459: "It is possible that a better performing architecture or training scheme could be constructed by optimizing hyperparameters with an out-of-sample subset; however, we show our model setup and design is sufficient for achieving the goals of this study."

*** In addition to the various other papers referenced in this review that should be cited in the paper but were not, the authors may also wish to read and cite the review articles by Reichstein et al. (2019) and McGovern et al. (2019). Citing prior applications of machine learning to hydrologic and related modeling in the study area would also be appropriate. Some examples that come to mind include Rasouli et al. (2012), Lima et al. (2015, 2016, 2017), Snauffer et al. (2018), Fleming et al. (2015), Hsieh et al. (2003), and Shrestha et al (2021).**

We thank the referee for the suggested references. We now note (text in **blue** indicates new text in an old sentence; all other points are entirely new text):

Line 24: "The use of deep learning (DL) has gained traction in geophysical disciplines as an active field of exploration in efforts to maximize the use of growing in situ and remote sensing datasets (Bergen et al., 2019; **Reichstein et al., 2019**; Shen, 2018)."

Line 98: “Notably, the CNN-LSTM architecture has been identified as being an architecture of potential or emergent interest for geoscientific applications involving spatiotemporal phenomena (Reichstein et al., 2019).”

Line 89: “In the geosciences, CNNs have gained popularity more recently with applications including long-term El-Nino forecasting (Ham et al., 2019), precipitation downscaling (Vandal et al., 2017), **hail prediction (Gagne et al., 2019),** and urban water flow forecasting (Assem et al., 2017).”

Line 122: “Here we introduce select concepts and methods which can be used to interpret DL models; further details for machine learning and deep learning interpretation in a geoscientific context can be found in McGovern et al. (2019).”

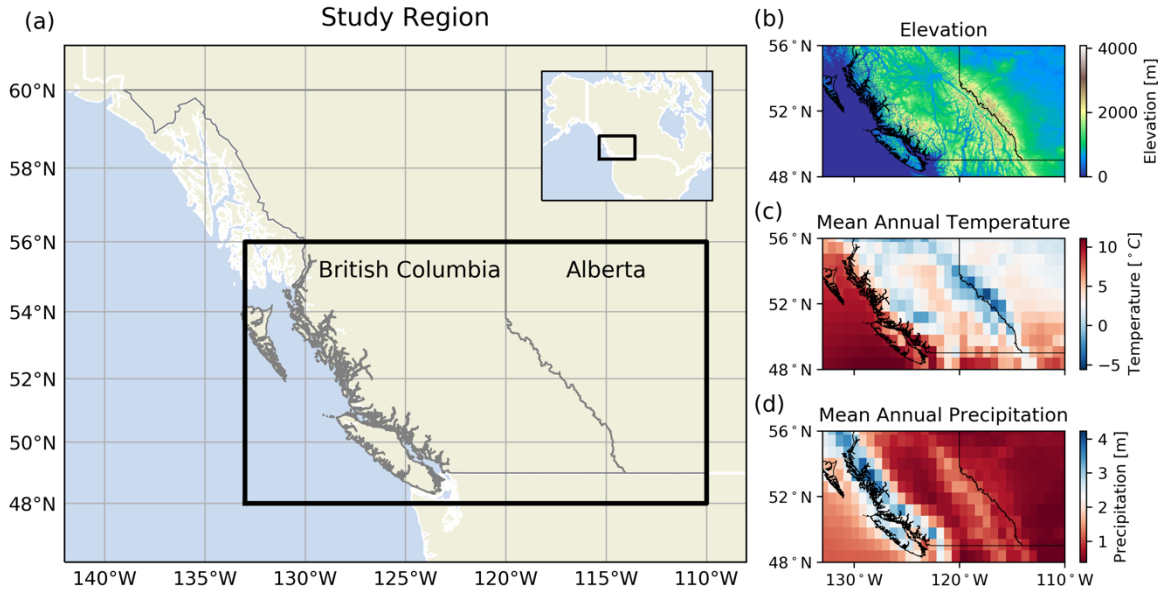
As previously noted in response to an earlier comment:

Line 38: “Early applications of machine learning in hydrology date back to the 1990s, with artificial neural network (ANN) models used for rainfall-runoff modelling (e.g. **Hsu et al., 1995; Maier and Dandy, 1996; Zealand et al., 1999) and a range of other hydrometeorological analysis such as flood forecasting (Fleming et al., 2015), improving gridded snow-water equivalent data products (Snauffer et al., 2018), and predicting total April-August streamflow (Hsieh et al., 2003).**”

Line 45: “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. For example: Bayesian neural networks, support vector regression, and Gaussian processes have been used for streamflow prediction at a single basin (Rasouli et al., 2012); quantile regression neural networks have been used for precipitation downscaling in British Columbia (Cannon, 2011) and estimation of rainfall intensity-duration-frequency curves across Canada (Cannon, 2018); online sequential extreme learning machines have been used for streamflow prediction in two basins (Lima et al., 2016, 2017); and random forest models have been used to identify temperature controls on maximum snow-water equivalence in Western North America (Shrestha et al., 2021). While ANNs and other non-deep machine learning architectures have a long history and continue to find useful applications in hydrology, DL has more recently become a promising area of investigation due to several key characteristics (Shen, 2018): DL models can automatically extract abstract features from large, raw datasets (Bengio et al., 2013), in contrast to labour-intensive manual feature extraction often required for non-deep models; and the existence of DL model architectures which are explicitly designed to learn complex spatial and/or temporal information, in particular convolutional neural networks (LeCun et al., 1990) and long short-term memory neural networks (Hochreiter and Schmidhuber, 1997).”

*** Figure 1 would be much better, especially for an international readership that is unlikely to be strongly familiar with the study area, if it was a multi-panel figure that additionally illustrated topography, mean annual temperature, mean annual precipitation, and perhaps mean April 1 snow water equivalent.**

We have updated this figure and its caption to include panels of elevation, mean annual temperature, and mean annual precipitation:



*** Lines 137-139: perhaps this passage merely is poorly written, but as it stands, the text implies a disturbing lack of understanding of the streamflow data being modeled. Naturalized flow data are flow data that have been adjusted for upstream water management activities – diversions, withdrawals, reservoir operations, etc. Data for stations upstream of dams are not necessarily naturalized, contrary to what is implied in this passage of the paper, and certainly in datasets like the HYDAT database used here, that step has not been undertaken and in many cases is unnecessary. Similarly, dams are not the only disturbance that result in non-natural streamflow data that would in principle require naturalization prior to use in a hydrologic modeling study of the sort done here; another obvious example is land use change.**

The word "naturalized" was an unfortunate typo, and it should have been "natural" flow ("natural" in the sense that the HYDAT system classifies stream gauges as either "natural" or "regulated"). We clarify this in the text:

Line 206: "HYDAT classifies stream gauge stations as either "regulated" (downstream of regulating structures such as a dam) or "natural" (upstream of regulating features). We use stations which are classified as natural and which are currently active."

Why not use the Reference Hydrometric Basin Network (RHBN) stations or something similar? There is no mention here at all of the RHBN station network, which has been very widely used for decades for hydrological analysis and modeling studies in Canada.

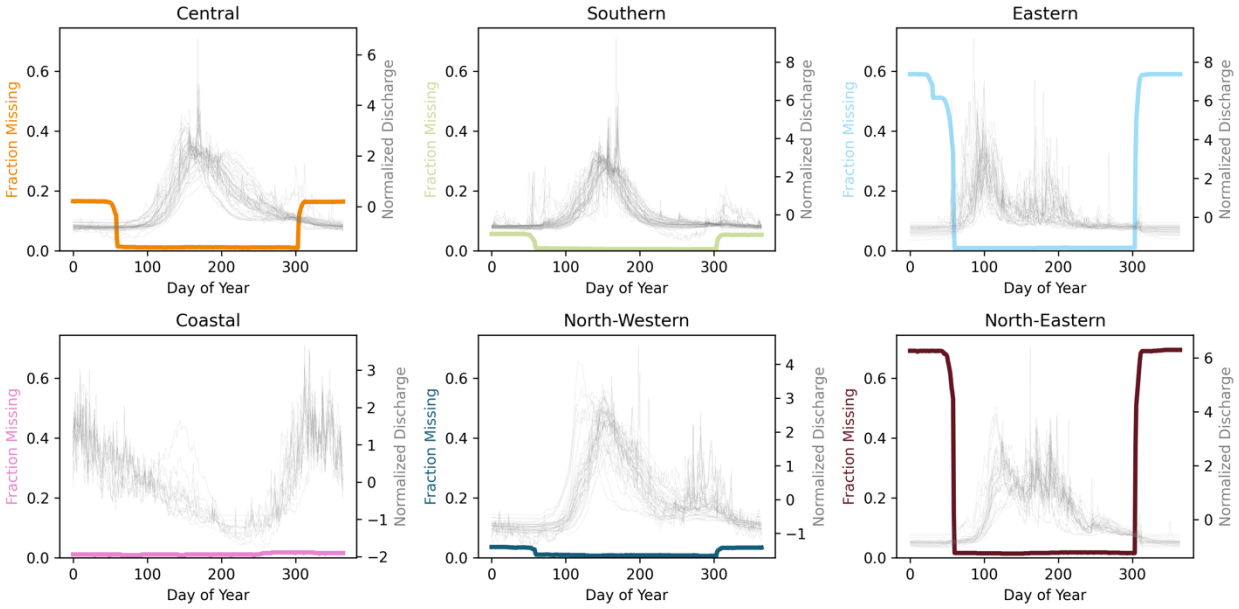
We now include the following on the RHBN:

Line 224: "The Reference Hydrometric Basin Network (RHBN) is a subset of the national stream gauge network which have long records and minimal human impacts that have been identified for use in climate change studies. Of the 226 stations used in our study, 213 are within the RHBN. The remaining 13 stations have long observational records and are not modified by regulating structures but may have more than minimal human impacts through other disturbances to the natural system such as land use changes. We provide station names, station numbers, and if they are a part of the RHBN network (Table S1)."

Also, I think quite a few hydrologists would raise their eyebrows at the specific data selection and processing procedures described in the first paragraph of section 3.1.

Unfortunately, this comment does not provide us with any information as to how we could improve our data selection and processing procedures. We will here comment on the steps we took.

One challenge is that we need temporally complete datasets, as the number of output neurons is constant through training. We recognize that the threshold of 40% missing data may lead to challenges, if the data is missing during periods of dynamic streamflow (in other words, the model would be missing how to learn when streamflow should substantially change due to meteorological forcing); however, this is not the case. A vast majority of data is missing between November and February, when temperatures are coldest and streamflow is more inhibited as compared to spring. These data are typically missing at stations which record data seasonally, rather than continuously, and the 40% threshold allows us to "forgive" seasonal stations which do not record in winter months. The following figure demonstrates how the most missing data occurs in the low-flow period.



*** The second paragraph of section 3.1 is also muddled. All that's needed here is a concise statement that hydrometric network density is much higher in southern than northern Canada, and so, for the purposes of this study, the authors focused on the former.**

We edit this point:

Line 220: "We further restrict the study region to stations south of 56° N because stream gauge density is greater below this latitude."

*** While the approach described on lines 159-170 is interesting and perhaps sufficient for the purposes of this study, overall it appears to be a naïve representation of spatiotemporal pattern formation in streamflow regimes in this study area. At an absolute minimum, some acknowledgement of prior work, and some caveats about the simple method and assumptions used here for regime classification, are needed. See in particular Halverson and Fleming (2015) and references cited therein. A particularly notable omission is that glacier-fed rivers are not identified as a distinct regime, whereas glacial cover is well-known to be a major control of streamflow dynamics in several areas within this region; see Moore et al. (2009), Fleming et al. (2016), Jost et al. (2012), and Bidlack et al. (2021).**

We recognize that we take a relatively simple approach in clustering stations into subdomains based only on seasonal hydrograph, latitude, and longitude. However, we use this clustering step not with the sole goal of finding stations which have the most similar physical and hydrological conditions (e.g. glacier cover, aspect, land use); rather, a key product of clustering is to find subsets of stations for which the model's learning can be more easily interpreted. It is desirable to identify clusters which are in large part determined by geographic location because one goal is to visualize where in space the model is learning to focus when predicting streamflow for each

cluster. When stations are nearby each other in space and the model is most sensitive in that small region, then we can better understand that the model is looking in the right place. When stations are spread over a larger area and clusters overlap more in space (e.g. if the importance of latitude and longitude are "watered down" by using other predictors in the clustering algorithm), the model may be sensitive over a large overlapping area for multiple clusters, and it becomes harder to interpret. Is the model focused on the watershed regions? Or is it just using the entire domain?

As noted to reviewer 1 who also had a similar comment on clustering: Consider two stations which are nearby one another, but have different characteristics such as drainage area, elevation, slope, aspect, and glaciation. In order to predict streamflow at each station, it should still be most important that the model focuses on areas near and within the two watersheds, respectively. For each station, the mapping through to streamflow from this 'most relevant information', then, may be different, but the sensitive areas should be similar. So, while clustering in the space of hydrologic variables other than geographic location may lead to small improvements in performance as measured by NSE by allowing the fine-tuned model to 'focus in' on more common details, it may make it more difficult to understand what the model is learning to do.

While we choose to not change our clustering method, we have added more context about the region's hydrology:

Line 230: "Streamflow throughout the study region varies strongly in space and time and reflects the varied topographic and climatic conditions in British Columbia and Alberta. Here we provide a brief, high-level overview of streamflow characteristics, and while it is not a complete summary of the full range of hydrologic conditions throughout the study region, we aim to highlight that streamflow through the region is heterogeneous in space and time. Streamflow at low-elevation coastal stations is primarily driven by rainfall, with monthly discharge maximized in November or December. In contrast, streamflow at stations that are at higher elevation, further north, or further inland transition to a snowmelt-dominated regime, with monthly discharge maximised in spring or early summer. Numerous glaciers exist in high elevation alpine areas throughout both the Coast Mountains along the west coast of British Columbia and the Rocky Mountains along the border between British Columbia and Alberta, and glacier runoff contributes to streamflow through late summer once the seasonal snowpack has melted (Eaton and Moore, 2010). East of the Rocky Mountains, the Prairie region in eastern Alberta is uniquely characterized by relatively flat topography with small surface depressions (LaBaugh et al., 1998). Water can pond and be stored in these depressions, leading to intermittent connectivity throughout many basins and drainage areas which may vary in time (e.g. Shook and Pomeroy, 2011)."

To comment on prior work that used clustering in the region:

Line 251: "Previous studies have used a range of techniques to cluster or summarize the diversity of spatiotemporal streamflow characteristics in the study region (e.g. Halverson and Fleming (2015) use complex networks to represent similarity between streamflow timeseries in the

Coastal Mountains, while Anderson and Radić (2020) use principal component analysis and Self-Organizing Maps to characterize summer streamflow through Alberta). In this study we use a relatively simple clustering approach, only considering seasonal streamflow, station latitude, and station longitude."

To comment on why we use this simpler clustering approach:

Line 274: "Our clustering approach does not explicitly consider input features such as land use, glacier coverage, drainage area, or elevation, but rather implicitly considers the expressions of these features in the seasonal hydrograph. The goal of this type of clustering is to define subsets of stream gauge stations that are nearby in space and share similar hydrographs. We prioritize proximity in space over an explicit representation of other important features (e.g. drainage area, elevation, glacier coverage) because a key goal of the study is to interpret where in space the DL models have learned to focus when predicting streamflow. As discussed in Sect. 4.3.1 and Sect. 4.5.1, having clusters of stream gauge stations which are nearby in space allows us to visualize if the trained models are learning to focus on the subregion of the input domain which overlaps with the watersheds where streamflow is being predicted."

*** Section 3.1: I think reproducibility requires that the hydrometric station list used here be shown to readers. A table in an appendix or supplementary materials would be fine.**

A table of station names, numbers, latitude, longitude, and if they are part of the RHBN network has been included in supplementary information (Table S1) and as 'station_table.csv' on Github.

*** Section 3.5: provide information about the latency of the ERA5 reanalysis product – is it available in near-real time? Some reanalysis products are, and some aren't. It's a crucial question if one were interested in operationalizing a hydrologic prediction system like this for actual use in flood forecasting or another similar practical hydrologic prediction application. If ERA5 products are not available in near-real time, then briefly but clearly state that limitation and its implications for wider use of the modeling framework introduced here.**

We include the following:

Line 317: "ERA5 data are available as a preliminary product 5 days behind real time, and as a final product 2 – 3 months behind real time (Hersbach et al., 2020). This latency has implications for model applications, as it may not be possible to use ERA5 data for real-time forecasting with the model in this study."

*** "data" = plural**

Having double checked all uses of "data", we found two which were incorrectly singular and these have been corrected.

*** Somewhere in Section 3 or 4 there needs to be an explicit and clear description of what the training vs. testing vs. validation datasets are. There is a very brief mention of training vs validation but it is inadequate. The reader is not provided with information about how the training vs validation split is made, nor whether another subset is reserved for out-of-sample hyperparameter selection. These are standard practices in machine learning, and information about them is needed for transparency, reproducibility, and credibility of the study.**

We edit and include the following text:

In original manuscript: 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. For all models, we use 1980 – 2000 for training, 2001 – 2010 for validation, and 2011 – 2015 for testing.

Added in updated manuscript, Line 416: "In other words, the training period is defined by daily streamflow from January 1, 1980 to December 31, 2000, with forcing data ranging from January 1, 1979 to December 30, 2000. The validation period uses streamflow data from January 1, 2001 to December 31, 2010, with forcing data ranging from January 1, 2000 to December 30, 2010. The testing period uses streamflow data from January 1, 2010 to December 31, 2015, with forcing data ranging from January 1, 2009 to December 30, 2015. We choose to separate the training/validation/testing datasets into non-overlapping time periods of streamflow so that model performance can be evaluated on out-of-sample streamflow examples. We choose to use a full decade for validation because we want to encourage the model to perform well across a range of conditions and not for one particular year or climate state, since oscillations in the climate system such as the El-Nino Southern Oscillation, the Pacific Decadal Oscillation, and the Pacific-North American atmospheric teleconnection influence streamflow through modifications to temperature, precipitation, and snow accumulation through the study region (e.g. Fleming and Whitfield, 2010; Hsieh et al., 2003; Hsieh and Tang, 2001; Whitfield et al., 2010). We also choose to use multiple years for testing so as to not bias our conclusions towards the conditions of a single year. Furthermore, we partition the training, validation, and testing data by year rather than by percentage of observations (i.e. the testing subset is chosen as 5 years, not 10% of observations) so that we do not bias our results by including or excluding parts of the year when the model performs better or worse than average. Overall, the training-validation-testing data split is approximately 59% - 27% - 14% of the total streamflow dataset. The input data are normalized so that each variable (maximum temperature, minimum temperature, precipitation) has a mean of zero and unity variance over the training period. The target data from each of the 226 stations are normalized so that each station's streamflow has a mean of zero and unity variance over the training period."

Line 459: "It is possible that a better performing architecture or training scheme could be constructed by optimizing hyperparameters with an out-of-sample subset; however, we show our model setup and design is sufficient for achieving the goals of this study."

*** A modern paper on machine learning applications to hydrologic prediction requires, in general, a performance comparison against some relevant benchmark model. Linear regression using precisely the same input dataset as the deep learning method introduced here is an obvious starting point and can provide a meaningful assessment of how much nonlinearity, interactions, etc contribute to the (presumably better) performance of the new technique. A conventional ANN and an LSTM would also be useful, if more ambitious, points of comparison.**

We agree with the reviewer and have now included a comparison of our CNN-LSTM model with an ensemble of linear models. The revisions read as following:

Line 862: "We compare our fine-tuned CNN-LSTM models against linear models to evaluate the extent to which the nonlinearities introduced by the CNN-LSTM approach improve streamflow predictions. We create an ensemble of 10 linear models for each cluster of stream gauge stations. Each linear model is a fully-connected ANN with an input layer, an output layer, and linear activation functions. We use the same training, validation, and testing data as in the CNN-LSTM approach. However, instead of structuring the input data as a video, each input observation is flattened and all values are input into the ANN. The target output is the next day of streamflow at all stations in the cluster. Therefore, for each model for cluster i, there are 420,480 input neurons (since each original observation is structured as a $365 \times 12 \times 32 \times 3$ video) and N output neurons (where N is the number of stations in cluster i). This approach was chosen in order to keep as much similarity as possible between the CNN-LSTM and linear model setup. The two approaches use the same input data, the same target data, and the same number of ensemble members, while the key difference is the nonlinearity and architecture of the CNN-LSTM model. We find that the CNN-LSTM model outperforms this simple linear benchmark, achieving a greater NSE at 222 out of 226 stations. The linear model has a minimum NSE of -13.33, a median NSE of 0.35, and a maximum NSE of 0.76, while the CNN-LSTM model has a minimum NSE of -0.7, a median NSE of 0.68, and a maximum NSE of 0.96."

The only significant attempt the paper makes at this is Table 2, which scours the peer-reviewed journal literature for examples of hydrologic models that have been developed previously for a few of the locations considered in this study. That comparison is interesting and probably worth including in the paper, but it also has limited meaningfulness as different date ranges etc were used in the studies. Moreover, Table 2 relies on a small handful of academic studies and misses a lot of existing models within the study area operated by pragmatic water-management organizations like a large government-owned hydroelectric utility (BC Hydro), a provincial ministry (BC River Forecast Center), regional water management authorities (e.g., the MIKE-SHE model operated in the Okanagan Basin), and so forth. Moreover, given that even the simplest machine learning architecture outperforms process-based models in most cases, the somewhat mixed results in Table 2 are a little surprising.

We recognize that this comparison may have limited meaningfulness as different date ranges were used in the study, which is why we emphasized in the text that it is not a direct comparison.

However, it is still valuable to at the very least comment on the performance of existing models in the peer-reviewed literature.

In Section 5 there is also a very brief verbal comparison against the LSTM-based work of Kratzert et al. (2018) but that study used a completely different set of basins and data, so again, the comparison is extremely approximate.

Yes, we are aware that Kratzert et al. (2018) use a different set of basins and data, which we note in the text. This does not mean that there is nothing to learn from prior regional-DL models. The purpose of this discussion in Section 5 is to say that the LSTM approach was improved by including temporally static catchment characteristics in the input, and noting that there would be ways to extend the CNN-LSTM approach to do this as well, outlining a potential avenue for future work.

I get that the purpose of this study is more around demonstrating a new technology, and perhaps delving a little into the question of explainability, but I suspect most readers would like to see more meaningful inter-model performance comparisons here.

We agree that inter-model comparison is important here, and we thank the referee for the idea to include a comparison to the linear benchmark (outlined above).

*** Estimating predictive uncertainty is a key element of a hydrologic prediction system. Figure 6 and its caption suggests that predictive uncertainty is quantitatively estimated here but is vague about the method. It appears that an ensemble of 10 different models is formed, and twice the standard deviation of the predictions from those 10 models on a given day is used as the de facto prediction bound for that day. This is a reasonable first-cut approach, I think. However, the method needs to be described in the methods section, and some capabilities and limitations need to be mentioned; I suspect that because weather uncertainty is not factored in (as far as I can tell from the manuscript as submitted) the ensemble spread will be substantially under-dispersive.**

These uncertainty bounds reflect the range of streamflow predictions due to the randomness in the initialization of weights of the network and through training, and do not reflect uncertainty in meteorological drivers. This point has been made clearer by including the following:

Line 475: "We compute NSE using the mean predictions across the ensemble members, and we quantify an uncertainty in the streamflow prediction as being twice the standard deviation across ensemble members. This uncertainty is due to randomness from the initialized parameters and through training. It is a measure of how different streamflow predictions may be even when using the same architecture and data, and it is not a measure of uncertainty in meteorological forcing. When and where this uncertainty is small (large) indicates that the models in the ensemble predict similar (different) streamflow values for that day. We evaluate performance from an ensemble mean rather than a single model's prediction, and so this uncertainty gives an indication of the magnitude of scatter around the ensemble mean."

*** The bar for explainability does not seem to be set very high here. The sensitivity analyses included in the paper are very useful, but they really amount to more of a plausibility test than an interpretability test. In particular, the paper demonstrates, though observing the CNN-LSTM responses to perturbations in the meteorological driving data, that its streamflow predictions (a) are most sensitive to weather in and near the basin as opposed to further away, and (b) are sensitive to temperature regimes, in particular, demonstrate hydrograph timing shifts corresponding to changes in snow accumulation and melt driven by temperature perturbations.**

Yes, the paper demonstrates the points (a) and (b), but it also goes further than that. We also demonstrate that the process of fine-tuning strongly influences the model's decision making by allowing it to (c) focus on smaller areas of the input (smaller A through fine-tuning), and (d) become more sensitive to perturbation near/within the watersheds being predicted and less so to areas further away (larger D through fine-tuning). "Interpretability" is more than revealing physical explanations of the input-output relationships, but is also building an understanding of the role of training steps. One question a person could ask themselves when training an ML model is: "When am I finished training?". One might compare NSE between a bulk and fine-tuned model and find them to be very close (as they are in this study, and others e.g. Kratzert et al. 2018). Only through (c) and (d) might we interpret how and if fine-tuning is improving model performance – perhaps not by making streamflow predictions which are more similar to observed values, but by better focusing in on the watershed areas.

Those results suggest the CNN-LSTM model is capturing key geophysical processes more-or-less correctly, but it does not clearly reveal physical explanations of the input-output relationships – only that the behaviors are consistent with some basic physical expectations. I think the paper is publishable without diving further into explainability, but the authors ought to phrase their outcomes a little more precisely around the question of interpretability and may wish to consider some additional sleuthing to demonstrate that the CNN-LSTM reveals physical processes. There is some precedent for this in machine learing-based streamflow modeling, and looking closely at those precedents may be useful to the authors; examples include Fleming (2007), Kratzert et al. (2018), and Fleming et al. (2021). Looking even more broadly across the literature than this would likely lead to even more suggestions of how to examine the geophysical relationships the model is capturing.

While it is not uncommon to center "interpretability" around the question of "which pixels are most important / relevant / sensitive for the model's decision making?" in the geophysical deep learning literature (e.g. Toms et al. (2020)), we have now added an additional analysis. We demonstrate that in glacier-fed rivers, August temperature perturbations are positively related to August mean streamflow (e.g. hotter temperatures lead to more flow), while this is not the case for non-glacier-fed rivers. Additionally, the strength of this relationship is positively (and non-linearly) related to the watershed glacier cover (greater percentage glaciation leads to flow being more sensitive to August temperature perturbations). This evidence supports the hypothesis that the model is learning physical processes (e.g. glacier-runoff contributions to streamflow, where melt is positively related to temperature) and is elaborated in the text:

Line 18: "We also demonstrate that modelled August streamflow in partially glacierized basins is sensitive to perturbations in August temperature, and that this sensitivity increases with glacier cover."

In Methods section:

Line 589: "Glacier runoff is a key contributor to streamflow in many watersheds in the study region, and compared to non-glacier-fed rivers, glacier-fed rivers have enhanced streamflow in late summer due to glacier runoff contributions after much of the seasonal snowpack has melted (e.g. Comeau et al., 2009; Jost et al., 2012; Moore et al., 2009; Naz et al., 2014). Additionally, glacier runoff counteracts variability in precipitation as enhanced (suppressed) glacier melt compensates for less (more) precipitation during hot and dry (cold and wet) years, leading to reduced interannual variability of total summer streamflow (Fountain and Tangborn, 1985; Meier and Tangborn, 1961). These effects lead to spatiotemporal patterns of summer streamflow in glacier-fed rivers which are markedly different than those in non-glacier-fed rivers (e.g. Anderson and Radić, 2020). Therefore, the model should learn a unique mapping of late summer climatic drivers to streamflow for glacier-fed rivers as compared to non-glacier-fed rivers, and the difference in these mappings can be exploited to interpret model learning. In particular, since temperature is a strong control of melt, we assume that mean August streamflow (Q_{Aug}) is positively related to mean August temperature (T_{Aug}) in basins with partial glacier coverage. Again, while this is a simplification of the actual glacier melt processes, it is a key assumption in widely used temperature index melt models and is supported by empirical evidence in the study region (Moore et al., 2009; Stahl and Moore, 2006). We introduce the following hypothesis: if the model is learning to represent physical processes which drive streamflow in August, then modelled Q_{Aug} in glacier-fed rivers should increase with increasing T_{Aug}, while modelled Q_{Aug} in non-glacier-fed rivers should not increase with increasing T_{Aug}. To test this hypothesis, we introduce a spatially uniform temperature perturbation to only days in August, ΔT_{Aug}, and add it to the maximum and minimum temperature channels. We then compute Q_{Aug} for each station. We perturb August temperatures from $-5^{\circ}\mathrm{C} \leq \Delta T_{Aug} \leq 5^{\circ}C$ with an increment of $1^{\circ}\mathrm{C}$ and use linear regression to estimate the sensitivity $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ for each station as:

$$Q_{Aug} = \frac{\partial Q_{Aug}}{\partial T_{Aug}} T_{Aug} + c \tag{17}$$

where $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ is calculated as the slope of the linear regression and c is a constant coefficient (intercept). We compute basin glacier cover, G, for each stream gauge station as:

$$G = \frac{A_{glaciers}}{A_{basin}} \tag{18}$$

where $A_{glaciers}$ is the total area of glaciers within the watershed boundaries and A_{basin} is the basin drainage area as reported in HYDAT (Environment and Climate Change Canada, 2018). To calculate $A_{glaciers}$, we determine which glacier outlines fall within the watershed boundaries and then sum their areas, where glacier locations and areas are taken from the Randolph Glacier Inventory Version 6 (RGI Consortium, 2017)."

In Results section:

Line 806: "When August temperatures are perturbed with $\Delta T_{Aug} > 0$, modelled mean August streamflow in partially glacierized watersheds increases, while when August temperatures are perturbed with $\Delta T_{Aug} < 0$, modelled mean August streamflow in partially glacierized watersheds decreases. This is indicated by $\frac{\partial Q_{Aug}}{\partial T_{Aug}} > 0$ for stations where watershed glacier cover is non-zero (Figure 11). In contrast, perturbations of mean August temperature (positive or negative) do not (or negligibly) influence modelled Q_{Aug} for stations where watersheds have no glacier coverage, which is indicated by $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ being narrowly distributed around zero for these stations (Figure 11). Additionally, we investigate how $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ varies for three ranges of watershed glacier cover, G, here defined as light glacier cover ($0\% < G \leq 1\%$), moderate glacier cover ($1\% < G \leq 10\%$), and substantial glacier cover ($10\% < G \leq 100\%$). We find that the median $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ increases as G increases from light, to moderate, to substantial glacier cover (Figure 11b), indicating that mean August streamflow is more sensitive to August temperature perturbations at higher glacier coverage.
"

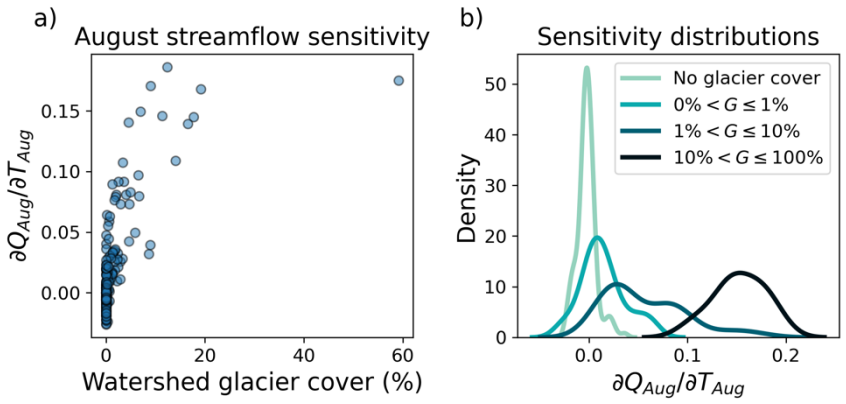


Figure 11: Modelled sensitivity of mean August streamflow to mean August temperature. a) $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ increases non-linearly with watershed glacier cover, G, indicating that greater watershed glacier coverage is related to more positive $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$. b) Probability distributions of $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ for different ranges of watershed glacier coverage, indicating that $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ for glacier-fed rivers is both greater than non-glacier-fed rivers, and greater at increasing glacier coverage. All probability distributions are normalized to have unity area."

In Discussion section:

Line 919: "When August temperatures are made warmer (cooler), modelled streamflow in partially glacierized watersheds increases (decreases) and the sensitivity of modelled August streamflow to these temperature perturbations is greater in more glacierized watersheds as compared to less glacierized watersheds (Figure 11). The positive relationship between Q_{Aug} and T_{Aug} in glacierized watersheds indicates that the model has learned that glacierized watersheds have an input to streamflow which is positively related to temperature in August, while non-glacierized watersheds do not. We interpret this result as the model learning to represent glacier runoff as a temperature-dependent source. Interestingly, the relationship between $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ and watershed glacier cover as derived from the sensitivity test of the CNN-LSTM model (Figure 11a), is similar in form to an empirically derived relationship between $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ and watershed glacier cover in British Columbia (Figure 5 in Moore et al. (2009), from analysis in Stahl and Moore (2006)). Both analyses identify a positive non-linear relationship between $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ and G when $G > 0$, while $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ is distributed around zero when $G = 0$. Note however that the raw values of $\frac{\partial Q_{Aug}}{\partial T_{Aug}}$ differ between our approach and that in Figure 5 of Moore et al. (2009) due to differing normalization schemes. While it is interesting that the model has learned the unique characteristics of temperature-driven August flow of glacierized watersheds, it also highlights a challenge when applying the CNN-LSTM model in its current realization for applications such as long-term forecasting under climate change. Under warmer future climate forcing, the model would associate higher temperatures with greater flow. However, projections of future glacier volume indicate that 70-90% of glacier ice volume will be lost by 2100 in Western Canada (Clarke et al., 2015; Marshall et al., 2011), and so it is expected that the learned temperature-flow relationship from the past will no longer hold under such conditions."

Line 994: "To investigate the learning of unique processes in partially glacierized basins, we focused on the sensitivity of August flow to August temperature. By increasing August temperature input to drive the model, the model responded by increasing August flow in partially glacierized basins while not increasing August flow in non-glacierized basins. The sensitivity of flow to temperature was found to be greater in more glacierized basins as compared to less glacierized basins."

*** Lines 629, "it is notable that the CNN-LSTM model achieves good streamflow simulation with only temperature and precipitation forcing data" – well, in practice the most widely applied hydrologic models tend to use only these two types of forcing because that's all that is usually available, so I guess this point might be worth mentioning here but it's not particularly "notable" to most streamflow modelers.**

We rephrase this line:

Line 938: "It is notable that the CNN-LSTM model achieves good streamflow simulation with only coarse resolution climate forcing data and localized streamflow data, with no knowledge of features such as basin characteristics, topography, or land use, and no explicit climate downscaling steps."

*** Lines 635-638: is it possible that, through its empirical and complex meteorological input-hydrologic output mappings – effectively, a transfer function linking the meteorological data to the point streamflow observations – the CNN-LSTM effectively downscaled the reanalysis data, at least to some degree? May be worth talking about here.**

Yes, this is possible, especially considering that CNNs have been used to map coarse resolution climate data to fine resolution climate data, indicating that sufficient information of high-resolution climate data is present within coarse resolution climate data (Vandal et al. (2017)). We now make the following point in the text:

Line 940: "Our model uses forcing data at relatively coarse spatial resolution (0.75° x 0.75°, or ~75 km resolution) as compared to studies identified in Table 2 (e.g. 0.0625° x 0.0.0625° in Shrestha et al. (2012); 10 km resolution in Eum et al. (2017)). Studies that employ a climate downscaling step first map coarse resolution climate data to fine resolution climate data, and then map the downscaled fine resolution climate data to streamflow. Here, the CNN-LSTM is effectively representing a single transfer function that maps coarse resolution climate data directly to streamflow, and it is possible that an effective downscaling of climate data is learned by the model. This indirect downscaling is plausible since statistical methods are often used for climate downscaling, including CNNs (Vandal et al., 2017)."

*** Lines 646-653: are the authors sure their method requires less data than an LSTM, as claimed here? Doesn't the CNN-LSTM still ultimately need data for all N basins? This passage needs further explanation/clarification.**

It is not that our method requires less data, but that our method leads to having fewer observations for training. Consider a scenario where an LSTM is being used to predict streamflow at 10 stations individually (1 station-day of streamflow per observation), compared with a CNN-LSTM which is being used to predict streamflow at all 10 stations simultaneously (10 station-days of streamflow per observation). Suppose that each station has 20 years of observations for training, meaning that there are (365 days / year) * (20 years) * (10 stations) station-days of streamflow in this dataset. The LSTM approach converts 1 station-day to 1 observation for training. The CNN-LSTM approach converts 10 station-days to 1 observation for training. This reduces the number of observations for training by the CNN-LSTM approach by a factor of 10 (i.e. reducing the number of observations for training, but not reducing the total data requirements). We do not wish to frame this as a "good" or "bad" thing, but rather it is something to consider when designing a model and how to train it.

We rephrase and include the following to improve clarity on this point:

Line 947: "In order for the model to learn the mapping between the meteorological forcing and streamflow, a sufficiently long data record is necessary for training. The CNN-LSTM architecture presented here predicts streamflow at multiple stations simultaneously. For a model which predicts at N stations simultaneously, one target observation is N station-days of streamflow. For a model which predicts at a single station (e.g. an LSTM with a single output neuron), one target observation is a single station-day of streamflow. For a given training dataset with M station-days of streamflow observations, the CNN-LSTM with N output neurons would have M/N observations for training, while the model with a single output neuron would have M observations for training. That the number of observations for training has been reduced is potentially detrimental to the model's performance. A potential solution to this problem could be to use transfer learning with a CNN-LSTM model pre-trained in a region with a sufficiently long streamflow record and then transferred to the new region of interest."

References:

Bidlack AL, Bisbing SM, Buma BJ, Diefenderfer HL, Fellman JB, Floyd WC, Giesbrecht I, Lally A, Lertzman KP, Perakis SS, Butman DE, D'Amore DV, Fleming SW, Hood EW, Hunt BPV, Kiffney PM, McNicol G, Menounos B, Tank SE. 2021. Climate-mediated changes to linked terrestrial and marine ecosystems across the Northeast Pacific Coastal Temperature Rainforest margin. Bioscience, doi.org/10.1093/biosci/biaa171.

Cannon AJ. 2010. A flexible nonlinear modelling framework for nonstationary generalized extreme value analysis in hydroclimatology. Hydrological Processes, 24, 673-685.

Cannon AJ. 2011. Quantile regression neural networks: implementation in R and application to precipitation downscaling. Computers and Geosciences, 37, 1277-1274.

Cannon AJ. 2018. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. Stochastic Environmental Research and Risk Assessment, 32, 3207-3225

Fleming SW. 2007. Artificial neural network forecasting of nonlinear Markov processes. Canadian Journal of Physics, 85, 279-294.

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, Goodbody AG. 2019. A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West. IEEE Access, 7, 119943-119964.

Fleming SW, Hood E, Dahlke HE, O'Neel S. 2016. Seasonal flows of international British Columbia-Alaska rivers: the nonlinear influence of ocean-atmosphere circulation patterns. Advances in Water Resources, 87, 42-55.

Fleming SW, Vesselinov VV, Goodbody AG. 2021. 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.

Halverson MJ, Fleming SW. 2015. Complex network theory, streamflow, and hydrometric monitoring system design. Hydrology and Earth System Sciences, 19, 3301-3318.

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.

Hsu K, Gupta HV, Sorooshian S. 1995. Artificial neural network modeling of the rainfall-runoff process. Water Resources Research, 31, 2517-2530.

Jost G, Moore RD, Menounos B, Wheate R. 2012. Quantifying the contribution of glacier runoff to streamflow in the upper Columbia River Basin, Canada. Hydrology and Earth System Sciences, 16, 849-860.

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.

Lima AR, Cannon AJ, Hsieh WW. 2015. Nonlinear regression in environmental sciences using extreme learning machines: a comparative evaluation. Environmental Modelling and Software, 73, 175-188.

Lima AR, Cannon AJ, Hsieh WW. 2016. Forecasting daily streamflow using online sequential extreme learning machines. Journal of Hydrology, 537, 431-443.

Lima AR, Hsieh WW, Cannon AJ. 2017. Variable complexity online sequential extreme learning machine, with applications to streamflow prediction. Journal of Hydrology, 555, 983-994.

McGovern A, Lagerquist R, Gagne DJ II, Jergensen GE, Elmore KL, Homeyer CF, Smith T. 2019. Making the black box more transparent: understanding the physical implications of machine learning. Bulletin of the American Meteorological Society, November, 2175-2199.

Moore RD, Fleming SW, Menounos B, Wheate R, Fountain A, Stahl K, Holm K, Jakob M. 2009. Glacier change in western North America: influences on hydrology, geomorphic hazards, and water quality. Hydrological Processes, 23, 42-61.

Rasouli K, Hsieh WW, Cannon AJ. 2012. Daily streamflow forecasting by machine learning methods with weather and climate Inputs. Journal of Hydrology, 414/415, 284-293.

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N, Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. Nature, 566, 195-204.

Shrestha RR, Bonsal BR, Bonnyman JM, Cannon AJ, Najafi MR. 2021. Heterogeneous snowpack response and snow drought occurrence across river basins of northwestern North America under 1.0*C to 4.0*C global warming. Climatic Change, 164, 40.

Snauffer AM, Hsieh WW, Cannon AJ, Schnorbus MA. 2018. Improving gridded snow water equivalent in British Columbia, Canada: multi-source data fusion by neural network methods. The Cryosphere, 12, 891-905.

References:

Anderson, S. and Radić, V.: Identification of local water resource vulnerability to rapid deglaciation in Alberta, Nat. Clim. Chang., 10(10), 933–938, doi:10.1038/s41558-020-0863-4, 2020.

Assem, H., Ghariba, S., Makrai, G., Johnston, P. and Pilla, F.: Urban Water Flow and Water Level Prediction based on Deep Learning, ECML PKDD 2017 Mach. Learn. Knowl. Discov. databases, 317–329, 2017.

Bengio, Y., Simard, P. and Frasconi, P.: Learning Long-term Dependencies with Gradient Descent is Difficult, Trans. Neur. Netw., 5(2), 157–166, doi:10.1109/72.279181, 1994.

Bengio, Y., Courville, A. and Vincent, P.: Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell., 35(8), 1798–1828, doi:10.1109/TPAMI.2013.50, 2013.

Bergen, K. J., Johnson, P. A., de Hoop, M. V and Beroza, G. C.: Machine learning for data-driven discovery in solid Earth geoscience, Science (80-.)., 363(6433), doi:10.1126/science.aau0323, 2019.

Cannon, A. J.: Quantile regression neural networks: Implementation in R and application to precipitation downscaling, Comput. Geosci., 37(9), 1277–1284, doi:https://doi.org/10.1016/j.cageo.2010.07.005, 2011.

Cannon, A. J.: Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes, Stoch. Environ. Res. Risk Assess., 32(11), 3207–3225, doi:http://dx.doi.org/10.1007/s00477-018-1573-6, 2018.

Clarke, G. K. C., Jarosch, A. H., Anslow, F. S., Radić, V. and Menounos, B.: Projected deglaciation of western Canada in the twenty-first century, Nat. Geosci., 8, 372 [online] Available from: https://doi.org/10.1038/ngeo2407, 2015.

Comeau, L. E. L., Pietroniro, A. and Demuth, M. N.: Glacier contribution to the North and South Saskatchewan Rivers, Hydrol. Process., 23(18), 2640–2653, doi:10.1002/hyp.7409, 2009.

Eaton, B. and Moore, R. D.: Regional Hydrology, in Compendium of forest hydrology and geomorphology in British Columbia, edited by R. G. Pike, T. E. Redding, R. D. Moore, R. D. Winkler, and K. D. Bladon, pp. 85–110, B.C. Ministry of Forests and Range, Victoria, British Columbia. [online] Available from: https://www.for.gov.bc.ca/hfd/pubs/Docs/Lmh/Lmh66.htm, 2010.

Environment and Climate Change Canada: Water Survey of Canada HYDAT data, [online] Available from: https://wateroffice.ec.gc.ca/mainmenu/historical_data_index_e.html, 2018.

Eum, H.-I., Dibike, Y. and Prowse, T.: Climate-induced alteration of hydrologic indicators in the Athabasca River Basin, Alberta, Canada, J. Hydrol., 544, 327–342, doi:https://doi.org/10.1016/j.jhydrol.2016.11.034, 2017.

Fleming, S. W. and Whitfield, P. H.: Spatiotemporal mapping of ENSO and PDO surface meteorological signals in British Columbia, Yukon, and southeast Alaska, Atmosphere-Ocean, 48(2), 122–131, doi:10.3137/AO1107.2010, 2010.

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, JAWRA J. Am. Water Resour. Assoc., 51(2), 502–512, doi:https://doi.org/10.1111/jawr.12259, 2015.

Fleming, S. W., Vesselinov, V. V and Goodbody, A. G.: Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach, J. Hydrol., 597, 126327, doi:https://doi.org/10.1016/j.jhydrol.2021.126327, 2021.

Fountain, A. G. and Tangborn, W. V: The Effect of Glaciers on Streamflow Variations, Water Resour. Res., 21(4), 579–586, doi:10.1029/WR021i004p00579, 1985.

Gagne II, D. J., Haupt, S. E., Nychka, D. W. and Thompson, G.: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms, Mon. Weather Rev., 147(8), 2827–2845, doi:10.1175/MWR-D-18-0316.1, 2019.

Halverson, M. J. and Fleming, S. W.: Complex network theory, streamflow, and hydrometric monitoring system design, Hydrol. Earth Syst. Sci., 19(7), 3301–3318, doi:http://dx.doi.org/10.5194/hess-19-3301-2015, 2015.

Ham, Y.-G., Kim, J.-H. and Luo, J.-J.: Deep learning for multi-year ENSO forecasts, Nature, doi:10.1038/s41586-019-1559-7, 2019.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. and Thépaut, J.-N.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146(730), 1999–2049, doi:10.1002/qj.3803, 2020.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Comput., 9(8), 1735–1780, doi:10.1162/neco.1997.9.8.1735, 1997.

Hsieh, W. W. and Tang, B.: Interannual variability of accumulated snow in the Columbia Basin, British Columbia, Water Resour. Res., 37(6), 1753–1759, doi:https://doi.org/10.1029/2000WR900410, 2001.

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.

Hsu, K., Gupta, H. V. and Sorooshian, S.: Artificial Neural Network Modeling of the Rainfall-Runoff Process, Water Resour. Res., 31(10), 2517–2530, doi:https://doi.org/10.1029/95WR01955, 1995.

Jost, G., Moore, R. D., Menounos, B. and Wheate, R.: Quantifying the contribution of glacier runoff to streamflow in the upper Columbia River Basin, Canada, Hydrol. Earth Syst. Sci., 16(3), 849–860, doi:10.5194/hess-16-849-2012, 2012.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22(11), 6005–6022, doi:10.5194/hess-22-6005-2018, 2018.

LaBaugh, J. W., Winter, T. C. and Rosenberry, D. O.: Hydrologic functions of prairie wetlands, Gt. Plains Res., 8(1), 17–37 [online] Available from: http://www.jstor.org/stable/24156332, 1998.

LeCun, Y., Boser, B., Denker, J. S., Howard, R. E., Habbard, W., Jackel, L. D. and Henderson, D.: Handwritten Digit Recognition with a Back-Propagation Network, in Advances in Neural Information Processing Systems, pp. 396–404, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA., 1990.

Lima, A. R., Cannon, A. J. and Hsieh, W. W.: Forecasting daily streamflow using online sequential extreme learning machines, J. Hydrol., 537, 431–443, doi:https://doi.org/10.1016/j.jhydrol.2016.03.017, 2016.

Lima, A. R., Hsieh, W. W. and Cannon, A. J.: Variable complexity online sequential extreme learning machine, with applications to streamflow prediction, J. Hydrol., 555, 983–994, doi:https://doi.org/10.1016/j.jhydrol.2017.10.037, 2017.

Maier, H. R. and Dandy, G. C.: The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters, Water Resour. Res., 32(4), 1013–1022, doi:10.1029/96WR03529, 1996.

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, Environ. Model. Softw., 15(1), 101–124, doi: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, Environ. Model. Softw., 25(8), 891–909, doi:https://doi.org/10.1016/j.envsoft.2010.02.003, 2010.

Marshall, S. J., White, E. C., Demuth, M. N., Bolch, T., Wheate, R., Menounos, B., Beedle, M. J. and Shea, J. M.: Glacier Water Resources on the Eastern Slopes of the Canadian Rocky Mountains, Can. Water Resour. J., 36(2), 109–134, doi:10.4296/cwrj3602823, 2011.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R. and Smith, T.: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning, Bull. Am. Meteorol. Soc., 100(11), 2175–2199, doi:10.1175/BAMS-D-18-0195.1, 2019.

Meier, M. F. and Tangborn, W. V: Distinctive characteristics of glacier runoff, US Geol. Surv. Prof. Pap, 424, B14–B16, 1961.

Moore, R. D., Fleming, S. W., Menounos, B., Wheate, R., Fountain, A., Stahl, K., Holm, K. and Jakob, M.: Glacier change in western North America: influences on hydrology, geomorphic hazards and water quality, Hydrol. Process., 23(1), 42–61, doi:10.1002/hyp.7162, 2009.

Naz, B. S., Frans, C. D., Clarke, G. K. C., Burns, P. and Lettenmaier, D. P.: Modeling the effect of glacier recession on streamflow response using a coupled glacio-hydrological model, Hydrol. Earth Syst. Sci., 18(2), 787–802, doi:10.5194/hess-18-787-2014, 2014.

Rasouli K, Hsieh WW, Cannon AJ. 2012. Daily streamflow forecasting by machine learning methods with weather and climate Inputs. Journal of Hydrology, 414/415, 284-293.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N. and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566(7743), 195–204, doi:http://dx.doi.org/10.1038/s41586-019-0912-1, 2019.

RGI Consortium: Randolph Glacier Inventory (RGI) - A Dataset of Global Glacier Outlines, Glob. L. Ice Meas. from Space, Boulder, Color. USA, Digit. Media, doi:https://doi.org/10.7265/N5-RGI-60, 2017.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning Internal Representations by Error Propagation, Institute for Cognitive Science, University of California, San Diego., 1985.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resour. Res., 54(11), 8558–8593, doi:10.1029/2018WR022643, 2018.

Shook, K. R. and Pomeroy, J. W.: Memory effects of depressional storage in Northern Prairie hydrology, Hydrol. Process., 25(25), 3890–3898, doi:https://doi.org/10.1002/hyp.8381, 2011.

Shrestha, R. R., Schnorbus, M. A., Werner, A. T. and Berland, A. J.: Modelling spatial and temporal variability of hydrologic impacts of climate change in the Fraser River basin, British Columbia, Canada, Hydrol. Process., 26(12), 1840–1860, doi:https://doi.org/10.1002/hyp.9283, 2012.

Shrestha RR, Bonsal BR, Bonnyman JM, Cannon AJ, Najafi MR. 2021. Heterogeneous snowpack response and snow drought occurrence across river basins of northwestern North America under 1.0*C to 4.0*C global warming. Climatic Change, 164, 40.

Snauffer AM, Hsieh WW, Cannon AJ, Schnorbus MA. 2018. Improving gridded snow water equivalent in British Columbia, Canada: multi-source data fusion by neural network methods. The Cryosphere, 12, 891-905.

Stahl, K. and Moore, R. D.: Influence of watershed glacier coverage on summer streamflow in British Columbia, Canada, Water Resour. Res., 42(6), doi:10.1029/2006WR005022, 2006.

Toms, B. A., Barnes, E. A. and Ebert-Uphoff, I.: Physically Interpretable Neural Networks for the Geosciences: Applications to Earth System Variability, J. Adv. Model. Earth Syst., 12(9), e2019MS002002, doi:https://doi.org/10.1029/2019MS002002, 2020.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R. and Ganguly, A. R.: DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution, eprint arXiv:1703.03126, arXiv:1703.03126 [online] Available from: https://ui.adsabs.harvard.edu/abs/2017arXiv170303126V, 2017.

Whitfield, P. H., Moore, R. D. (Dan), Fleming, S. W. and Zawadzki, A.: Pacific Decadal Oscillation and the Hydroclimatology of Western Canada—Review and Prospects, Can. Water Resour. J. / Rev. Can. des ressources hydriques, 35(1), 1–28, doi:10.4296/cwrj3501001, 2010.

Zealand, C. M., Burn, D. H. and Simonovic, S. P.: Short term streamflow forecasting using artificial neural networks, J. Hydrol., 214(1), 32–48, doi:https://doi.org/10.1016/S0022-1694(98)00242-X, 1999.