Referee Comment #1:

This manuscript presents an interesting application of deep learning approach for modelling streamflow responses across 226 streamflow gauges in southwestern Canada. This paper is well written and mostly easy to follow. I find the application of DL approach for streamflow simulation to be quite innovative and worthy addition to the growing body of literature in this field. The paper will be of widespread interest to the community.

We thank the referee for their comments and are glad they find our study to be innovative and interesting.

Overall, the paper offers plenty of interesting work, however, some effort is needed to communicate the results more effectively and highlighting key findings and novelty in view of the journal audience, e.g., better describe temperature and spatial perturbations results by linking with previous studies. Additional discussion is also needed on what the DL method brings to the table in comparison to the traditional process based models. Furthermore, while the application of DL methods for streamflow simulation is interesting, it is not entirely clear how this approach could be used for real world applications. There are also questionable choices on some of the methods and data used.

We thank the referee for the constructive criticism and have now addressed these key comments and revised the manuscript accordingly. The specific points are responded to where they are elaborated on in "Major Comments".

I also find the most figure captions lacking in details and could be expanded to provide more details. This will avoid readers having to scroll up and down the paper to understand the details in figures.

We now add more detail and expand most Figure captions. In particular, the revised captions read as following:

Line 640:

Figure 4:
"NSE values are greatest (indicating the best model performance) throughout mainland British Columbia, and are smallest (indicating the worst model performance) in south-eastern Alberta. A is smallest (indicating small sensitive areas) in the south-west and north-west coastal regions in British Columbia, and is largest (indicating large sensitive areas) throughout the rest of British Columbia and near the Alberta border."

Line 670:

Figure 5:

"The central and southern clusters show the least amount of change between the bulk and fine-tuned models, while all other clusters increase NSE through fine-tuning (indicating improved model performance)."

"In the central cluster, the variance of D across model runs decreases through fine-tuning, indicating improved consistency between the fine-tuned central models. In all other clusters, D increases, indicating improved separation between information which is near/within basins as compared to information which is further away."

"In all clusters, A decreases through fine-tuning, indicating that fine-tuned models are sensitive to smaller areas of the input as compared to the bulk models."

Line 706:

Figure 7:
"The (a) central, (b) southern, (d) coastal, and (e) north-western clusters are generally most sensitive in the areas nearest the basins where streamflow is being predicted. The (c) eastern and (f) north-eastern clusters are most sensitive to perturbations both near the stations being predicted and further away along the west coast."

Line 726:

Figure 8:
"… and the cluster watershed regions are shown in Fig. A7. While all clusters are more sensitive to perturbations within/near their watershed regions, the coastal cluster demonstrates the greatest difference in sensitivity between within/near the watersheds and the rest of the domain."

Line 794:

Figure 9:
"In the (a) central, (b) southern, (c) eastern, (e) north-western, and (f) north-eastern clusters, $\Delta T > 0^{\circ}C$ leads to an earlier freshet with a smaller peak flow, while $\Delta T < 0^{\circ}C$ leads to a later freshet with a larger peak flow. In the (d) coastal cluster, $\Delta T > 0^{\circ}C$ leads to enhanced streamflow in winter and fall and suppressed streamflow in summer, while $\Delta T < 0^{\circ}C$ leads to suppressed streamflow in winter and fall and enhanced streamflow in summer."

Line 1451:

Figure A5:
"ΔNSE is most positive (indicating the greatest improvement through fine-tuning) along the west coast and northern regions of both British Columbia and Alberta. ΔA is most negative (indicating that fine-tuning reduces the size of the sensitive areas) along the west coast, in northern British Columbia, and throughout Alberta."

Line 1523:

Figure A7:
"Watershed boundaries are shown in white and are accessed through the Water Survey of Canada (Environment and Climate Change Canada, 2016)."

I hope these comments are helpful and I look forward to reading the revised manuscript. My detailed comments are given below.

We thank the referee for the numerous constructive comments which have helped to improve this study.

Major comments

It is not clear what the application of DL method bring to the table in comparison to the traditional process based hydrologic models.

We recognize that we could be clearer on this point in Sect. 1. As such, we have revised the first paragraph as follows (italicized are new, rather than rewritten/clarified points from the initial submission):

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). In hydrology, DL can provide alternative or complementary approaches to supplement traditional process-based modelling (Hussain et al., 2020; Marçais and de Dreuzy, 2017; Shen, 2018; Shen et al., 2018; Van et al., 2020). *Particularly notable are DL models which have been found to outperform traditional hydrological models applied at regional scale, including those for streamflow prediction at daily temporal scale (Kratzert et al., 2018, 2019b), at hourly temporal scale (Gauch et al., 2021), and at ungauged basins (Kratzert et al., 2019a). These recent DL-based studies have emphasized the development of lumped hydrological models; however, progress has not yet been made toward distributed DL hydrological models (Gauch and Lin, 2020).* In contrast, traditional process-based approaches have made substantial towards distributed hydrological models (Freeze and Harlan, 1969; Marsh et al., 2020; Pomeroy et al., 2007). Nevertheless, as input and target data are becoming available at increasingly finer spatiotemporal resolution, process-based modellers are having to address the rising computational requirements and human labour required to represent the relevant hydrological processes across larger spatial scales (Marsh et al., 2020). *A key opportunity exists, then, to develop a DL hydrological model which can utilize spatially discretized forcing data at regional scale.*"

We also note key benefits of DL as opposed to non-deep machine learning:

Line 53: "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)."

The use of machine learning in hydrology, and the advent of deep learning in hydrology in the last several years, has gained substantial research interest with numerous works exploring how and why deep learning offers new and exciting ideas both in addition to and in complement to traditional hydrological models (in particular, see Shen 2018 and Shen et al. 2018). As we refer to numerous studies that:

- use deep learning which outperform regional-scale traditional hydrological models (e.g. Kratzert et al. 2018, Kratzert et al. 2019a, and in the revised submission, Gauch et al. 2021);
- use deep learning in the geosciences more broadly (e.g. Vandal et al. 2017, Ham et al. 2019, Gange et al. 2019, McGovern et al. 2019); and
- use non-deep learning in Canadian hydrometeorology (e.g. Cannon 2011, Cannon 2018, Lima et al. 2016 and 2017, Shrestha et al. 2021, Snauffer 2018),

we do not think it is necessary to further explore or elaborate in more detail why DL is an attractive complement to traditional modelling approaches as this content has been addressed in the cited literature.

We hope the referee agrees we have made it clearer in the revised submission that DL models have outperformed numerous process-based models, motivating their use and further investigation, but at the same time the most successful DL models follow a lumped hydrological modelling approach. Therefore there is an opportunity to explore DL models which can take advantage of spatially discretized forcing datasets.

This is an important question as the application of DL methods may be limited to predicting within the range of training datasets. Additionally, while the authors outlined the development and evaluation of DL method for streamflow simulation as their objectives, it is not stated how the DL method could be used for real world applications beyond the proof of concept type approach presented in this paper.

We now more clearly state the advantages of DL for real world applications. One of the key aspects in real world applications is whether one can trust the model predictions. The application of DL models is limited to making predictions during periods when we can trust the model's predictions. This "period of trust" is the training period only if we cannot trust the model beyond when it was trained. We can build trust in model predictions outside of this period in different ways:

Trusting models for near-term forecasting: In our study, we evaluate the model performance on the testing set (2011 – 2015), which follows the training/validation period (up to 2010). This provides evidence that the models have learned enough from the training period to successfully extrapolate to the near-term.

Trusting models for long-term forecasting: Traditional modelling approaches are used for long-term forecasting (e.g. climate change projections) under the assumption that the representation of the underlying physics will not change between the model training/validation period and the future. In other words, we may trust traditional models to work in the future (or at the very least, know the limits of our trust in these models in the future) because we understand why they work. To trust DL models for future projections, we need to understand what they are learning. Our study makes progress on this front, and in the revised submission, we emphasize more why this is important and what sorts of applications become available when we better trust and understand DL models in hydrology.

We address these points, and potential future applications and areas of study arising from the success of the CNN-LSTM approach and the use of globally available climate reanalysis data, in the following points added to the updated submission:

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."

Line 1004: "Considering that ERA5 climate reanalysis has global spatial coverage and is temporally complete to 1979, there are many opportunities to investigate the transferability of this approach to different regions, the use of different predictor variables, and the use of different spatial and temporal resolution of both input and target data."

Beyond the applications in the above text (e.g. reconstructing missing historical streamflow, investigating transferability of this approach to additional regions/scales), we noted in the original manuscript (and clarified in the resubmission) that the modelling setup may be conducive for transfer learning, which is another avenue of application which could help overcome challenges arising from data limitations:

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."

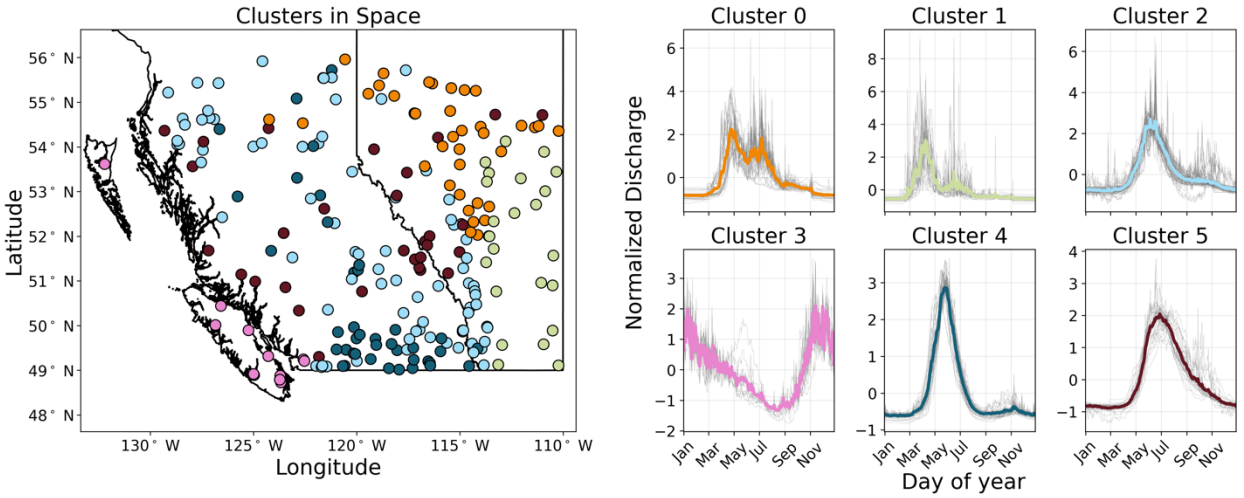
The clustering method divided the study region into six clusters based on seasonal streamflow, latitude and longitude variables in order to fine-tune the model training. However, there are a number of studies in the region which describe the spatial heterogeneity of the region. For instance, streamflow responses in the lee- and windward side of coast and rocky mountains, as well as mountainous and interior plains are know to be quite different (e.g., Moore 1991; Shrestha et al. 2012). Therefore, I would think including variables like slope and aspect will be able to better characterize the spatial heterogeneity and provide clusters that better capture the variability in the streamflow response. Better clustering can potentially improve the model fine-tuning and model performance in several regions, especially in the Eastern slopes of the Rocky Mountains where the model performed relatively poorly.

This is a valid point raised by the referee and is one we have considered during the study design.

The clustering is not used to just find stations which are most hydrologically similar; if this were the sole goal of clustering, then we agree with the referee and should use a greater number of relevant variables such as slope and aspect (and drainage area, elevation, glacier coverage, etc.) which better capture the heterogeneity of the region. A key product of clustering is to find subsets of stations which can be predicted by an *interpretable* model. In our case, it is desirable to identify clusters which are in large part determined by geographic location because one of our main goals is to determine where in space the model is learning to focus when predicting streamflow at each cluster.

If multiple clusters overlap in space, it is less easily detectable if the model is learning to focus on different physically relevant areas, or if it is learning to map similarly large areas of the input through to the output. To illustrate this point, we here cluster only the seasonal streamflow (i.e. without the geographical information). In this case, clusters 2, 4, and 5 span similar large regions of the input space, and also overlap considerably with cluster 0. For clusters 2, 4, and 5 in particular, we would expect the model to be sensitive to perturbation throughout most of British Columbia, and so the sensitivity heat maps would be very similar. We would then not be able to tell if the model is really learning different things for clusters 2, 4, and 5. By emphasizing geographic location in clustering, as we do in the paper, we can set ourselves up to train models

that can be better interpreted because of their sensitivity to different geographic areas in the domain.



Another example to justify our point: consider two stations that are nearby one another, but have different characteristics such as slope, aspect, and percent glaciation in the watershed. 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, it makes it more difficult to understand what the model is learning and why it is making decisions.

To add additional context about the region's hydrology, we revised the following text:

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)."

Revisions that discuss prior work:

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."

It is surprising to see that the study used 0.75° x 0.75° resolution ERA5 reanalysis data, especially given that the authors stated finer resolution climate data may improve model performance (L637). I wonder why the authors did not used to the finer resolution data readily available for the region (e.g., Werner et al. 2019)?

We recognize that there are multiple high quality datasets available for our study region. One key reason to use ERA5 data specifically, and not other regional datasets which are available for this region, is because of ERA5's global spatial coverage. By training on ERA5 data, models can more easily be adapted for applications such as hindcasting or for transfer learning across regions. We did not pursue these applications here, but we comment on their potential for future work.

There are also practical computational reasons to use coarser resolution climate data. There is a balance between "how much information there is to learn from" (related to the size/dimension of individual observations for training, i.e. the amount of information in a single weather video) and "how much learning can be done during training" (related to the number of observations available for training); generally speaking, the more information there is to learn from, the more learning that needs to be done. When predicting streamflow at multiple stations simultaneously,

the number of observations for training is reduced, and so it is beneficial to also reduce the size/dimension of individual training samples. The simplest way to do this is by using coarser resolution climate data, which we find to be sufficient for meeting our goals.

It is common in hydrological modelling to first perform downscaling of the input data to the model resolution. In other words, there is a mapping from coarser resolution climate data, to finer resolution climate data, to streamflow (e.g. there is a transfer function from coarse resolution climate data to fine resolution climate data, and then another transfer function from fine resolution data to streamflow). Here, the intention is that the effect of downscaling is learned in the mapping from coarse resolution climate data to streamflow directly (e.g. a single transfer function from coarse resolution climate data to streamflow). That CNNs have been used to map coarse-resolution climate data to high-resolution climate data suggests that key information of high-resolution climate data is present within coarse resolution data (e.g. Vandal et al., 2017), and so it could be possible that an 'implicit downscaling' is learned during the mapping from coarse climate data to streamflow. We now include this 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)."

The authors described the DL methods as if the study is on image/video processing. While the methods may be same as image/video processing, there is a need to rephrase section 4 in terms of hydro-climatic modelling.

The benefit of using terminology from video/image processing (e.g. 'video', 'frame', 'channel', 'pixel') is that it offers a succinct way to describe the input data and its structure while maintaining accuracy (e.g. it is simpler to refer to a 'frame' rather than 'one day of three weather predictors'). However, we agree that the connection between the image processing and hydro-climatic terminology should be improved. To improve clarity, we add the following description in the first paragraph of Sect. 4.3:

Line 381: "To ensure consistency between terminology in both image processing and hydro-climatic modelling, a 'weather video' refers to 365 days of the three weather predictors, a 'frame' or 'image' in a weather video refers to one day of the three weather predictors, a 'channel' in a 'frame' or 'image' refers to one day of one weather predictor, and a 'pixel' refers to one grid cell."

Specific comments

L95-110: The objective and novelty need to be revised by clearly describing what the DL method bring to the table compared to the process-based models, and how the DL method could be used for real-world application.

See response to the first two major comments.

L135: What are the range of basin areas for the selected stations?

The basin drainage areas span approximately 5 orders of magnitude (minimum area: ~6 km^2, maximum area: 133,000 km^2). Stream gauge elevation and drainage area are now visualized in Figure A3 and referred to in Sect. 3.2 (Streamflow clusters):

Line 265: "The elevation and drainage area of stations for each cluster is shown in Fig. A3."

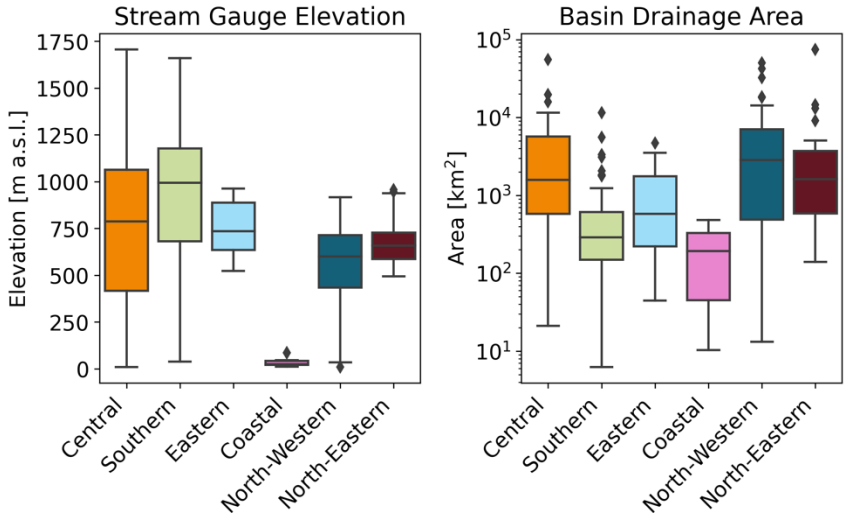


Figure A3: Elevation and drainage area of stations within each of the identified clusters. Station elevation is calculated from a digital elevation model from the Shuttle Radar Topography Mission (SRTM) at 90 m resolution (Farr et al., 2007). Drainage area is taken from the Environment Canada HYDAT database (Environment and Climate Change Canada, 2018). The coastal cluster is at the lowest elevation and with the smallest drainage areas. Clusters in mainland British Columbia (central, southern, and north-western) span wide ranges of elevation and drainage area, while clusters in Alberta (eastern and north-eastern) span narrower ranges of elevation and drainage area.

L138: Naturalized flow generally means regulated flow adjusted with regulation/abstraction removed. Correct term is natural flow.

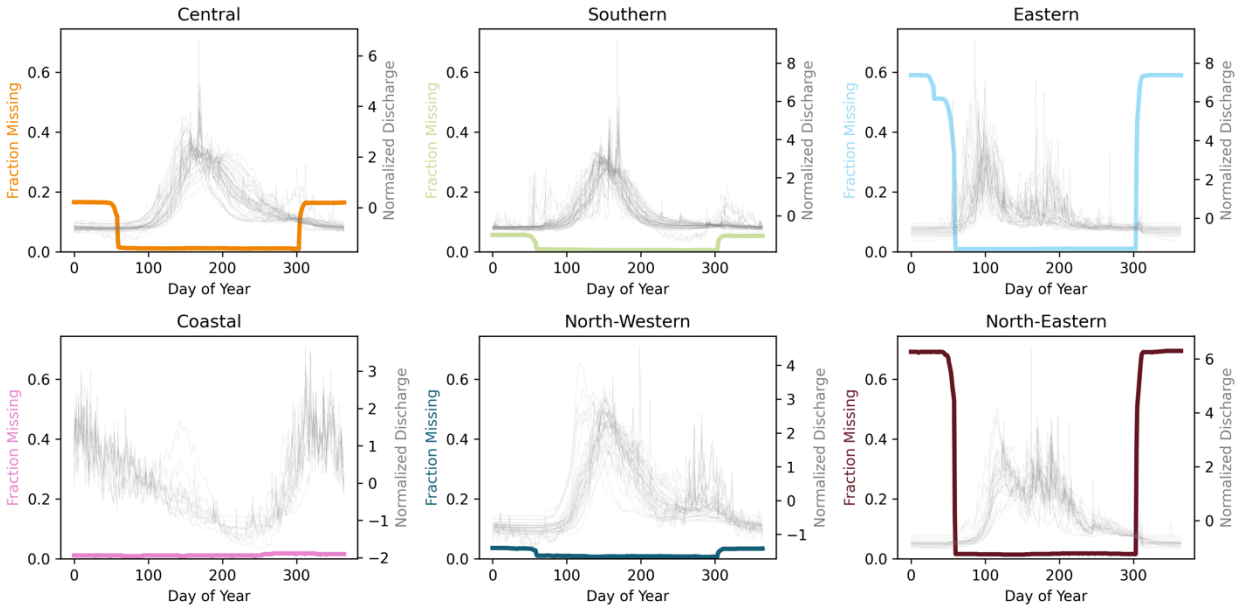
We have fixed this typo and note for clarity:

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 that are classified as natural and that are currently active.”

L140-145: 40% missing data can lead to challenges in model setup. Wondering if the model performance was inferior for basins with missing data than basins with complete data sets?

One challenge is that we need temporally complete datasets since the number of output neurons is constant through training. We recognize that the threshold of 40% missing data may lead to challenges, especially 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 and summer. These data are typically missing at stations which record data seasonally, rather than continuously, and the 40% threshold allows us to include seasonal stations which do not record in winter months. The following figure demonstrates how the most missing data occurs in the low-flow period.

It is true that the two worst performing clusters (eastern and north-eastern clusters; Figure 5) are also those with the most missing data (below). However, the missing data occurs during low-flow periods and as such is not likely to be driving the poor performance, since NSE is largely determined by predictions in spring and summer.



L158-170: As stated earlier including slope and aspect may improve the cluster selection and model performance.

See above response in 'Major Comments'.

Figure 2: State in figure caption how the discharge values are normalized. Similarly, the authors need to provide more details in all Figure captions.

In Figure 2, the following was added to the caption:

Line 297: "Seasonal discharge of each station is normalized to have a mean of zero and unity variance".

Figure captions were expanded as addressed prior to the "Major Comments" above.

L189: It is not clear how the gridded weather data is mapped to the streamflow stations, are the nearest grid cells or mean values from several grid cells taken?

All values of the gridded weather data ("weather video") are mapped to all stream gauge stations (see: Figure 3). The specific mapping from weather-to-streamflow for each stream gauge is learned through training. We clarify this point in Sect. 4.3 when the weather videos are first described in detail (new text in italics):

Line 390: "One year-long weather video is used as an input to predict the next day of streamflow at the 226 stream gauge stations; *in other words, all grid cells of temperature and precipitation are mapped to streamflow at all stream gauge stations.*"

L315-316: Since previous 365 days of data are required, is Jan. 1 1980 is the first day used for streamflow training?

The referee is correct that Jan 1 1980 is the first day of streamflow used. We add:

Line 414: "Since 365 days of previous temperature and precipitation are used to predict streamflow, and since the ERA5 data begins on December 1, 1979, the first day of streamflow used is January 1, 1980."

To clarify the dates for predictors/predictands and to further justify our decisions when creating the training/validation/testing sets:

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."

L364: The spatial perturbation section is hard to follow, how was the amplitude of 1 used in perturbation of climate fields?

To improve the clarity of this section and to show how the amplitude of 1 is used in the perturbation of the climate fields, we added equations to define perturbed daily temperature and precipitation fields ($T_{max,p}$, $T_{min,p}$, and P_p) as:

Line 512:

$$p(x,y) = \beta * e^{\frac{-1}{2}\left[\frac{(x-x_p)^2}{\sigma_x^2}+\frac{(y-y_p)^2}{\sigma_x^2}\right]}$$
$$T_{max,p}(x,y,t) = T_{max}(x,y,t) + p(x,y)$$
$$T_{min,p}(x,y,t) = T_{min}(x,y,t) + p(x,y)$$
$$P_p(x,y,t) = P(x,y,t) + p(x,y)$$

We also write that:

Line 516: "…T_{max}, T_{min}, and P are the unperturbed normalized daily maximum temperature, minimum temperature, and precipitation, respectively."

Line 517: "The amplitude of the perturbation was chosen to be 1 since the climate variables are normalized to have unity variance across the training period. This way each climate variable is perturbed by a maximum of a single standard deviation."

L411: Also say temperature perturbations are constant throughout the time period.

We have edited the text:

Line 567: “To test the hypothesis, we add a spatially *and temporally* uniform temperature perturbation…”.

L425: on what basis/reference was the criteria for freshet timing defined?

One challenge in choosing a definition of “freshet timing/onset” is the number of definitions which are used in the literature, such as:

Zhang et al., 2001: The date when the increase in daily streamflow across 4 days is greater than the average from January to July

Woo and Thorne, 2003: The first day flow is more than double than the flow of the prior day

Burn et al., 2004: The first day flow exceeds 1.5 times the average of the previous 16 days

Vincent et al., 2015: The date when the cumulative sum of the difference between daily mean streamflow and its climatology reaches a minimum in the hydrological year

We found our definition to be robust across the snowmelt-dominated cluster (e.g. cluster-ensemble-mean streamflow fluctuations prior to the spring rising limb were not large enough to be mis-identified as the freshet onset) and, in our view, easier to visually and conceptually understand as meaning “When has spring snowmelt substantially increased flow?”. We now include two additional references (Woo and Thorne, 2003; Burn et al., 2004) when we introduce our definition to emphasize the diversity of freshet definitions used previously in literature (Line 580).

The results in Figure 4b have not been adequately described in the text.

We have added the following description in the text:

Line 738: “The central, coastal, and north-western stations have smaller sensitive areas, while the southern, central, and north-eastern stations have larger sensitive areas (Fig. 4b and Fig. 5c). … Notably, the clusters which are sensitive to the smallest areas of the input (central, coastal, and north-western, Fig. 4b) all experience a substantial decrease in A through fine-tuning (Fig. A4b and Fig. 5c). This indicates that fine-tuning may be necessary for the model to focus on small areas of the input space.”

Figure 6: name the basins for which these example results are presented.

The names and station IDs are now included in the title of Figure 6. Additionally, for all stations used we include station names, numbers, latitude, longitude, and if they are part of the Reference Hydrometric Basin Network in Table S1.

L492: Given that streamflow at a hydrometric station is response to precipitation and temperature over the entire drainage basin, it is to be expected that there are higher sensitivity in response by including areas near and within the station. This need to be clarified, in the context of how big the drainage basins are, and whether the inclusion of precipitation and temperature variables from a wider region improved the model performance.

The referee is correct that a greater sensitivity in response to perturbations is expected near/within the basin; however, this should only be true if the model has learned that the grid cells near/within the basin are more important to streamflow at that station as compared to grid cells which are further away. Since all grid cells are mapped to all stream gauge stations (clarified in response to a point above), the model needs to learn which grid cells are most relevant. The model automatically learns that grid cells near/within basins are more important (i.e. higher sensitivity), which indicates the model's ability to learn physically relevant and interpretable information.

L562: How is the intensity of freshet calculated?

The intensity of the freshet was calculated as the peak spring flow, as defined after Equation 16 in the original submission. However, upon reflection, we have edited the text and now refer to this quantity as the "freshet peak flow", still defined after Equation 16, rather than "intensity".

L562-580: The results in Figures 9 and 10 seem to be consistent with previous climate change impact studies in the region. This is quite promising and is perhaps one of results the authors can highlight further. I suggest expanding the discussion in this section by linking with previous climate impacts studies.

We have added the following paragraph:

Line 907: "When the input temperature series is made warmer (cooler), the freshet onset timing and peak flow advances (delays) and decreases (increases) (Fig. 9 and Fig. 10). This finding is consistent with previous studies of climate change impacts in the region. For example, Shrestha et al. (2012) used the macro-scale Variable Infiltration Capacity (VIC) hydrological model forced by a suite of global climate models in the Fraser River Basin (which spans the central cluster in our study), finding that spring peak flows occur earlier in the year and with lower magnitude under a warmer future climate (Shrestha et al., 2012). Schnorbus et al. (2014) used the VIC model to project streamflow in the Peace, Campbell, and Columbia River watershed in British Columbia (primarily in the north-western, coastal, central, and southern clusters in our study) under a range of climate change scenarios (Schnorbus et al., 2014). They found greater spring flows and lower summer flows in the snowmelt dominated locations, while the coastal location was projected to experience enhanced winter flows and depressed summer flows. It is promising that not only does the CNN-LSTM model perform well in the historical period (e.g. the test period of 2011 - 2015), but produces conceptually similar projections for a warmer climate as compared to existing physically-based models."

L589: Rephrase the sentence, it appears as if previous studies also used deep learning.

The referee's interpretation of that sentence is correct. The studies discussed in that paragraph also used deep learning, but in different regions and with different model architectures.

L626-627: While it is true that the non-contributing areas may have played a part in DL results in parts of eastern cluster, the cited studies are outside of the study region and not directly comparable. Also non-contributing areas may not be a factor for the entire region. There are maps available which outline the extent of non-contributing areas.

We have adjusted our language in this section to emphasize that the effect of non-contributing areas is not the only possible explanation, but is at least a possible explanation to some degree within the eastern cluster.

It is a good point raised by the referee that non-contributing areas may not be a factor for all stations. We further investigate this by calculating the fraction of non-contributing areas for all stations in the eastern cluster using data from the 'Areas of Non-Contributing Drainage within Total Gross Drainage Areas of the AAFC Watersheds Project – 2013' (found at this link, with citation below: https://open.canada.ca/data/en/dataset/adb2e613-f193-42e2-987e-2cc9d90d2b7a). We find that across the eastern cluster:

- Minimum fraction of non-contributing areas across all stations: 0 (only 4 of 34 basins have no non-contributing areas)
- Minimum fraction of non-contributing areas across the 30 basins with non-zero non-contributing area: 0.01
- Mean fraction of non-contributing areas: 0.20
- Maximum fraction of non-contributing areas: 0.79

We add:

Line 882: "In the eastern cluster, 30 out of 34 basins have non-contributing areas, ranging from 1% to 79% of the total basin area with a mean of 20% of the total basin area not contributing to streamflow on average (Government of Canada, 2020)."

Furthermore, we provide code which reproduces this step of the analysis on Github (https://github.com/andersonsam/cnn_lstm_era/blob/master/non_contributing_areas.ipynb).

Table 2 heading: State the period of test set used. Also it should be clarified in the heading that various validation periods were used in reference models.

We have added the period of our test set (2011 – 2015) and that various validation periods were used in the reference models (Line 1010).

Discussion and Conclusions: the changes suggested above also applies to the results and discussion section.

The changes made to the discussion and conclusions sections have been listed above.

References

Moore, R. D., 1991: Hydrology and water supply in the Fraser River basin. Water in Sustainable Development: Exploring Our Common Future in the Fraser River Basin, 21–40.

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

Werner, A. T., M. A. Schnorbus, R. R. Shrestha, A. J. Cannon, F. W. Zwiers, G. Dayon, and F. Anslow, 2019: A long-term, temporally consistent, gridded daily meteorological dataset for northwestern North America. Scientific Data, 6, 180299, https://doi.org/10.1038/sdata.2018.299.

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.

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.

Burn, D. H., Abdul Aziz, O. I. and Pietroniro, A.: A Comparison of Trends in Hydrological Variables for Two Watersheds in the Mackenzie River Basin, Can. Water Resour. J. / Rev. Can. des ressources hydriques, 29(4), 283–298, doi:10.4296/cwrj283, 2004.

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: National hydrometric network basin polygons, [online] Available from: https://open.canada.ca/data/en/dataset/0c121878-ac23-46f5-95df-eb9960753375, 2016.

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.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D. and Alsdorf, D.: The Shuttle Radar Topography Mission, Rev. Geophys., 45(2), doi:10.1029/2005RG000183, 2007.

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.

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

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.

Gauch, M. and Lin, J.: A Data Scientist's Guide to Streamflow Prediction, [online] Available from: http://arxiv.org/abs/2006.12975 (Accessed 10 May 2021), 2020.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J. and Hochreiter, S.: Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network, Hydrol. Earth Syst. Sci., 25(4), 2045–2062, doi:10.5194/hess-25-2045-2021, 2021.

Government of Canada: Areas of Non-Contributing Drainage within Total Gross Drainage Areas of the AAFC Watersheds Project - 2013, [online] Available from: https://open.canada.ca/data/en/dataset/adb2e613-f193-42e2-987e-2cc9d90d2b7a, 2020.

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.

Hussain, D., Hussain, T., Khan, A. A., Naqvi, S. A. A. and Jamil, A.: A deep learning approach for hydrological time-series prediction: A case study of Gilgit river basin, Earth Sci. Informatics, 13(3), 915–927, doi:10.1007/s12145-020-00477-2, 2020.

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.

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, Water Resour. Res., 55(12), 11344–11354, doi:10.1029/2019WR026065, 2019a.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23(12), 5089–5110, doi:http://dx.doi.org/10.5194/hess-23-5089-2019, 2019b.

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.

Marçais, J. and de Dreuzy, J.-R.: Prospective Interest of Deep Learning for Hydrological Inference, Groundwater, 55(5), 688–692, doi:https://doi.org/10.1111/gwat.12557, 2017.

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, Geosci. Model Dev., 13(1), 225–247, doi: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, vol. 21, pp. 2650–2667., 2007.

Schnorbus, M., Werner, A. and Bennett, K.: Impacts of climate change in three hydrologic regimes in British Columbia, Canada, Hydrol. Process., 28(3), 1170–1189, doi:10.1002/hyp.9661, 2014.

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.

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., Fang, Z., Fang, K., Li, D., Li, X. and Tsai, W.-P.: HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community, Hydrol. Earth Syst. Sci., 22(11), 5639–5656, doi:10.5194/hess-22-5639-2018, 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.

Van, S. P., Le, H. M., Thanh, D. V., Dang, T. D., Loc, H. H. and Anh, D. T.: Deep learning convolutional neural network in rainfall–runoff modelling, J. Hydroinformatics, 22(3), 541–561, doi:10.2166/hydro.2020.095, 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.

Vincent, L. A., Zhang, X., Brown, R. D., Feng, Y., Mekis, E., Milewska, E. J., Wan, H. and Wang, X. L.: Observed Trends in Canada's Climate and Influence of Low-Frequency Variability Modes, J. Clim., 28(11), 4545–4560, doi:10.1175/JCLI-D-14-00697.1, 2015.

Woo, M.-K. and Thorne, R.: Streamflow in the Mackenzie Basin, Canada, Arctic, 56(4), 328–340 [online] Available from: http://www.jstor.org/stable/40513072, 2003.

Zhang, X., Harvey, K. D., Hogg, W. D. and Yuzyk, T. R.: Trends in Canadian streamflow, Water Resour. Res., 37(4), 987–998, doi:https://doi.org/10.1029/2000WR900357, 2001.