https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Evaluation and interpretation of convolutional-recurrent networks for regional hydrological modelling

Sam Anderson[1], Valentina Radic[1]

[1]Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, V6T 1Z4, Canada

Correspondence to: Sam Anderson (sanderson@eoas.ubc.ca)

Abstract. Deep learning has emerged as a useful tool across geoscience disciplines; however, there remain outstanding questions regarding the suitability of unexplored model architectures and how to interpret model learning for regional scale hydrological modelling. Here we use a convolutional-recurrent network, a deep learning approach for learning both spatial and temporal patterns, to predict streamflow at 226 stream gauges across the region of southwestern Canada. The model is forced by gridded climate reanalysis data and trained to predict observed daily streamflow between 1979 and 2015. To interpret the model learning of both spatial and temporal patterns, we introduce two experiments with evaluation metrics to track the model's response to perturbations in the input data. The model performs well in simulating the daily streamflow over the testing period, with a median Nash-Sutcliffe Efficiency (NSE) of 0.68 and 35% of stations having NSE > 0.8. When predicting streamflow, the model is most sensitive to perturbations in the input data prescribed near and within the basins being predicted, demonstrating that the model is automatically learning to focus on physically realistic areas. When uniformly perturbing input temperature timeseries to obtain relatively warmer and colder input data, the modelled freshet timing and intensity changes in accordance with the transition timing from below- to above-freezing temperatures. The results demonstrate the suitability of a convolutional-recurrent network architecture for spatiotemporal hydrological modelling, making progress towards interpretable deep learning hydrological models.

1 Introduction

The use of deep learning (DL) has gained traction in geophysical disciplines (Bergen et al., 2019; Gagne II et al., 2019; Ham et al., 2019), including hydrology, providing alternative or complementary approaches to supplement traditional process-based modelling (Hussain et al., 2020; Kratzert et al., 2018, 2019a, 2019b; Marçais and de Dreuzy, 2017; Shen, 2018; Van et al., 2020). While substantial progress has been made towards distributed process-based hydrological models, input and target data are becoming available at increasingly higher spatiotemporal resolution leading to greater computational requirements and human labour (Marsh et al., 2020). DL in hydrology has emerged in recent years as an active field of exploration in efforts to maximize the use of growing in situ and remote sensing datasets (Kratzert et al., 2018; Shen, 2018; Shen et al., 2018).

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

Early applications of machine learning in hydrology date back to the 1990s, with artificial neural network (ANN) models used for rainfall-runoff modelling (Maier and Dandy, 1996, 2000; Zealand et al., 1999). ANN models aim to approximate functions that connect input data (e.g. weather data), represented by input neurons, to output or target data (e.g. streamflow data), represented by output neurons, through a series of hidden layers, each containing hidden neurons. The training of these models, i.e. the tuning of model parameters in the functions interconnecting each layer, aims to minimize the distance between model output and observed target data. However, ANNs struggle to represent more complex temporal or spatial information, and so ANN models are often highly tuned for a specific basin and cannot be applied elsewhere. In contrast, the design of DL, generally referring to large multi-layer networks applied directly to large, raw datasets, has led to both improved representation of more complex functions and better learning of spatial and/or temporal patterns within the data (Shen, 2018).

As a type of DL architecture, long short-term memory (LSTM) networks are designed to learn sequential relationships on a range of scales. Learning is achieved by including four gates through which information can flow, the outputs of which interact with the memory state of the network. LSTMs have had particular success in natural language processing (NLP), including applications of text prediction (Karpathy et al., 2015), language translation (Sutskever et al., 2014), image captioning (Kiros et al., 2014), and video-to-text conversion (Venugopalan et al., 2015). In hydrology, Kratzert et al (2018) demonstrated the effectiveness of LSTMs for rainfall-runoff modelling, using the previous 365 days of basin-averaged weather variables to predict the next day of streamflow at a stream gauge station. They have shown that a single LSTM can be trained on hundreds of basins, and then fine-tuned for either each region or each basin, oftentimes outperforming standard hydrological models. Additionally, LSTM models trained on many basins have been shown to also outperform standard hydrological models for prediction at ungauged basins, demonstrating the potential for LSTM models to be used as regional hydrological models (Kratzert et al., 2019a). However, while addressing the need to learn complex sequential information, this approach does not explicitly consider spatial information, and as such has been primarily used as a lumped hydrological model with basin-averaged values or point-observations as input.

Another type of DL architecture, convolutional neural networks (CNNs) are specifically designed to learn spatial information. Learning is achieved through convolving an input with a layer of filters made up of trainable parameters. The development of CNNs was largely driven by image classification applications (Krizhevsky et al., 2012). In the geosciences, CNNs have gained popularity only relatively recently with applications including long-term El-Nino forecasting (Ham et al., 2019), precipitation downscaling (Vandal et al., 2017), and urban water flow forecasting (Assem et al., 2017). Importantly, CNNs have been combined with LSTMs to encode both spatial and temporal information. Sequential CNN-LSTM models have been used to map input videos to class labels or text captions, where frames in the video are passed first through a CNN, the output of which is then passed through an LSTM (Donahue et al., 2017). Alternatively, LSTM models with convolutional rather than fully connected (or 'dense') layers have also been used to encode spatiotemporal information for applications including precipitation nowcasting (Shi et al., 2015). In hydrology, CNN (and particularly combined CNN-LSTM) models have seen fewer applications to date as compared to the LSTM approach, with recent work developing 1D CNNs for rainfall-runoff modelling (Hussain et al., 2020; Van et al., 2020).

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

Historically, hydrological model development has emphasized understanding and incorporating physical processes in order to improve model performance (Freeze and Harlan, 1969; Hedstrom and Pomeroy, 1998; Marsh et al., 2020; Painter et al., 2016; Pomeroy and Gray, 1990). Considering the emphasis on process-based modelling within the hydrological community (Bahremand, 2016) and the multifaceted challenges surrounding water management (Milly et al., 2008; Wheater and Gober, 2013), it is important that DL-based hydrological models are interpretable and trustworthy in addition to being successful in simulating accurate streamflow. How a model is interpreted, and what it means to interpret a DL model, depends on the model architecture (e.g. ANN, CNN, LSTM), the task the model is performing (e.g. regression, classification), and the research questions being asked with the model. Several methods to interpret DL models have been developed, as outlined below.

One approach to interpret CNN models is to visualize the regions in the input that are most important for decision making, which can be done for both classification and regression problems. Techniques such as class activation mapping (CAM) and gradient class activation mapping (Grad-CAM) utilize deep feature maps to determine the regions most important for classification (Selvaraju et al., 2020). Another technique, layerwise relevance propagation (LRP), backpropagates from a single output neuron through the trained network to identify the input region which is most relevant for determining the value of the output neuron (Bach et al., 2015). For LRP, the propagation rules used depend on model architecture (Arras et al., 2019; Bach et al., 2015; Toms et al., 2020). In contrast to these 'white-box' approaches that interpret the model through explicit use of the model parameters, 'black-box' approaches do not utilize internal network states for interpretation. For example, techniques such as occlusion (Zeiler and Fergus, 2014) and randomized image sampling explanation (RISE) (Petsiuk et al., 2018) iteratively gray- or zero-out subsets of an input image and measure how a model's predictions change due to this perturbation. Occlusion and RISE can identify the area in the input where the model's predictions are most sensitive to perturbation, which can be interpreted as being the most important information for the model to have in order to make its prediction.

Recurrent networks can be challenging to interpret as the relevance of any feature in the network depends on the processing of previous features. LSTMs have often been interpreted by analysing their internal states (Shen, 2018). For example, Karpathy et al (2015) visually inspect cell states of an LSTM trained for natural language processing applications to identify states which track various recognizable text features, such as quotations and line length. Most states, however, were found to be uninterpretable (Karpathy et al., 2015). A similar approach has been taken for interpreting LSTMs in hydrology; for example, Kratzert et al. (2018) discuss cell states as being comparable to storages in traditional hydrological models. They show that the evolution of one cell state closely resembles the dynamics of a snowpack, increasing when temperatures are below freezing and quickly depleting when temperatures rise above freezing (Kratzert et al., 2018). More recently, LRP has been adapted for the LSTM architecture (Arras et al., 2019); however, to our knowledge there are no examples of its use in the geoscientific literature.

We aim to create a relatively simple and interpretable DL model which maps spatiotemporal weather fields, represented by gridded climate data at a relatively coarse (~75 km) spatial resolution, to streamflow at multiple stream gauge stations

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

across a region. By explicitly encoding spatial information, we aim to develop a DL analogue of a distributed hydrological model on a regional scale which predicts streamflow on a regional scale without the need for climate downscaling. Our specific objectives for this paper are to: 1) evaluate how well the sequential convolutional-recurrent model performs when predicting daily streamflow simultaneously at multiple stream gauge stations across a region, 2) investigate if the model has learned to focus on the areas within or near the watersheds where streamflow is being predicted, and 3) investigate if the model has learned physically interpretable temporal links which drive the timing and intensity of the spring freshet in snowmelt dominated basins. The first objective is related to evaluating the accuracy of the model's predictions, while the latter two objectives relate to model interpretability. We do not undergo an exhaustive parameter search to create the best or most complex model; rather, we develop a model with relatively few predictor variables which is sufficient for achieving these objectives.

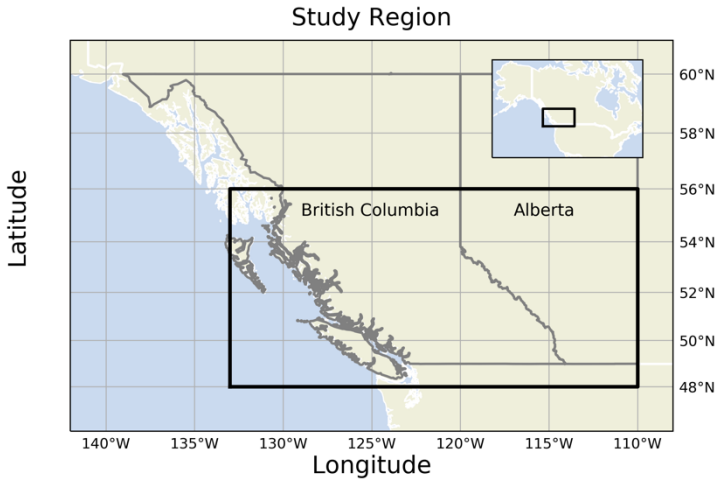
The paper is structured in the following way: in Sect. 2, we discuss the study region. In Sect. 3, we outline the datasets used, and detail our decision making for choosing the input and output variables. In Sect. 4, we outline our methods, including the architecture, training, and evaluation of the model, and describe the experiments developed for interpreting the model's learning. In Sect. 5, we present and discuss the results of our analysis, and we present a summary and conclusions in Sect. 6.

2 Study region

We chose the south-central domain of Western Canada as our study region, containing large sections of the provinces of British Columbia (BC) and Alberta (AB) (Fig. 1). This region contains a range of hydroclimate regimes, allowing for our modelling approach to be evaluated across a range of conditions. In winter, relatively warm and moist Pacific air is advected into the study region, leading to frequent rainfall events at low elevations along the west coast of British Columbia where the maritime climate is wetter and milder as compared to much of the rest of the study region. While most precipitation typically falls as rain at lower elevations, substantial winter snowfall leads to seasonally continuous snow and substantial glacier coverage at higher elevations in the coastal region (Trubilowicz et al., 2016). Cooler winter conditions through much of the rest of the province allow for the accumulation of a seasonal snowpack (Moore et al., 2010). In contrast, winters in Alberta are colder and drier given the influence of Arctic air masses. Substantial snowfall can occur in Alberta when comparably moist Pacific air crosses the Rockies and interacts with cold and dry Arctic air (Sinclair and Marshall, 2009; Vickers et al., 2001), but most precipitation in Alberta falls as rain in the spring and early summer. The seasonal streamflow characteristics are described in Sect. 3.2.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU



130

Figure 1: The study region in Western Canada. The black box outlines the study region in both the main figure and the inset. The provincial borders of British Columbia and Alberta are shown in grey. The inset shows the broader context of the study region in North America. Made with the Python library Cartopy (Met Office, 2018) with data from Natural Earth.

3 Data

135 ### 3.1 Streamflow data

We use daily streamflow data (in [m^3 s^{-1}]) extracted from Environment and Climate Change Canada's Historical Hydrometric Data website (HYDAT) (Environment and Climate Change Canada, 2018). We only use stations which are upstream of dams (measuring naturalized flows) and which are currently active. Many stream gauges do not record data every day of the year, so we only select stream gauges which have no more than 40% of daily data missing between 1979 and 2015
140 and for which no more than 1 year is missing more than 40% of data. For all missing data, we fill the daily value with the average value of that day between 1979 – 2015. If all years are missing that day (which is true for some stations which do not record in low flow periods), we fill the missing day with the minimum seasonal streamflow. The threshold of 40% is chosen to allow for relatively dense spatial coverage of stations across the study region and is acceptable considering that most missing data are during low-flow seasons when rainfall and snowmelt are not strongly driving streamflow dynamics. It is acceptable
145 to allow for one year with greater than 40% missing data because it substantially increases the station density. There are 279 stream gauge stations in Alberta which are active and have naturalized flow. Of these, 120 meet the aforementioned criteria; however, only 66 meet the stricter criteria of having all years with less than 40% missing data. In BC, there are 288 active and naturalized stream gauges; of these, 145 meet the less strict criteria and 108 meet the stricter criteria. Missing data is a common

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

feature in geoscientific datasets which poses a challenge for the use of machine learning models (Karpatne et al., 2019), and so creating a suitably large training dataset may require pre-processing steps like those outlined above.

We further restrict the study region to stations south of 56° N because stream gauge density is greater below this latitude. Out of the total 265 stations that meet the less strict criteria, 226 stream gauge stations are south of 56° N. This means that north of 56° N, 15% of stream gauges span 33% of the study region, and so this restriction reduces memory and computational requirements, as well as increases the spatial stream gauge density. As our goal is not to develop a model at continental scale, but to investigate how well the model can learn different streamflow regimes, the above-described data selection process is justified for our application. The 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.

3.2 Streamflow clusters

To optimally facilitate the investigation of the model's learning with the methods described in Sect. 4, we cluster the stations in our study region into subdomains based on both similarity in streamflow regime and proximity in space. The clustering input (observation) for each station is a vector where the first 365 dimensions are the seasonal streamflow, the next 182 dimensions are repeated values of latitude, and the final 182 dimensions of repeated values of latitude (Fig. A1 in Appendix A). This clustering input is designed to give seasonal streamflow and geographic location similar weight in the clustering algorithm, where approximately one half of the input is seasonal streamflow, one quarter is latitude, and one quarter in longitude. By clustering in this way, the stream gauges that belong to the same cluster are likely to have similar streamflow and experience similar climatic conditions. Seasonal streamflow is normalized at each station to have zero mean and unity variance, while latitude and longitude are each normalized across all stations to have a mean of zero and unity variance. We use agglomerative hierarchical clustering with Ward's method (Hastie et al., 2009) to identify six subdomains or clusters of streamflow (Fig. 2). The number of clusters chosen (six in this case) is determined from the dendrogram (Fig. A2). We refer to the clusters as north-western, north-eastern, central, southern, eastern, and coastal, as labelled in Fig. 2.

There are key differences between the streamflow regimes identified by the clustering (Fig. 2). Only the lower-elevation coastal stream gauges are characterized by low summer flows and high winter flows which are driven by winter rainfall events; all other clusters differ from one another largely in the timing and intensity of both the spring freshet (the first streamflow peak in a year) and a second rainfall-driven peak occurring in spring, summer, or fall. The eastern and north-eastern clusters are characterized by relatively early spring freshet, followed by rainfall-driven streamflow peaks in early summer. The southern, central, and north-western stations are characterized by a later and more sustained spring runoff, in part due to a longer-lasting snowpack which accumulates from the relatively higher rates of winter precipitation in British Columbia.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

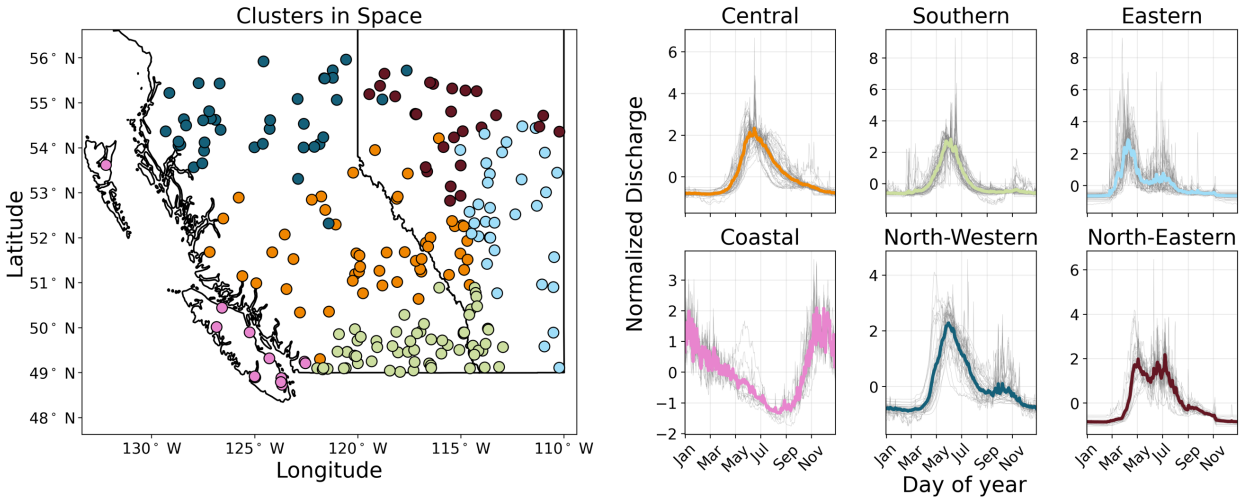


Figure 2: The seasonal streamflow cluster patterns and their locations in space. The colour of the stream gauge in the left panel corresponds to the seasonal streamflow cluster pattern on the right. The background grey curves in the cluster pattern panels are the cluster members.

3.3 Weather data

As input weather variables to the model, we select daily fields of precipitation, maximum temperature, and minimum temperature, extracted from ERA5 reanalysis (Hersbach et al., 2020) from the European Centre for Medium-Range Weather Forecasts (ECMWF). Data are downloaded at 6-hourly temporal resolution and 0.75° x 0.75° spatial resolution for the time period 1979 – 2015. Our selection of variables is based on the assumption that the combination of precipitation and temperature is sufficient for estimating both how precipitation contributes to streamflow as rainfall, and for estimating the onset, intensity, and longevity of the spring freshet through the seasonal accumulation and ablation of a snowpack. We recognize that the underlying physics which governs streamflow throughout the year is more complex than these comparably simple assumptions (e.g. interactions between surface water and ground water (Hayashi et al., 2016), evapotranspiration (Penman and Keen, 1948), snow redistribution from wind (Pomeroy and Li, 2000)); however, we are assuming that temperature and precipitation from reanalysis data can act as proxies from which most of the information can be inferred (e.g. Essou et al., 2016). While additional variables could be used as climatic drivers of streamflow (e.g. solar radiation, evaporation, wind), we opt to use a simpler model with fewer input variables as a proof of concept and to achieve our goals stated in Sect. 1.

ERA5 reanalysis was chosen as it is globally available from 1979 through the present (once complete it will be available from 1950 onwards), and has been shown to compare well against other reanalysis products (Hersbach et al., 2020).

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

Importantly, the precipitation output from ERA5 has been found to typically outperform the earlier ERA-Interim reanalysis in the northern Great Plains region, which experiences a similar climate to the Prairie region in our study area (Xu et al., 2019). We downloaded total precipitation (variable name: 'tp'; parameter ID: 228) and near-surface air temperature (variable name: '2m Temperature'; parameter ID: 500011), from which daily total precipitation, daily maximum temperature, and daily minimum temperature are calculated.

4 Methods

Here we summarize our methods before providing details of each key step. We use a sequential CNN-LSTM model to map weather predictors to streamflow at multiple stream gauge stations simultaneously. As input data, we use the past 365 days of weather, covering the whole study region, in order to predict streamflow of the next day at N stream gauge stations (Fig. 3 and Table 1). The CNN learns the spatial features in each day of the input, while the LSTM model learns the temporal relationship between these features in order to predict streamflow. After the model is trained, we evaluate its performance against the observed streamflow over a testing period, which is independent of the training period. Finally, we introduce two experiments to investigate the model's learning. The first experiment is focused on interpreting the learning of spatial links between the predictors and streamflow, while the second experiment is focused on the learning of links between temperature and the snowmelt-driven freshet. This section will provide a brief overview of both the CNN and LSTM architectures, followed by a description of our CNN-LSTM model design and training, and finally with a description of metrics and experiments developed for the model evaluation and interpretation.

Table 1: Details of the model layers.

Layer Type	Description	Output Shape	Number of Parameters
Convolutional	32 filters, 1x1 size	365 x 11 x 31 x 32	128
Convolutional	16 filters, 3x3 size	365 x 11 x 31 x 16	4624
Convolutional	16 filters, 3x3 size	365 x 11 x 31 x 16	2320
Max Pooling	2x2 pool size	365 x 5 x 15 x 16	0
Convolutional	32 filters, 3x3 size	365 x 5 x 15 x 32	4640
Convolutional	32 filters, 3x3 size	365 x 5 x 15 x 32	9248
Max Pooling	Global	365 x 32	0
Dropout	0.1 dropout rate	365 x 32	0
LSTM	80 units	80 x 1	36160
Dense	As many neurons as stream gauges	N x 1	N*81

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

4.1 CNN overview

The CNN is constructed using two main types of layers: convolutional and pooling. Convolutional layers are made up of multiple filters, which are constructed by trainable weights. Each filter convolves across an input layer to produce an output image which is then passed through a nonlinear activation function. Mathematically, a single output neuron can be calculated from a single filter as:

$$y_{CNN} = g\left(\sum_{i,j,k} W_{CNN}^{i,j,k} x_{CNN}^{i,j,k} + b_{CNN}\right) \quad (1)$$

where y_{CNN} is the value of one neuron in the output layer, g is the nonlinear activation function, W_{CNN} are the weights of the filter, x_{CNN} is the region of the input layer, b_{CNN} is the bias value of the output neuron, and i, j, and k correspond to width (e.g. number of pixels along the x-direction of the image), height (e.g. number of pixels along the y-direction of the image), and depth (e.g. number of channels of the image) of the input, respectively. Pooling layers reduce image resolution, which reduces memory requirements of the network; for example, a 2x2 max-pooling layer will reduce the number of pixels by a factor of 4 by outputting only the maximum value of each 2x2 region of the input. CNN architectures often have a repeating structure of several convolutional layers followed by pooling layer. Through training, the convolutional layers learn the spatial features present with more abstract features being learned at deeper layers, and the pooling layers reduce images to smaller and smaller sizes. The output feature vector is encoded with the learned spatial information from the input.

4.2 LSTM overview

The LSTM network output is determined by the interaction between two internal states: the cell state $c(t)$ which acts as the memory of the network, and the hidden state $h(t)$ which is an intermediate output of the network. Both states are updated at each time step t ($1 \leq t \leq n$) by a series of gates through which information can flow: the forget gate f_t, the input gate i_t, the potential cell update $\tilde{c}_t$, and the output gate o_t. Each time step of the input is concatenated with the hidden state as calculated in the prior time step before being passed through the network; in this way, learned information from previous time steps is used to calculate the next output. In the following equations, weights ($\mathbf{W}$) and biases ($\boldsymbol{b}$) are the learnable parameters in the network:

$$\boldsymbol{f_t} = \sigma\left(\mathbf{W_f}[\boldsymbol{x_t}, \boldsymbol{h_{t-1}}] + \boldsymbol{b_f}\right) \quad (2)$$
$$\boldsymbol{i_t} = \sigma(\mathbf{W_i}[\boldsymbol{x_t}, \boldsymbol{h_{t-1}}] + \boldsymbol{b_i}) \quad (3)$$
$$\tilde{\boldsymbol{c}}_{\boldsymbol{t}} = \tanh(\mathbf{W_c}[\boldsymbol{x_t}, \boldsymbol{h_{t-1}}] + \boldsymbol{b_c}) \quad (4)$$
$$\boldsymbol{o_t} = \sigma(\mathbf{W_o}[\boldsymbol{x_t}, \boldsymbol{h_{t-1}}] + \boldsymbol{b_o}) \quad (5)$$

where x_t is the input vector to the LSTM at time t, tanh is a hyperbolic tangent function, σ is a sigmoid function, and square brackets indicate concatenation. The cell state at time t is determined by the prior cell state and the interactions with the outputs of the forget, input, and potential cell update, while the hidden state at time t is determined by the new cell state and the output gate:

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

$$\boldsymbol{c_t} = \boldsymbol{c_{t-1}} \odot \boldsymbol{f_t} + \boldsymbol{i_t} \odot \widetilde{\boldsymbol{c_t}} \quad (6)$$

$$\boldsymbol{h_t} = \tanh(\boldsymbol{c_t}) \odot \boldsymbol{o_t} \quad (7)$$

where $\odot$ denotes elementwise multiplication. The final hidden state, h_n, is passed through a dense layer constructed of fully connected neurons. The activation of this dense layer is linear, and so the final output is a linear transformation of the final hidden state:

$$\boldsymbol{y_{flow}} = \mathbf{W_d}\boldsymbol{h_n} + \boldsymbol{b_d} \quad (8)$$

This output, y_{flow}, is a vector of normalized streamflow for a single day at multiple stream gauge stations. The length of y_{flow} is encoded with the learned sequential information from the input series.

4.3 Sequential CNN-LSTM architecture

An overview of the model architecture is shown in Fig. 3, and information of each layer is presented in Table 1. We use a sequential CNN-LSTM model in order to simultaneously map the previous 365 days of temperature and precipitation to the next day of streamflow at multiple stream gauges throughout the study region (i.e. days 1 through 365 of weather are used to predict day 366 of streamflow). Daily weather images are constructed where the height and width of the image correspond to the number of grid cells along latitude and longitude, respectively, and with three channels corresponding to normalized maximum temperature (T_{max}), minimum temperature (T_{min}), and precipitation (P). Yearly weather videos are constructed from the past 365 days of weather images, where each frame in the video is a weather image. One year-long weather video is used as an input to predict the next day of streamflow at the 226 stream gauge stations. Each frame in the video is passed independently through the CNN, which converts each of the 365 frames into a feature vector of length 32. This feature vector is a representation of the learned spatial features found in that frame of weather. There are 365 feature vectors generated from one year-long weather video, since there are 365 days in the input video. Then, this series of feature vectors is passed through an LSTM, which learns the sequential relationship between the learned spatial features and outputs a final hidden vector, h_n (Equation 7), with length 80. This hidden vector contains information of the sequential relationships between the spatial features and is next passed through a dense layer with linear activation to connect to the final output neurons (y_{flow}, Equation 8). In other words, the 80 values in the hidden vector are linearly combined to predict a single day of streamflow at each individual station.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

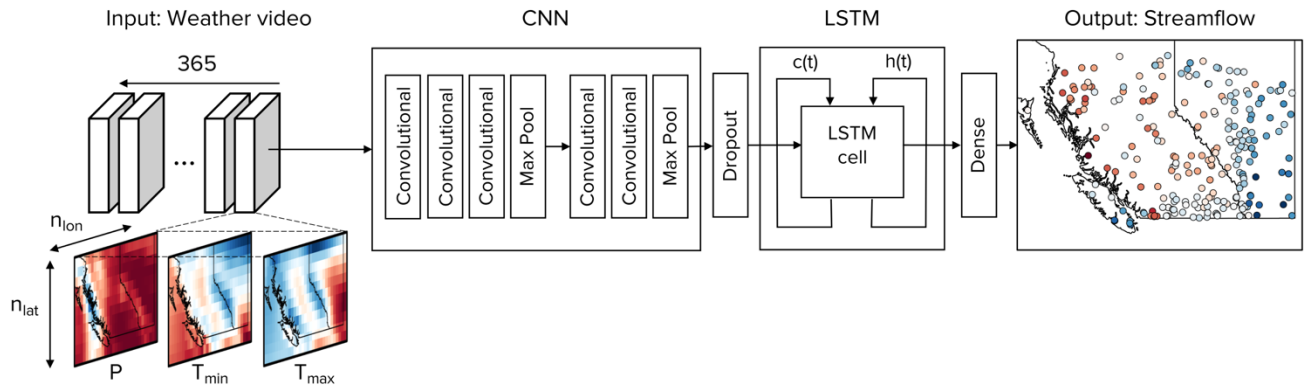


Figure 3: Overview of the model architecture. The model input is a weather video with 365 frames/images, each corresponding to one day of weather from ERA5 reanalysis in the past year. Each frame in the video has three channels corresponding to precipitation (P), maximum temperature (T_{max}), and minimum temperature (T_{min}). Each channel has dimensions of $n_{lat} \times n_{lon}$, where n_{lat} is the number of grid cells in the vertical direction (latitude) and n_{lon} is the number of grid cells along the horizontal direction (longitude). Each frame in the weather video is passed through a CNN, and each weather video generates a sequence of 365 feature vectors. A dropout layer is used between the CNN and LSTM for regularization. The sequence of 365 feature vectors is then passed through an LSTM and a dense linear transformation to output the next day of modelled streamflow at N stations (i.e. streamflow at day 366 at N stations). Within the LSTM cell, $c(t)$ is the cell state and $h(t)$ is the hidden state.

Training the DL model requires a balance of having sufficient complexity to learn the mapping from weather to streamflow, but without being so complex that the model overfits to the training set and performs poorly on the validation set. With that in mind, we designed this architecture (Table 1) considering the following: 1) the number of pooling layers is limited by the relatively small input images (spatial size of 12 x 32); 2) the number of filters in the deepest layer of the CNN determines the length of the spatial feature vector (input to LSTM); and 3) the number of parameters in a single LSTM layer goes linearly with the length of the spatial feature vector and quadratically with the number of LSTM units. In addition to these general guiding principles, we found that a single LSTM layer with more units performed better than multiple LSTM layers with fewer units, as had previously been used when predicting streamflow at a single station (Kratzert et al., 2018). The success of a single LSTM layer with more units is likely because we map the LSTM hidden state to multiple stream gauges (a higher-dimensional space) rather than a single neuron, and so more units are required for this mapping to work well. Additionally, we found that including 32 filters of size 1x1 as the first layer improved model performance.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

4.3.1 Training

We use fine-tuning (Razavian et al., 2014) to train our model in two steps:

1. Bulk training: a model is trained on all 226 stream gauge stations in the region.
2. Fine-tuning: the bulk model is further trained at stations from each of the six clusters (Fig. 1) in the following way. The bulk model is copied six times, with one copy used for each cluster, but the last dense layer in the bulk model is removed and replaced with a new dense layer which has as many neurons as stations in that cluster. Weights in the new dense layer are randomly initialized. Each fine-tuned model is then trained further on only the stations in that cluster.

For both bulk and fine-tuned models, early stopping is used to reduce overfitting. We use a dropout layer with a dropout rate of 0.1 between the CNN and LSTM layers for regularization (Srivastava et al., 2014). We use batch sizes of 64, a learning rate of 10^{-4}, mean squared error loss, and Adam optimization (Kingma and Ba, 2017). For all models, we use 1979 – 2000 for training, 2001 – 2010 for validation, and 2011 – 2015 for testing. We use the Keras (Chollet, 2015) and Tensorflow (Abadi et al., 2016) libraries in Python (Van Rossum and Drake, 2009), and we use Google Colab for access to a cloud GPU. In total, we train an ensemble of 10 bulk models and further fine-tune each one, yielding an ensemble of 10 bulk models and an ensemble of 10 fine-tuned models per cluster. Training a single bulk model on a single cloud GPU in Colab takes on the order of tens of minutes.

4.4 Evaluation of model performance

We evaluate how well streamflow is simulated by the bulk and fine-tuned models with the Nash-Sutcliffe Efficiency (NSE) (Nash and Sutcliffe, 1970). For each station, we calculate NSE over the test period for both the bulk and fine-tuned models, using the ensemble mean as the final model output. NSE is defined as:

$$NSE = 1 - \frac{\sum_{t=1}^{t=T}(Q_m^t - Q_o^t)^2}{\sum_{t=1}^{t=T}(Q_o^t - \overline{Q_o})^2} \quad (9)$$

where T is the total number of time steps in the test series, Q_m^t is the modelled streamflow for that station at that time, Q_o^t is the observed streamflow for that station at that time, and $\overline{Q_o}$ is the mean observed streamflow for that station over the whole test period. The overall performance of both the bulk and fine-tuned models is evaluated by the median NSE of all stations as evaluated over the test period. When $NSE = 1$, the modelled streamflow is exactly equal to the observed streamflow, while $NSE < 0$ indicates very poor model performance as more variability would be captured if the streamflow was represented with its mean value than with the modelled streamflow.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

4.5 Interpretation of model learning

4.5.1 Spatial perturbations

We interpret the model's learning of spatial links by testing the following hypothesis: if the model is learning physical processes that drive streamflow at a given stream gauge station, then the modelled streamflow at that station should be most sensitive to perturbations in the watershed or vicinity of that station, and less sensitive to perturbations further from that station. To test this hypothesis, we perturb small spatial regions of the input weather video and determine how sensitive the predicted streamflow of each cluster of stream gauge stations (Fig. 1) is to the areas which are perturbed. To evaluate the regions of the input space which are most important for streamflow predictions at each stream gauge, we take the following steps, each of which will be discussed in more detail:

1. Perturb the input video
2. Evaluate how much the modelled streamflow prediction changes at each station
3. Define a sensitivity heat map for this perturbation for each station
4. Iterate through steps 1 – 3 for each day in the test series until the heat map no longer substantially changes from further perturbations
5. Evaluate if the sensitive areas are representative of physically realistic learning for each streamflow cluster

Steps 1 – 4 are similar to the occlusion (Zeiler and Fergus, 2014) and RISE (Petsiuk et al., 2018) algorithms in that we iteratively perturb the input and generate sensitivity heat maps based on how the output changes. The RISE approach zeroes out portions of the input image, which here would be equivalent to setting a portion of the input to be the mean weather values since the input variables are normalized to have zero mean; therefore, the difference between the perturbed and unperturbed input would depend on how close the input variables are to their mean values in each day. Instead, here we perturb the input by adding or subtracting a 2D gaussian curve from the input video, which alters each day in the input no matter if it is near the mean or not. We developed this method, as opposed to using already established methods such as occlusion, RISE, or LRP, because our method is both agnostic of model architecture, and is grounded in a physical understanding of the key processes taking place (i.e. the perturbations are adding in synthetic warm/wet or cold/dry areas and we determine if and how this changes streamflow in the perturbed basins).

In step 1, we define a gaussian perturbation, p, as:

$$p(x,y) = \beta * e^{-\frac{1}{2}\left[\frac{(x-x_p)^2}{\sigma_x^2}+\frac{(y-y_p)^2}{\sigma_x^2}\right]} \quad (10)$$

where x and y are longitude and latitude (in degrees), x_p and y_p are the longitude and latitude of a randomly selected point within the study domain (in degrees), σ_x and σ_y are the standard deviations of the gaussian in the x- and y-directions (in degrees), and β is a multiplicative factor which has equal probability of being either 1 or -1 for each perturbation. The gaussian has an amplitude of 1 and standard deviations are 1.5 pixels in both the x- and y-direction. The amplitude determines the strength of the perturbation and the standard deviations determine the extent. The amplitude was chosen to be 1, equating to

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

a maximum perturbation of a single standard deviation of each input variable. σ_x and σ_y were chosen so that the gaussian perturbation is small relative to both the height and width of the input weather frame, but larger than a majority of basins. This perturbation is added to every channel (predictor variable) and frame in the input video. An example of a perturbation, a perturbed maximum temperature field, and the perturbed streamflow response is shown in Fig. A3.

In step 2, we pass the perturbed video through the trained model, and calculate the absolute value of the difference between the unperturbed and the perturbed modelled streamflow for each stream gauge. From this difference at each station, we can quantify how important the perturbed area is for the model's decision making. We quantify the importance in step 3 by defining a sensitivity heat map for each stream gauge station:

$$s^i(x,y) = \left|Q_m^i - Q_{m,p}^i\right| * p(x,y) \tag{11}$$

where $s^i(x,y)$ is the sensitivity heat map of stream gauge i, Q_m^i is the unperturbed modelled streamflow at the stream gauge i, $Q_{m,p}^i$ is the perturbed modelled streamflow at the stream gauge i, and $p(x,y)$ is the perturbation. Each perturbation produces 226 different sensitivity heat maps, corresponding to one heat map for each stream gauge station.

In step 4, we iterate through the first three steps for each day in the test series until the mean sensitivity heat maps converge, here taken as when the relative error between heat maps of subsequent perturbations falls below 0.5%. Then, for each streamflow cluster, we calculate the mean sensitivity heat map across all stream gauges in the cluster, all iterations, and all days in the test series as:

$$S(x,y) = \frac{1}{N} * \frac{1}{m} * \frac{1}{q} \sum_{k=1}^{q} \sum_{j=1}^{m} \sum_{i=1}^{N} s^{i,j,k}(x,y) \tag{12}$$

where $S(x,y)$ is the mean heat map, q is the number of stream gauges in this cluster, m is the number of days in the test set, N is the number of iterations, and $s^{i,j,k}(x,y)$ is the sensitivity heat map corresponding to iteration i of day j at stream gauge k.

Finally, in step 5, we identify the values in the cluster heat map which are either: 1) within the watershed or within 1 pixel of distance from the watershed boundaries, or 2) further than 1 pixel in distance from the watershed boundaries (where 1 pixel has size 0.75° x 0.75°). If the model is focusing on the areas which are within or near the cluster's basins, then we expect the sensitivity within or near the basins to have a higher mean sensitivity and a substantially different distribution of sensitivity than the distribution of sensitivity outside or far from the basins. To evaluate how different the distributions of sensitivity are, we calculate the Kolmogorov-Smirnov D-statistic (Chakravarti et al., 1967) to compare the distribution of heat map pixels which are within or near the cluster's watershed boundaries, with the distribution of pixels which are not within or near the cluster's watershed boundaries (i.e. all other pixels in the domain). The D-statistic is a measure of how different two distributions are, where a value of 0 indicates perfectly overlapping distributions, while a value of 1 indicates entirely non-overlapping distributions. The D-statistic, D, is calculated as:

$$D = \max\left|F_{in}(S) - F_{out}(S)\right| \tag{13}$$

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

where $F_{in}(S)$ is the cumulative density function (CDF) of sensitivity within/near the cluster's watersheds, and $F_{out}(S)$ is the CDF of sensitivity outside/far from the cluster's watersheds. Watershed boundaries are accessed through the Water Survey of Canada (Environment and Climate Change Canada, 2016).

Additionally, we characterize the heat maps by the value A, here defined as the area fraction of the sensitivity heat map which is more than the half-maximum sensitivity. Smaller values of A (closer to 0) indicate that the model is focused on a smaller portion of the input area, while large values of A (closer to 1) indicate that larger portions of the input video are important for the model's prediction at that station. D and A are calculated from the ensemble mean heat map of each cluster and each station, respectively.

4.5.2 Temperature perturbations

We assume that the transition from below- to above-freezing temperatures is strongly related to the onset of snowmelt and thus the timing of the freshet. While the assumption is a simplification of processes dictated by the surface energy balance, the use of positive temperatures as successful indicators for the warming and melting of snow is a common assumption of positive-degree-day models in simulating snow and glacier melt across multiple spatial scales (e.g. Hock, 2003; Radic et al., 2014). Therefore, for interpreting the model's learning, we introduce the following hypothesis: if the model is learning physical processes which are driving streamflow over the course of one year, and since snowmelt is a key contributor to streamflow, then the modelled freshet should occur once temperatures in the forcing data have transitioned from below- to above-freezing. To test the hypothesis, we add a spatially uniform temperature perturbation, ΔT, to both the maximum and minimum temperature channels, i.e. the same temperature change as measured in degrees Celsius is added to every pixel and every day in the test period. With this perturbation we create a new test set which has either warmer or colder temperature channels than the original, but the same precipitation channel. We pass this new test set through the model and compute the mean seasonal flow for each cluster, where the mean is derived across all years in the test set and all stations in the cluster. We perform these steps for the range $-5°\mathrm{C} \leq \Delta T \leq 5°C$ with an increment of $1°\mathrm{C}$ to test how the modelled streamflow responds under a range of warmer or cooler conditions.

Then, for each cluster region and for each temperature perturbation, we identify when the 30-day running mean of daily minimum temperature and maximum temperature transition from being below- to above-freezing temperatures:

$$T_{max}(t_{0,max}) = 0°C \quad (14)$$

$$T_{min}(t_{0,min}) = 0°C \quad (15)$$

where $t_{0,max}$ and $t_{0,min}$ indicate the day when maximum and minimum temperatures warm above freezing, respectively. The timing of a freshet has been previously defined in different ways, each with the goal of indicating when the spring snowmelt is strongly contributing to streamflow (e.g. Vincent et al., 2015; Zhang et al., 2001). For each cluster and temperature perturbation, we define the freshet timing ($t_{freshet}$) as the day when the 30-day running mean of modelled streamflow rises to be halfway between the winter minimum flow (Q_{min}) and spring maximum flow (Q_{max}):

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

$$Q(t_{freshet}) = \frac{Q_{max} - Q_{min}}{2} \quad (16)$$

For each cluster and temperature perturbation, we also define the intensity or magnitude of the freshet to be the spring maximum flow, Q_{max}. By perturbing temperatures in the range of $-5°C \leq \Delta T \leq 5°C$, we can track how well the model is learning the links between the temperature transitions and the intensity and timing of the snowmelt-driven freshets.

5 Results

5.1 Evaluation of NSE

For each station, we derive ensemble-mean streamflow for the bulk model runs and fine-tuned model runs. The median fine-tuned NSE calculated over the test period is 0.68, and 35% of stream gauges have NSE > 0.8 (Fig. 4a). We compare the performance of the bulk versus fine-tuned models by looking at the difference in NSE between the bulk and fine-tuned models (ΔNSE), evaluated across stations for each cluster (Fig. 5a) and in space (Fig. A4a). We find that overall, there is a small increase in NSE, with only a median $\Delta NSE = 0.02$. The best performing stations are those in the central, southern, and north-western clusters, all of which have snowmelt dominated streamflow regimes throughout BC (Fig. 2). For these clusters, which represent a majority of stations, there is relatively little change in NSE between the bulk and fine-tuned models (Fig. 5a). The eastern cluster, which is made up of stations in the relatively arid Prairie region, has the worst overall performance and shows only slight improvements after fine-tuning. The coastal cluster, which is made up of rainfall dominated stations along the west coast, has a relatively narrow range of NSE and shows the largest improvement from fine-tuning. The north-eastern cluster, which is characterized as having comparable snowmelt- and rainfall-driven peaks in spring and summer, respectively, also shows a notable improvement from fine-tuning. Importantly, the median NSE is relatively consistent across model runs in the fine-tuned ensemble, with a range of only 0.05 across all 10 fine-tuned model runs. This result indicates that in terms of NSE, the fine-tuned model runs perform similarly as evaluated across the whole region.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

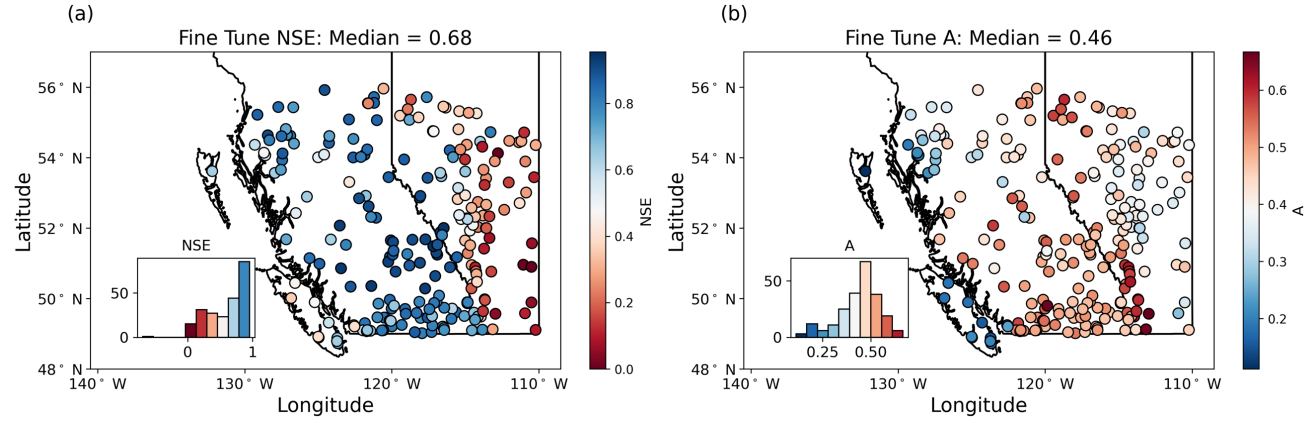
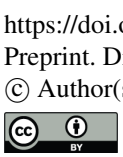


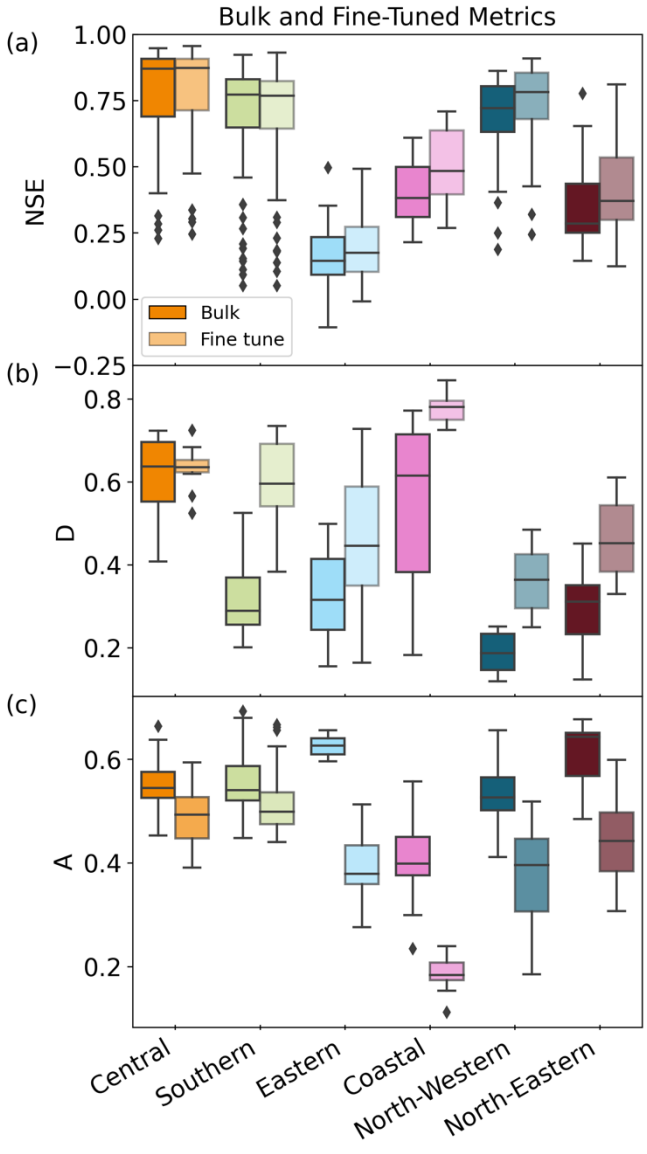
Figure 4: NSE as calculated over the test period for the fine-tuned model. Inset shows a histogram of NSE values of these stations. The colourmap is clipped at NSE = 0 for better visualization and is justified since only two stations have $NSE < 0$.

To illustrate the model's performance in simulating different streamflow regimes, we compare the fine-tuned model output between a station with a snowmelt dominated regime and a station with a rainfall dominated regime (Fig. 6). The snowmelt dominated station is well simulated by the ensemble-mean ($NSE = 0.87$), capturing the timing and magnitude of many daily or weekly scale streamflow peaks; however, the 2σ interval is not consistently narrow throughout the year (Fig. 6a). Rather, it is smallest in the low-flow periods and freshet, and then larger during the recession over spring and summer. The modelled streamflow for the rainfall dominated station yields a lower NSE than the snowmelt dominated station ($NSE = 0.59$), despite the ensemble-mean being able to correctly model the timing of most rainfall-induced onset, peaks, and decay. However, the peak magnitude in streamflow is often under- or over-estimated, particularly for the largest observed peaks (Fig. 6b). The 2σ interval is relatively narrow throughout the year, indicating that the 10 fine-tuned models output relatively similar streamflow.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021



Hydrology and Earth System Sciences
Discussions
Open Access
EGU



475 **Figure 5: Comparison of metrics for both the bulk and fine-tuned models for each cluster of stream gauge stations, coloured according to the clusters shown in Figure 2.** a) NSE of modelled streamflow of each stream gauge station ($n = 226$ across all six clusters). For readability, the y-axis was clipped at $NSE = -0.25$; however, one station in the Eastern cluster is below this threshold for both bulk and fine-tuned models ($NSE = -2.04$ in the bulk model and $NSE = -0.70$ in the fine-tuned model). This station is still included in all analyses and is only not shown here for readability. b) The D-statistic
480 for each model run for both bulk and fine-tuned model types ($n = 226$ for each cluster). c) Sensitive area of the input as evaluated for each stream gauge station ($n = 226$ across all six clusters).

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

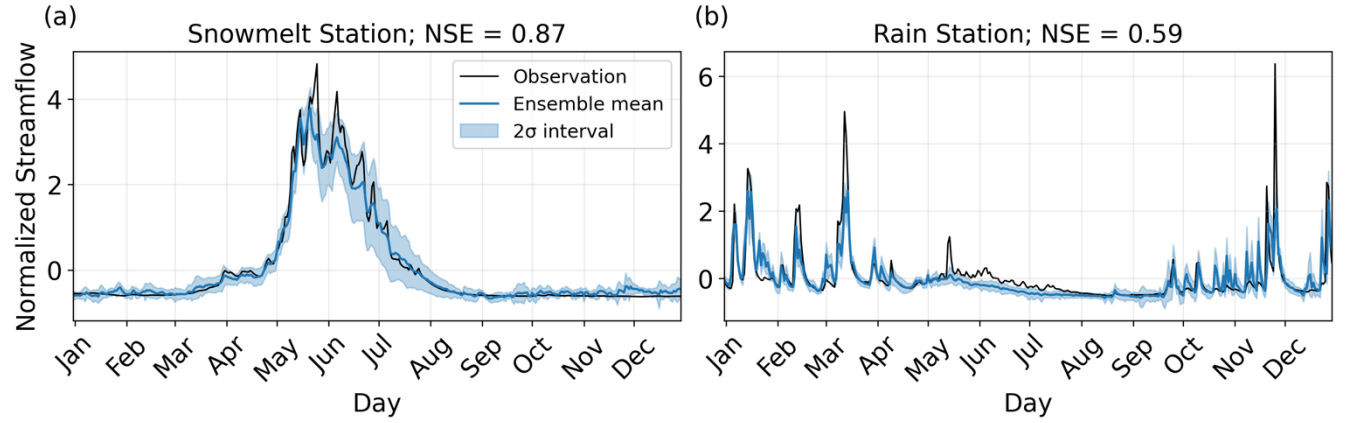


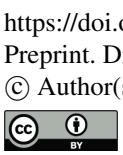
Figure 6: Examples of observed and modelled streamflow for one year at two stations of different streamflow regimes. The ensemble mean is the mean across the 10 model runs, and the shaded area is plus/minus two standard deviations across the 10 model runs. The station with the fifth-highest NSE is chosen in each of the southern and coastal clusters, which are snowmelt- and rain-dominated clusters, respectively. An arbitrary year in the test set is chosen. We chose the fifth-best performing station as to show more typical model performance for these clusters.

5.2 Evaluation of interpretability

5.2.1 Spatial perturbations

Stream gauge stations, watershed boundaries, and sensitivity heat maps for each cluster are displayed in Fig. 7. The central, southern, and coastal regions are generally most sensitive in the areas near and within the watersheds of the cluster, which means that information nearby the watersheds is most important for the model to predict streamflow. The models' predictions are relatively insensitive to perturbations further from the watersheds as indicated by the low values of $S(x, y)$ in Alberta for the coastal cluster and in northern British Columbia for the southern and central clusters. This result indicates that information far from the watersheds is less important for the models' decision making. In contrast, the eastern and north-eastern clusters are sensitive both within the watersheds of the cluster and at a second sensitive area along the west coast of British Columbia. These findings indicate that models for these latter clusters are relying on links across space and time which may be present between the input and output datasets, but which are not physically driving streamflow; consequently, long-term forecasting is likely not appropriate as these links may not hold in the future.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021



Hydrology and Earth System Sciences
Discussions
Open Access
EGU

505

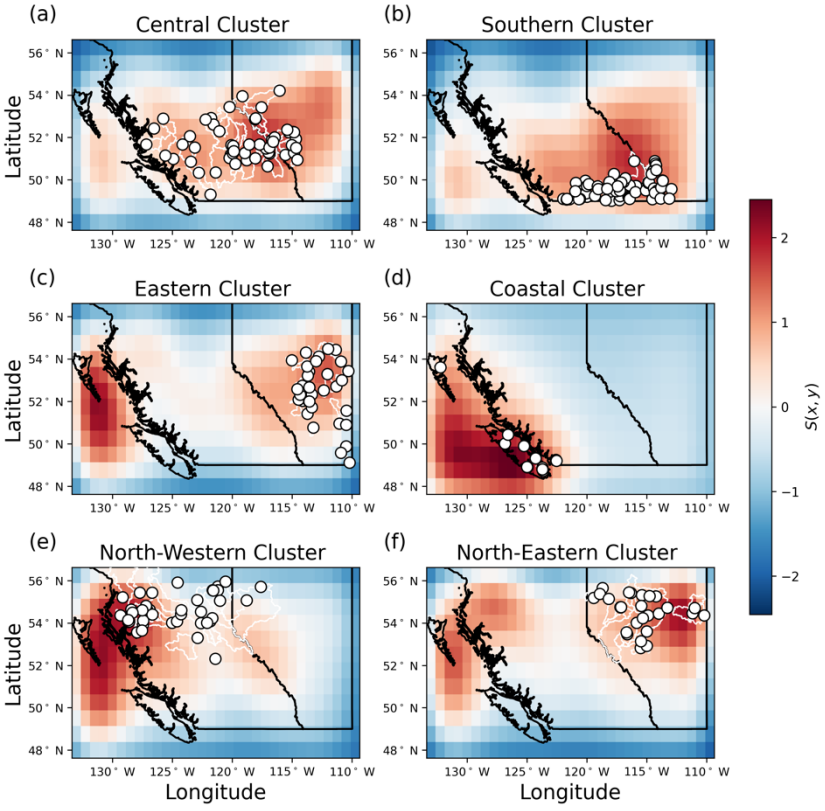


Figure 7: Ensemble mean sensitivity heat maps. For each cluster, the mean sensitivity heat map $S(x, y)$ is calculated across all stations and all days in the test period.

510

The comparison of sensitivity between regions which are within/near the watersheds versus areas which are outside/far from the watersheds are summarized for each cluster (Fig. 8). The steps to calculate the D-statistic for one cluster is shown in Fig. A5. The difference between the within/near and outside/far sensitivity distributions are relatively large for the snowmelt-dominated stations in the central, southern, eastern, and north-eastern clusters (D values of 0.69, 0.66, 0.54, and 0.52, 515 respectively), with the mean sensitivity being higher within/near than outside the watersheds (Fig. 8). The sensitivity distributions are also different for the coastal cluster (D = 0.77), where regions within/near the coastal watersheds are substantially more sensitive to perturbations than the regions outside the watersheds. Stations in the north-western cluster have the lowest D value relative to other clusters (D = 0.40), with the sensitivity near/within the watersheds not being substantially different from the sensitivity outside the watersheds.

520

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

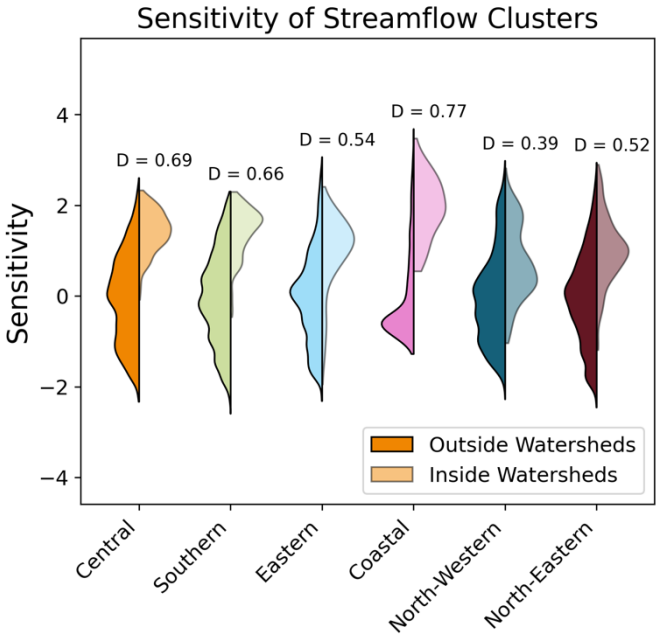


Figure 8: The sensitivity distributions for inside/near and outside of the cluster watersheds. Distribution pairs are labelled by their corresponding D-statistic from the Kolmogorov-Smirnov test. Clusters are coloured according to Fig. 2.

525 The D-statistic is further evaluated by comparing both the bulk and fine-tuned models (Fig. 5b). All clusters except for the north-eastern cluster show an increase in D through fine-tuning, and in particular, southern stations show the largest increase in D. Because D is a metric which indicates how different the inside/outside sensitivity distributions are from one another, the widespread increase in D through fine-tuning indicates that fine-tuning helps the model separate more relevant information (within/near watershed regions having higher sensitivity) from less relevant information (outside/far from 530 watershed regions having lower sensitivity).

We also compare the values of A between the bulk and fine-tuned models per cluster (Fig. 5c) and in space (Fig. A4b). Here, all clusters have their mean A decrease with fine-tuning and the median difference in A between the bulk and fine-tuned models is $\Delta A = -0.09$, meaning that the fine-tuned models are sensitive to smaller areas of the input relative to the bulk model. While the central cluster only shows a minor decrease in A, all other clusters show a larger decrease. Because A is a 535 metric which indicates the area that is most sensitive to perturbation, the widespread decrease in A indicates that the process of fine-tuning generally helps the model to focus on smaller areas of the input space.

It has previously been shown that fine-tuning an LSTM-based hydrological model can lead to a moderate improvement in performance which is heterogeneous in space (Kratzert et al., 2018). Here we build on this understanding to show also that fine-tuning substantially influences what the model is learning. Comparing the results between clusters in terms of NSE, D,

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

and A, we find that the process of fine-tuning does not impact model performance equally across all clusters (Fig. 5). Specifically, the central cluster is the least impacted by fine-tuning, as indicated by the relatively small differences in NSE, D, and A between the bulk and fine-tuned models. The bulk model focuses on large areas of the input, and since the central cluster spans a relatively large area of the input space (the largest area of all the clusters), the bulk model is already effective in learning the weather-to-streamflow mapping and further fine-tuning does not substantially change its learning. On the other hand, the coastal cluster is more impacted by fine-tuning, with simulated streamflow more closely matching observations (increase in NSE) with more realistic learning (increase in D and a substantial decrease in A). In other words, fine-tuning has made the model focus on smaller regions nearby the watersheds, which has led to better performance. Considering the processes which drive streamflow, fine-tuning has minimal impact on NSE at southern and central clusters (which are melt dominated flow), and most improves NSE at the coastal, north-western, north-eastern, and eastern clusters (where a rain-driven flow peak is present in the seasonal hydrograph) (Fig. 2 and Fig. 5a). Seeing as fine-tuning also more substantially decreases A at these latter clusters, and that rainfall occurs over smaller spatial scales as compared to snowmelt, we suggest that the process of fine-tuning allows the model to better learn rainfall-runoff processes.

Interestingly, it is not necessarily true that the model run which performs best according to NSE is the model which has best learned to focus within the watersheds being predicted. All models in the fine-tuned ensemble achieve relatively similar performance as evaluated by the median NSE (range of 0.05 across the 10 models) while there is a much larger range of median D (range of 0.25 across the 10 models). Furthermore, NSE and D are not significantly correlated (correlation coefficient $R = 0.05$ with p-value > 0.1). In fact, the model with the highest median NSE has the lowest median D ($NSE = 0.69$ and $D = 0.41$). The range in D across the models in the ensemble and the lack of significant correlation between NSE and D indicate that while all model runs can output streamflow to a similar degree of accuracy, the internal learning processes are different among model runs.

5.2.2 Temperature perturbations

The intensity and timing of the modelled freshet, as well as the timing of transition from below- to above-freezing temperature, reveals characteristic patterns of snowmelt for the snowmelt dominated clusters (Fig. 9). When no temperature perturbation is added (e.g. $\Delta T = 0$), the snowmelt driven streamflow in southern, central, north-western, north-eastern, and eastern clusters experience a large increase in modelled streamflow after temperatures increase above freezing. For these snowmelt driven streamflow regimes, a positive temperature perturbation ($\Delta T > 0°C$) advances the timing of the freshet's onset while the intensity of the freshet decreases (Fig. 10). A possible physical interpretation of this result is that a warmer climate would lead to both a smaller fraction of precipitation falling as snow rather than as rain and a shorter cold season, leading to a thinner seasonal snowpack, an earlier onset to melt, and less water to drive streamflow in spring. Similarly, a decrease in temperature ($\Delta T < 0°C$) delays the timing of the freshet's onset while the intensity of the freshet increases (Fig. 10). Notably, fall and winter streamflow is suppressed when temperatures are lowered, and enhanced when temperatures are

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

raised in the more rainfall driven coastal and north-western clusters (approximately before April and after December; Fig. 9d and 9e). These results are consistent with the rationale that a colder climate would lead to both a larger fraction of precipitation falling as snow rather than rain and a longer cold season, building a deeper snowpack which can deliver a larger volume of water to streamflow in spring. Similarly, when temperatures are raised ($\Delta T > 0°C$), winter and fall streamflow increases, which can physically be explained by more precipitation falling as rain which leads to a faster streamflow response. Importantly, while the timing of the freshet onset relative to the timing of above-freezing maximum/minimum temperatures may not be the same for all clusters (Fig. 9 and Fig. 10), all clusters respond similarly to a change in the timing of this temperature transition, and this response is consistent with a physical understanding of the drivers of streamflow.

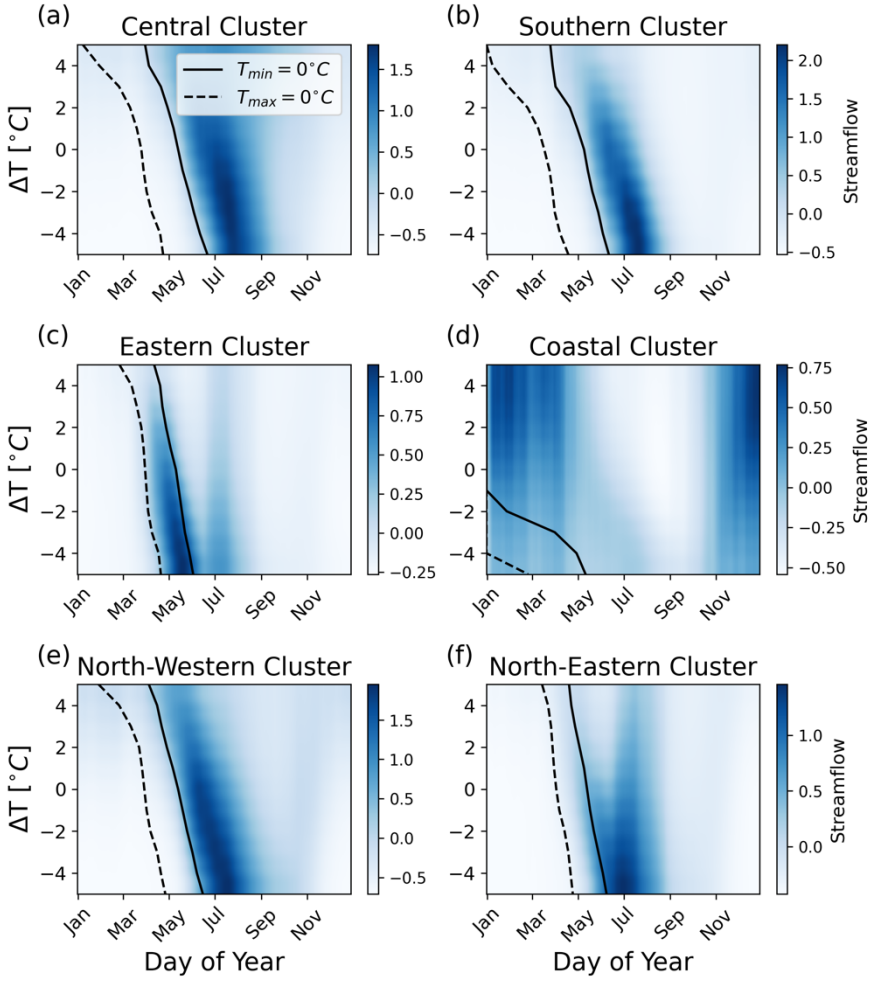


Figure 9: Modelled streamflow for a range of temperature perturbations, averaged across all stream gauges in each cluster and all years in the test set. Black lines indicate the transition from below-freezing to above-freezing maximum (dashed) and minimum (solid) daily temperatures.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

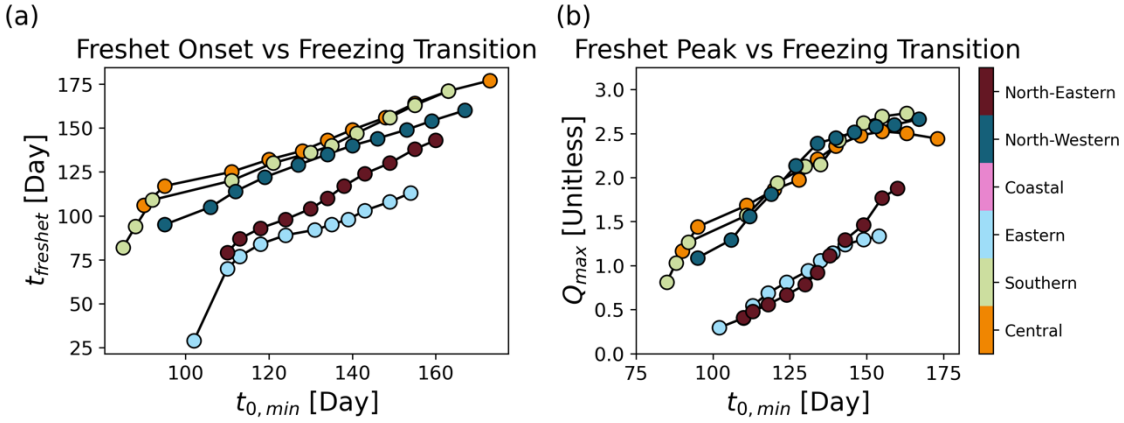


Figure 10: Modelled a) freshet onset timing and b) peak magnitude are both positively correlated with the day that minimum temperatures rise above freezing. The coastal cluster is not shown as it is dominated by winter rainfall rather than a spring freshet.

6 Discussion

Our model achieves comparable performance to previous studies which have used deep learning for modelling streamflow across a region using meteorological inputs; for example, LSTMs have been used to achieve median NSE of 0.72 across 241 catchments in the United States (Kratzert et al., 2018) with the worst performance in the more arid regions (similar to our model's poor performance in the eastern cluster). Additionally, Kratzert et al. (2019) achieved a median NSE of 0.63 across 531 catchments in the United States and found that model performance was improved (median NSE of 0.74 across 531 catchments) when catchment characteristics were included in the input to incorporate information related to the climate, topography, vegetation, and subsurface. The CNN-LSTM modelling framework introduced here could be extended to include spatially distributed variables which are constant in time (such as topography and permeability) by using these variables as additional channels in each frame of the input video, for example.

One of our key findings is that the model performs well (high NSE values) in all clusters except for the eastern cluster. We compare our results with findings from process-based models previously used to predict streamflow at 45 of the same stations as in our study (Table 2). We identify studies which modelled streamflow at daily temporal resolution, and evaluated performance using NSE or correlation between observed and modelled streamflow over at least a one-year period. In total we selected 45 stations for this comparison, keeping in mind that this is not an exhaustive comparison of all studies across the region, nor do we claim that the identified studies are necessarily directly comparable with our results as each define calibration and validation periods differently. As our goals are to explore the CNN-LSTM model architecture and interpret its decision making, and not necessarily to outperform existing process-based models, we do not need to compare every individual station

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

to process-based models. The intercomparison shows that the CNN-LSTM model performance is at least similar to, and often outperforms, many process-based models existing in the literature for the southern, central, coastal, north-western, and north-eastern clusters (Table 2). Our model achieves higher values of NSE or correlation at 37 of the 45 identified stream gauge stations, indicating generally good performance in all clusters except for the eastern cluster. It is notable that the CNN-LSTM model achieves this performance with only coarse resolution temperature and precipitation as input.

The eastern cluster is unique among our six clusters in terms our model's poor performance, and also in terms of the region's hydrology and a lack of studies in the literature at the same locations and with the same metrics as our study for comparison. Our model's inability to successfully simulate streamflow for the eastern cluster (Fig. 5a) could be due to the unique hydrological conditions in the Prairie region which cannot be learned from the past 365 days of temperature and precipitation alone. Prairie topography is characterized by small surface depressions which result in intermittent water connectivity and variable sized drainage basins (Shaw et al., 2012). Rainfall and snowmelt are stored in upstream depressions rather than contributing to streamflow when the depressions are not full, and so there may not be a substantial streamflow response even if there is rainfall or snowmelt within the basin. Additionally, there is hysteresis between contributing area and water storage within these depressions, meaning that the area which contributes to streamflow is determined by both the presence of depressional storage and if the storage is increasing (wetting) or decreasing (drying) (Shook and Pomeroy, 2011). Because storage in ponds can vary substantially on both seasonal and decadal timescales (Hayashi et al., 2016; Shaw et al., 2012), it is likely that the CNN-LSTM model has insufficient input information to successfully learn the complex rainfall-runoff behaviour observed in the eastern basins (e.g. both the seasonal and decadal fluctuations of depressional storage cannot be learned from only a seasonal-scale input time series). Several studies have developed process-based models for application to different stations than ours in the Prairie region, which have outperformed our DL-based approach. While our model had a median $NSE = 0.17$, process-based models achieve higher values, for example, $NSE > 0.4$ (Mengistu and Spence, 2016; Unduche et al., 2018) and $NSE > 0.7$ (Muhammad et al., 2019).

Considering the complexity of hydrology in real catchments and the dependence of streamflow on locally resolved processes, it is notable that the CNN-LSTM model achieves good streamflow simulation with only temperature and precipitation forcing data. The model's performance is likely limited by its inability to learn processes beyond those which could be more easily inferred from the streamflow response to temperature and precipitation, such as advective fluxes (e.g. wind transport of snow), evaporative fluxes (e.g. sublimation of the snowpack, evapotranspiration), and interactions between ground and surface water. Additionally, our model uses forcing data at comparably 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)). While it is notable that the CNN-LSTM model achieves good performance without climate downscaling, it is also possible that the absence of information of locally resolved meteorological data limits the model's learning and performance. Nevertheless, the model could still be successful in regions where the importance of such processes is less than those which can be inferred from coarse resolution temperature and precipitation alone. For example, winters are generally warmer and wetter in British Columbia (where the model performs better) as compared to Alberta (where the model

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

performs worse), which may limit the importance of processes such as blowing snow transport and sublimation by increasing cohesion of the snowpack (Pomeroy and Gray, 1990). Additionally, characteristics such as topography and subsurface geology play an important role in determining how surface and ground water interact. However, since geology and topography remain essentially constant in time through the study period, the controls they exert on streamflow should remain consistent through the study period as well. As such, their role in determining streamflow at each station may be encoded as part of the mapping between the meteorological forcing and streamflow response at each station so long as it is consistent through the study period.

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, and so the number of observations used for training is decreased by a factor N as compared to using an LSTM model at each station individually, where N is the number of stream gauge stations (e.g. for each year of daily streamflow at N stations, the CNN-LSTM model is trained with 365 observations, while an LSTM model would be trained with $N \times 365$ observations). For a given architecture, having a smaller training set means that there are fewer computational requirements to train a single epoch; however, this is only beneficial so long as the training dataset is still large enough to achieve good model performance. There may be regions of interest where the streamflow record is not as extensive as in our study region and thus training only within the region of interest may not produce sufficiently accurate streamflow predictions. 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.

7 Summary and conclusions

This study investigated the applicability of a sequential CNN-LSTM model for regional scale hydrological modelling, where the model was forced by gridded climate data to predict streamflow at multiple stream gauge stations simultaneously. We focused on using a relatively simple deep learning model, with the input data represented by temperature and precipitation reanalysis for the period 1979 – 2015 given on relatively coarse spatial resolution (0.75° x 0.75°). In particular, we investigated how well the model learns different streamflow regimes and how physically realistic the model's learning was for each streamflow regime. To reach these goals, the model was trained, validated, and tested on a set of stream gauge stations across Western Canada, initially partitioned into six clusters based on the similarity in both seasonal streamflow regime and proximity in space. A set of metrics was introduced and developed to evaluate the model performance and to investigate the model's learning, in particular the learning of spatial and temporal processes taking place at the weather-to-streamflow model's mapping. We summarize the major findings as follows:

1) The model successfully simulated streamflow at multiple stations simultaneously, with a median NSE of 0.68 and 35% of stations having NSE > 0.8. The best model performance was for stations with snowmelt dominated streamflow in British Columbia, and the worst performance was for the eastern cluster of stations in the Prairie region.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

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

2) For a majority of stations, the model was most sensitive to perturbations in the input data prescribed near and within the basins being predicted, demonstrating that the model's spatial learning focused on areas where the physical drivers of streamflow are occurring. For the eastern and north-eastern clusters, the model was sensitive to perturbations that are relatively far from the watersheds where streamflow was being predicted, thus linking the streamflow to weather fields far away (> 500 km apart). In these cases, the model may be more appropriate for short term rather than long term prediction, as the learned links over far distances may not hold in the future.
3) To investigate the learning of temporal patterns, we focused on the timing and intensity of the spring freshet. By uniformly perturbing temperature input to drive the model with warmer and colder climates relative to present, the model responded by changing the intensity and timing of the freshet in accordance with the timing of the transition from below- to above-freezing temperatures.
4) Fine-tuning by streamflow regime led to modest improvements in model performance as evaluated by NSE, but allowed the model to focus on areas near and within the watersheds where streamflow is being predicted. We conclude that fine-tuning is most beneficial for directing the model to focus on processes taking place on relatively smaller spatiotemporal scales (e.g. rainfall driven as opposed to snowmelt driven streamflow).

The CNN-LSTM model presented has been able to explicitly incorporate both spatial and temporal information for predicting streamflow across a region. In addition to successfully simulating streamflow across a range of streamflow regimes, we are able to interpret key aspects of the model's learning. Interpretability of model learning builds trust in the model's predictions, leading to potentially further applications whether for prediction, or as a compliment to process based models, or in further exploration of processes which are poorly understood or represented in process based models.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

Table 2: Comparison to select process-based models. Metrics for comparison are Nash-Sutcliffe Efficiency (NSE) or correlation coefficient (R). For our CNN-LSTM model, the performance metrics are calculated from the test set. For the reference models, performance metrics are as reported for the validation set. In bold is the better performing (higher) value between the two models.

Station Name	Station ID	Cluster	Reference NSE	CNN-LSTM NSE	Reference R	CNN-LSTM R	Reference
Bridge River below Bridge Glacier	08ME023	Central	**0.95**	0.94			(Stahl et al., 2008)
Lillooet River near Pemberton	08MG005	Central	0.70	**0.88**			(Whitfield et al., 2002)
Quesnel River near Quesnel	08KH006	Central	0.83	**0.90**			(Shrestha et al., 2012)
North Thompson River at McLure	08LB064	Central	0.85	**0.92**			(Shrestha et al., 2012)
South Thompson River at Chase	08LE031	Central	0.87	**0.91**			(Shrestha et al., 2012)
Thompson River near Spences Bridge	08LF051	Central	0.89	**0.93**			(Shrestha et al., 2012)
Harrison River near Harrison Hot Springs	08MG013	Central	0.66	**0.83**			(Shrestha et al., 2012)
Columbia River at Nicholson	08NA002	Central			0.899	**0.980**	(Bingeman et al., 2006)
Kicking Horse River at Golden	08NA006	Central	0.77	**0.91**	0.884	**0.961**	(Bingeman et al., 2006; Schnorbus et al., 2011)
Columbia River at Donald	08NB005	Central	0.91	**0.96**	0.924	**0.984**	(Bingeman et al., 2006; Schnorbus et al., 2011)
Goldstream River below Old Camp Creek	08ND012	Central			0.689	**0.961**	(Bingeman et al., 2006)
Duncan River below B.B. Creek	08NH119	Central			0.863	**0.959**	(Bingeman et al., 2006)

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

Illecillewaet River at Greeley	08ND013	Central			0.906	**0.959**	(Bingeman et al., 2006)
Gold River above Palmer Creek	08NB014	Central			0.813	**0.957**	(Bingeman et al., 2006)
Split Creek at the Mouth	08NB016	Central			0.744	**0.954**	(Bingeman et al., 2006)
Miette River near Jasper	07AA001	Central	**0.86**	0.84			(Chernos et al., 2020)
Athabasca River near Jasper	07AA002	Central	**0.93**	0.90			(Chernos et al., 2020)
Athabasca River at Hinton	07AD002	Central	**0.91**	0.88			(Chernos et al., 2020)
Athabasca River near Windfall	07AE001	Central	0.80	**0.86**			(Eum et al., 2017)
Englishman River near Parksville	08HB002	Coastal	**0.65**	0.59			(Whitfield et al., 2002)
Athabasca River at Athabasca	07BE001	North-Eastern	0.75	**0.81**			(Eum et al., 2017)
Pembina River at Jarvie	07BC002	North-Eastern	0.48	**0.63**			(Eum et al., 2017)
Stuart River near Fort St. James	08JE001	North-Western	0.82	**0.86**			(Shrestha et al., 2012)
Fraser River at Shelley	08KB001	North-Western	0.75	**0.87**			(Shrestha et al., 2012)
Omineca River near the Mouth	07EC002	North-Western	0.81	**0.89**			(Schnorbus et al., 2011)
Parsnip River above Misinchinka River	07EE007	North-Western	0.81	**0.87**			(Schnorbus et al., 2011)
Pine River at East Pine	07FB001	North-Western	0.71	**0.84**			(Schnorbus et al., 2011)
Murray River above Wolverine River	07FB006	North-Western	0.67	**0.86**			(Schnorbus et al., 2011)
Murray River near the Mouth	07FB002	North-Western	0.58	**0.86**			(Schnorbus et al., 2011)
Sukunka River near the Mouth	07FB003	North-Western	0.78	**0.80**			(Schnorbus et al., 2011)
Kuskanax Creek near Nakusp	08NE006	Southern			0.819	**0.968**	(Bingeman et al., 2006)

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

Kaslo River below Kemp Creek	08NH005	Southern			0.864	**0.938**	(Bingeman et al., 2006)
Barnes Creek near Needles	08NE077	Southern			0.797	**0.911**	(Bingeman et al., 2006)
Mather Creek below Houle Creek	08NG076	Southern			0.795	**0.868**	(Bingeman et al., 2006)
St. Mary River below Morris Creek	08NG077	Southern			0.871	**0.926**	(Bingeman et al., 2006)
Fry Creek below Carney Creek	08NG130	Southern			0.816	**0.936**	(Bingeman et al., 2006)
Keen Creek below Kyawats Creek	08NH132	Southern			0.774	**0.931**	(Bingeman et al., 2006)
Lemon Creek above South Lemon Creek	08NJ160	Southern			0.784	**0.945**	(Bingeman et al., 2006)
Kootenay River at Kootenay Crossing	08NF001	Southern	0.75	**0.89**			(Schnorbus et al., 2011)
Kootenay River at Fort Steele	08NG065	Southern	0.85	**0.86**			(Schnorbus et al., 2011)
Elk River at Fernie	08NK002	Southern	0.81 **0.92**	0.73			(Schnorbus et al., 2011) (Chernos et al., 2017)
Elk River near Natal	08NK016	Southern	0.75 **0.91**	0.73			(Schnorbus et al., 2011) (Chernos et al., 2017)
Fording River at the Mouth	08NK018	Southern	0.72 **0.84**	0.71			(Schnorbus et al., 2011) (Chernos et al., 2017)
Slocan River near Crescent Valley	08NJ013	Southern	0.78	**0.88**			(Schnorbus et al., 2011)
Salmo River near Salmo	08NE074	Southern	0.73	**0.83**			(Schnorbus et al., 2011)

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021


Hydrology and Earth System Sciences
Discussions

Data availability

All data used in this study are publicly available. ERA reanalysis data are available from the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach et al., 2020). Streamflow data are available from the Environment Canada HYDAT database (Environment and Climate Change Canada, 2018). Basin outlines are available from the Water Survey of Canada (Environment and Climate Change Canada, 2016). Provincial borders used in mapping are available from Statistics Canada (Statistics Canada, 2016).

Code availability

Code used in this study is available on Github (https://github.com/andersonsam/cnn_lstm_era).

Author contributions

SA designed and conducted the study. SA and VR analysed the results and wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This research was funded through the National Science and Engineering Research Council of Canada (NSERC).

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, CoRR, abs/1603.0 [online] Available from: http://arxiv.org/abs/1603.04467, 2016.

Arras, L., Arjona-Medina, J., Widrich, M., Montavon, G., Gillhofer, M., Müller, K.-R., Hochreiter, S. and Samek, W.: Explaining and Interpreting LSTMs BT - Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, edited by W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, pp. 211–238, Springer International Publishing, Cham., 2019.

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.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W.: On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation, PLoS One, 10(7), e0130140 [online] Available from: https://doi.org/10.1371/journal.pone.0130140, 2015.

Bahremand, A.: HESS Opinions: Advocating process modeling and de-emphasizing parameter estimation, Hydrol. Earth Syst. Sci., 20(4), 1433–1445, doi:10.5194/hess-20-1433-2016, 2016.

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.

Bingeman, A. K., Kouwen, N. and Soulis, E. D.: Validation of the Hydrological Processes in a Hydrological Model, J. Hydrol. Eng., 11(5), 451–463, doi:10.1061/(ASCE)1084-0699(2006)11:5(451), 2006.

Chakravarti, I. M., Laha, G. G. and Roy, J.: Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, Hoboken., 1967.

Chernos, M., MacDonald, R. and Craig, J.: Efficient semi-distributed hydrological modelling workflow for simulating streamflow and characterizing hydrologic processes, Conflu. J. Watershed Sci. Manag., 1(3), doi:10.22230/jwsm.2018v1n3a6, 2017.

Chernos, M., MacDonald, R. J., Nemeth, M. W. and Craig, J. R.: Current and future projections of glacier contribution to streamflow in the upper Athabasca River Basin, Can. Water Resour. J. / Rev. Can. des ressources hydriques, 45(4), 324–344, doi:10.1080/07011784.2020.1815587, 2020.

Chollet, F.: Keras, GitHub Repos., 2015.

Donahue, J., Hendricks, L. A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K. and Darrell, T.: Long-Term Recurrent Convolutional Networks for Visual Recognition and Description, IEEE Trans. Pattern Anal. Mach. Intell., 39(4), 677–691, doi:10.1109/TPAMI.2016.2599174, 2017.

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.

Essou, G. R. C., Sabarly, F., Lucas-Picher, P., Brissette, F. and Poulin, A.: Can Precipitation and Temperature from Meteorological Reanalyses Be Used for Hydrological Modeling?, J. Hydrometeorol., 17(7), 1929–1950, doi:10.1175/JHM-D-15-0138.1, 2016.

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.

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.

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.

Hastie, T., Tibshirani, R. and Friedman, J. H.: The elements of statistical learning: data mining, inference, and prediction, 2nd ed., Springer, New York. [online] Available from: https://adams.marmot.org/Record/.b41452057#:~:text=Hastie%2C T.%2C Tibshirani%2C,New York%3A Springer., 2009.

Hayashi, M., van der Kamp, G. and Rosenberry, D. O.: Hydrology of Prairie Wetlands: Understanding the Integrated Surface-Water and Groundwater Processes, Wetlands, 36(2), 237–254, doi:10.1007/s13157-016-0797-9, 2016.

Hedstrom, N. R. and Pomeroy, J. W.: Measurements and modelling of snow interception in the boreal forest, Hydrol. Process., 12(10-11), 1611–1625, doi:https://doi.org/10.1002/(SICI)1099-1085(199808/09)12:10/11<1611::AID-HYP684>3.0.CO;2-4, 1998.

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.

Hock, R.: Temperature index melt modelling in mountain areas, J. Hydrol., 282(1), 104–115, doi:https://doi.org/10.1016/S0022-1694(03)00257-9, 2003.

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.

Karpathy, A., Johnson, J. and Li, F.-F.: Visualizing and Understanding Recurrent Networks, CoRR, abs/1506.0 [online] Available from: http://arxiv.org/abs/1506.02078, 2015.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A. and Kumar, V.: Machine Learning for the Geosciences: Challenges and Opportunities, IEEE Trans. Knowl. Data Eng., 31(8), 1544–1554, doi:10.1109/TKDE.2018.2861006, 2019.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, 2017.

Kiros, R., Salakhutdinov, R. and Zemel, R. S.: Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, CoRR, abs/1411.2 [online] Available from: http://arxiv.org/abs/1411.2539, 2014.

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.

Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in Advances in Neural Information Processing Systems 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, pp. 1097–1105, Curran Associates, Inc. [online] Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf, 2012.

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.

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.

Mengistu, S. G. and Spence, C.: Testing the ability of a semidistributed hydrological model to simulate contributing area, Water Resour. Res., 52(6), 4399–4415, doi:https://doi.org/10.1002/2016WR018760, 2016.

Met Office: Cartopy: a cartographic python library with a matplotlib interface, 2018.

Milly, P. C. D., Betancourt, J., Falkenmark, M., Hirsch, R. M., Kundzewicz, Z. W., Lettenmaier, D. P. and Stouffer, R. J.: Stationarity Is Dead: Whither Water Management?, Science (80-.)., 319(5863), 573–574 [online] Available from: http://www.jstor.org.ezproxy.library.ubc.ca/stable/20053240, 2008.

Moore, R. D., Spittlehouse, D. L., Whitfield, P. H. and Stahl, K.: Weather and Climate, 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. 47–84, Victoria, British Columbia., 2010.

Muhammad, A., Evenson, G. R., Stadnyk, T. A., Boluwade, A., Jha, S. K. and Coulibaly, P.: Impact of model structure on the accuracy of hydrological modeling of a Canadian Prairie watershed, J. Hydrol. Reg. Stud., 21, 40–56, doi:https://doi.org/10.1016/j.ejrh.2018.11.005, 2019.

Nash, J. E. and Sutcliffe, J. V: River flow forecasting through conceptual models part I — A discussion of principles, J. Hydrol., 10(3), 282–290, doi:https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Painter, S. L., Coon, E. T., Atchley, A. L., Berndt, M., Garimella, R., Moulton, J. D., Svyatskiy, D. and Wilson, C. J.: Integrated

surface/subsurface permafrost thermal hydrology: Model formulation and proof-of-concept simulations, Water Resour. Res., 52(8), 6062–6077, doi:https://doi.org/10.1002/2015WR018427, 2016.

Penman, H. L. and Keen, B. A.: Natural evaporation from open water, bare soil and grass, Proc. R. Soc. London. Ser. A. Math. Phys. Sci., 193(1032), 120–145, doi:10.1098/rspa.1948.0037, 1948.

Petsiuk, V., Das, A. and Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models, CoRR, abs/1806.0 [online] Available from: http://arxiv.org/abs/1806.07421, 2018.

Pomeroy, J. W. and Gray, D. M.: Saltation of snow, Water Resour. Res., 26(7), 1583–1594, doi:https://doi.org/10.1029/WR026i007p01583, 1990.

Pomeroy, J. W. and Li, L.: Prairie and arctic areal snow cover mass balance using a blowing snow model, J. Geophys. Res. Atmos., 105(D21), 26619–26634, doi:https://doi.org/10.1029/2000JD900149, 2000.

Radic, V., Bliss, A., Beedlow, A. C., Hock, R., Miles, E. and Cogley, J. G.: Regional and global projections of twenty-first century glacier mass changes in response to climate scenarios from global climate models, Clim. Dyn., 42(1–2), 37–58, doi:http://dx.doi.org/10.1007/s00382-013-1719-7, 2014.

Razavian, A. S., Azizpour, H., Sullivan, J. and Carlsson, S.: CNN Features off-the-shelf: an Astounding Baseline for Recognition, CoRR, abs/1403.6 [online] Available from: http://arxiv.org/abs/1403.6382, 2014.

Van Rossum, G. and Drake, F. L.: Python 3 Reference Manual, CreateSpace, Scotts Valley, CA., 2009.

Schnorbus, M. A., Bennett, K. E., Werner, A. T. and Berland, A. J.: Hydrologic Impacts of Climate Change in the Peace, Campbell and Columbia Watersheds, British Columbia, Canada, Victoria, British Columbia., 2011.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, Int. J. Comput. Vis., 128(2), 336–359, doi:10.1007/s11263-019-01228-7, 2020.

Shaw, D. A., Vanderkamp, G., Conly, F. M., Pietroniro, A. and Martz, L.: The Fill–Spill Hydrology of Prairie Wetland Complexes during Drought and Deluge, Hydrol. Process., 26(20), 3147–3156, doi:https://doi.org/10.1002/hyp.8390, 2012.

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.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W. and WOO, W.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, in Advances in Neural Information Processing Systems 28, edited by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, pp. 802–810, Curran Associates, Inc. [online] Available from: http://papers.nips.cc/paper/5955-convolutional-lstm-network-a-machine-learning-approach-for-precipitation-nowcasting.pdf, 2015.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

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.

Sinclair, K. E. and Marshall, S. J.: Temperature and vapour-trajectory controls on the stable-isotope signal in Canadian Rocky Mountain snowpacks, J. Glaciol., 55(191), 485–498, doi:DOI: 10.3189/002214309788816687, 2009.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A Simple Way to Prevent Neural Networks from Overfitting, J. Mach. Learn. Res., 15(1), 1929–1958, 2014.

Stahl, K., Moore, R. D., Shea, J. M., Hutchinson, D. and Cannon, A. J.: Coupled modelling of glacier and streamflow response to future climate scenarios, Water Resour. Res., 44(2), doi:10.1029/2007WR005956, 2008.

Statistics Canada: Boundary Files, 2016 Census., [online] Available from: https://open.canada.ca/data/en/dataset/a883eb14-0c0e-45c4-b8c4-b54c4a819edb, 2016.

Sutskever, I., Vinyals, O. and Le, Q. V: Sequence to Sequence Learning with Neural Networks, CoRR, abs/1409.3 [online] Available from: http://arxiv.org/abs/1409.3215, 2014.

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.

Trubilowicz, J. W., Shea, J. M., Jost, G. and Moore, R. D.: Suitability of North American Regional Reanalysis (NARR) output for hydrologic modelling and analysis in mountainous terrain, Hydrol. Process., 30(13), 2332–2347, doi:https://doi.org/10.1002/hyp.10795, 2016.

Unduche, F., Tolossa, H., Senbeta, D. and Zhu, E.: Evaluation of four hydrological models for operational flood forecasting in a Canadian Prairie watershed, Hydrol. Sci. J., 63(8), 1133–1149, doi:10.1080/02626667.2018.1474219, 2018.

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.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R. J., Darrell, T. and Saenko, K.: Sequence to Sequence - Video to Text, CoRR, abs/1505.0 [online] Available from: http://arxiv.org/abs/1505.00487, 2015.

Vickers, G., Buzza, S., Schmidt, D. and Mullock, J.: The Weather of the Canadian Prairies, NAV CANADA. [online] Available from: http://www.navcanada.ca/EN/media/Publications/Local Area Weather Manuals/LAWM-Prairies-1-EN.pdf, 2001.

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.

Wheater, H. and Gober, P.: Water security in the Canadian Prairies: science and management challenges, Philos. Trans. Math. Phys. Eng. Sci., 371(2002), 1–21 [online] Available from: http://www.jstor.org/stable/42583068, 2013.

Whitfield, P. H., Cannon, A. J. and Reynolds, C. J.: Modelling Streamflow in Present and Future Climates: Examples from the Georgia Basin, British Columbia, Can. Water Resour. J. / Rev. Can. des ressources hydriques, 27(4), 427–456, doi:10.4296/cwrj2704427, 2002.

Xu, X., Frey, S. K., Boluwade, A., Erler, A. R., Khader, O., Lapen, D. R. and Sudicky, E.: Evaluation of variability among different precipitation products in the Northern Great Plains, J. Hydrol. Reg. Stud., 24, 100608, doi:https://doi.org/10.1016/j.ejrh.2019.100608, 2019.

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.

Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, in Computer Vision -- ECCV 2014, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, pp. 818–833, Springer International Publishing, Cham., 2014.

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.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

Appendix A

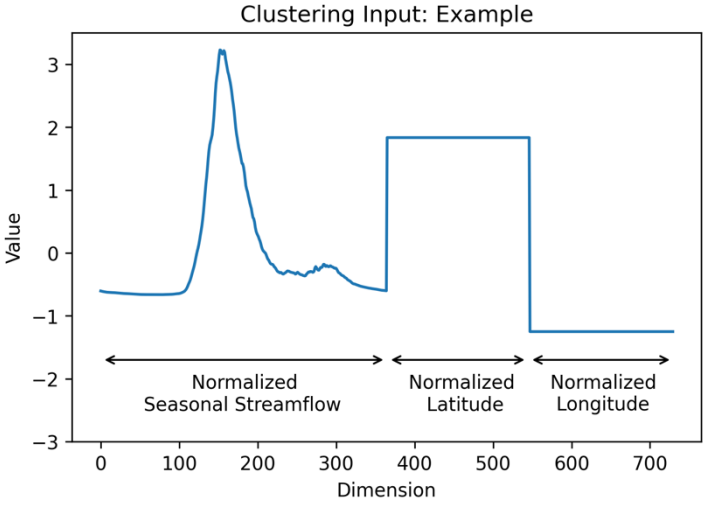


Figure A1: Example of one observation input into clustering algorithm. The first half of the input vector is the seasonal streamflow (normalized at each station), the third quarter is latitude (normalized across all stations), and the fourth quarter is longitude (normalized across all stations).

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

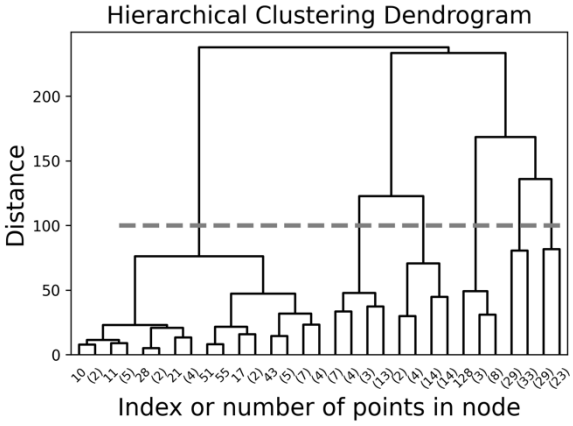


Figure A2: Dendrogram of stream gauge clustering. The dashed grey line indicates the level at which we grouped the cluster members.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

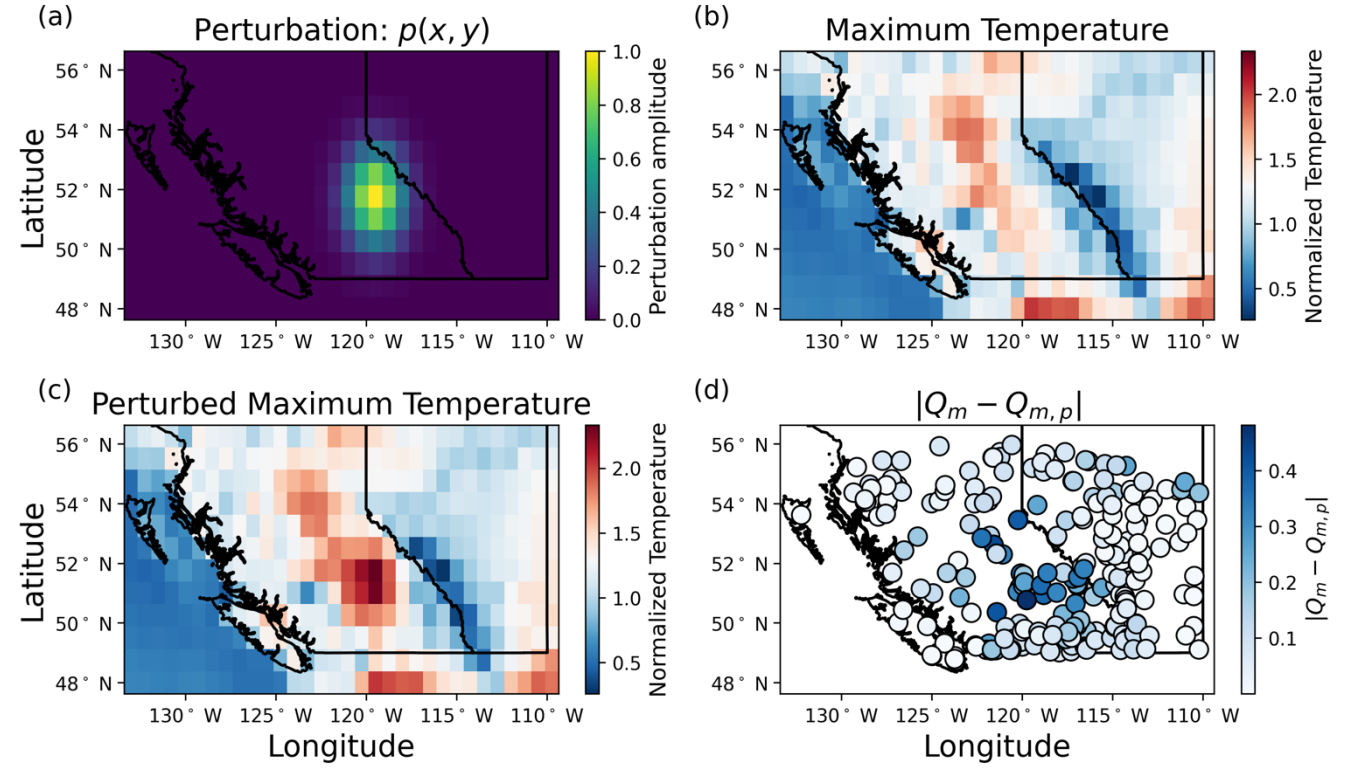
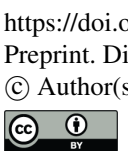


Figure A3: Perturbing one day's maximum temperature field. a) The perturbation to be added to the test weather data. b) The unperturbed maximum temperature field for 3 August 2011. c) The perturbed maximum temperature field for 3 August 2011. d) The magnitude of the difference between the unperturbed streamflow and the perturbed streamflow, averaged across all model runs.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021



Hydrology and Earth System Sciences
Discussions

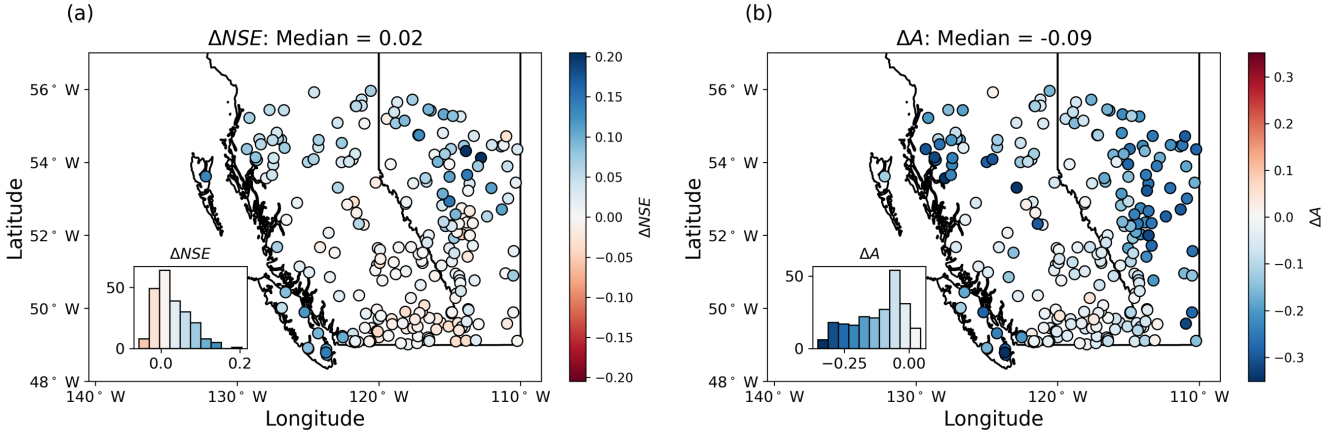


Figure A4: The difference in model performance between the bulk model and fine-tuned model for both a) NSE, and b) sensitive area A. Negative values indicate that fine-tuning reduced NSE or A, while positive values indicate that fine-tuning increase NSE or A. One stream gauge station experienced a substantial increase in NSE from fine tuning ($\Delta NSE >$ 1.0), but the histogram and colour bar are clipped for readability.

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

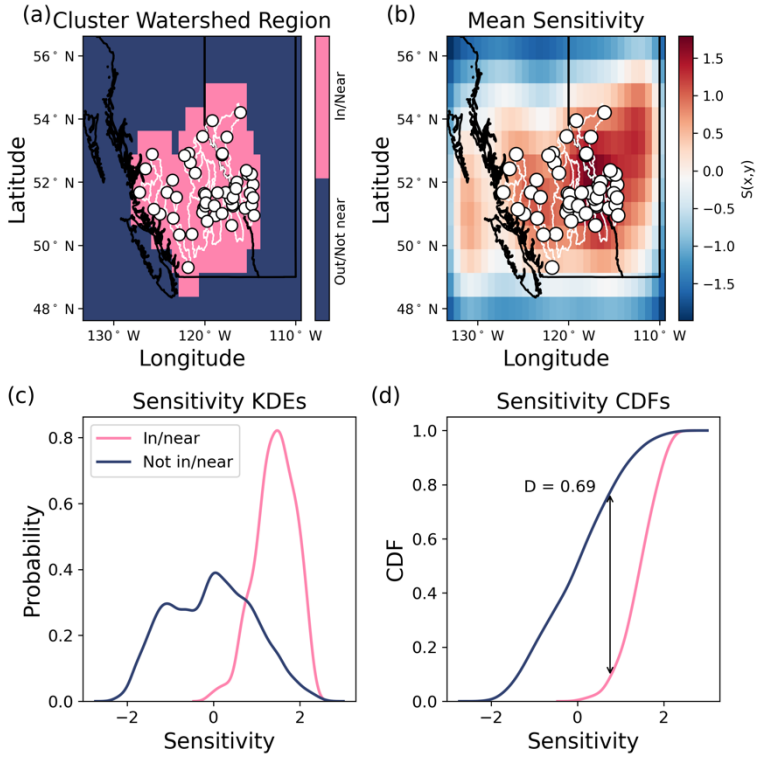


Figure A5: Steps of calculating the D-statistic from the Kolmogorov-Smirnov test. A) The mask of pixels which are either within/near the cluster watersheds (pink), or are within 1 pixel in distance from the watershed boundaries (blue). B) The mean sensitivity evaluated over the test period for the ensemble of models, with red indicating more sensitive and blue indicating less sensitive. C) The sensitivity distributions (within/near in pink and not within/near in blue), calculated by kernel density estimation (KDE) using Gaussian kernels and Scott's rule for kernel bandwidth calculation. D) The Kolmogorov-Smirnov D-statistic is calculated from the sensitivity cumulative density functions (CDFs).

https://doi.org/10.5194/hess-2021-113
Preprint. Discussion started: 11 March 2021

Hydrology and Earth System Sciences
Discussions

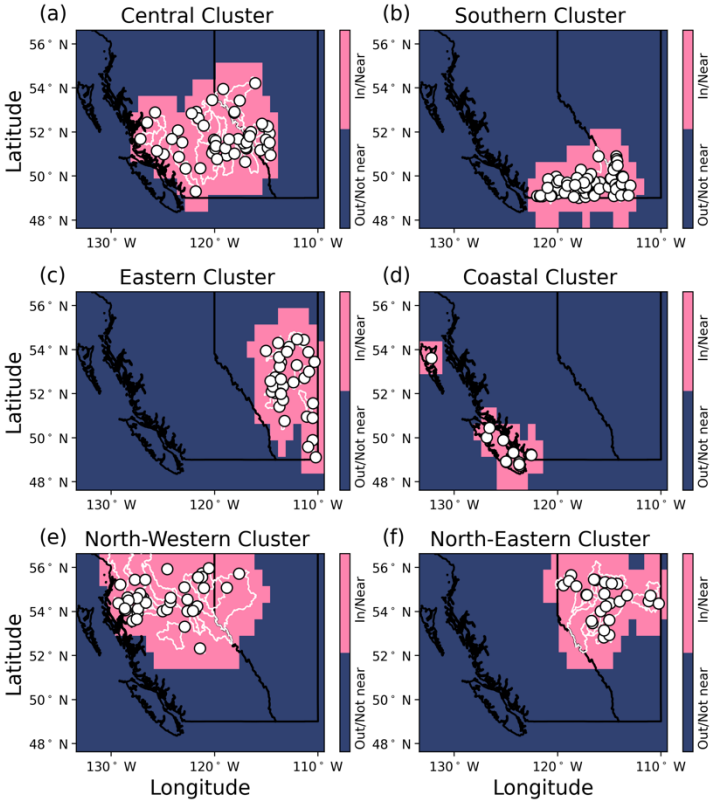


Figure A6: The areas of the study region which are within/near watersheds (pink) of each cluster, and those which are outside/far from the watersheds (blue).