Snowpack-Driven Streamflow Predictability under Future Climate: Contrasting Changes across Two Western Canadian River Basins

RAJESH R. SHRESTHA,a YONAS B. DIBIKE,a AND BARRIE R. BONSALb

a Watershed Hydrology and Ecology Research Division, Environment and Climate Change Canada, University of Victoria, Victoria, British Columbia, Canada
b Watershed Hydrology and Ecology Research Division, Environment and Climate Change Canada, Saskatoon, Saskatchewan, Canada

(Manuscript received 3 November 2021, in final form 8 April 2022)

ABSTRACT: Anthropogenic climate change–induced snowpack loss is affecting streamflow predictability, as it becomes less dependent on the initial snowpack conditions and more dependent on meteorological forecasts. We assess future changes to seasonal streamflow predictability over two large river basins, Liard and Athabasca in western Canada, by approximating streamflow response from the Variable Infiltration Capacity (VIC) hydrologic model with the Bayesian regularized neutral network (BRNN) machine learning emulator. We employ the BRNN emulator in a testbed ensemble streamflow prediction system by treating VIC-simulated snow water equivalent (SWE) as a known predictor and precipitation and temperature from GCMs as ensemble forecasts, thereby isolating the effect of SWE on streamflow predictability. We assess warm-season mean and maximum flow predictability over 2041–70 and 2071–2100 future periods against the 1981–2010 historical period. The results indicate contrasting patterns of change, with the predictive skills for mean flow generally declining for the two basins, and marginally increasing or decreasing for the headwater subbasins. The predictive skill for maximum flow declines for the relatively warmer Athabasca basin and improves for the colder Liard basin and headwater subbasins. While the decreasing skill for the Athabasca is attributable to substantial loss in SWE, the improvement for the Liard and headwaters can be attributed to an earlier maximum flow timing that reduces the forecast horizon and offsets the effect of SWE loss. Overall, while the future change in SWE does affect the streamflow prediction skill, the loss of SWE alone is not a sufficient condition for the reduction in streamflow predictability.

SIGNIFICANCE STATEMENT: The purpose of this study is to evaluate potential changes in seasonal streamflow predictability in relation to snowpack change under future climate. This is highly relevant because snowpack storage provides a means of predicting available freshet water supply, as well as peak flow events in cold regions. We use a machine learning model as an emulator of a hydrologic model in a testbed ensemble prediction system. Our results provide insights on hydroclimatic controls and interactions that affect future streamflow predictability across two river basins in western Canada. We conclude that besides snowpack, predictability depends on a number of other factors (basin/subbasin characteristics, streamflow variables, and future periods), and the loss of snowpack alone is not a sufficient condition for the reduction in streamflow predictability.

KEYWORDS: Climate change; Climate prediction; Hydrologic models; Model evaluation/performance; Neural networks; Seasonal forecasting; Snowpack; Streamflow; Water resources

1. Introduction

Snowmelt-driven spring and summer runoff is an important source of freshwater supply in the cold regions of the world, contributing to more than 50% of runoff in many areas of the Northern Hemisphere (Qin et al. 2020), including parts of the western North America (Li et al. 2017; Shrestha et al. 2021). The delayed release of the accumulated winter snowpack as snowmelt runoff supplements water supply during summer when demands are the highest (Li et al. 2017). In many river basins, the snowmelt-driven runoff contributes a large fraction of reservoir storage demand, such as for agriculture and hydropower generation (Barnett et al. 2005; Qin et al. 2020), which would otherwise be unmet by instantaneous rainfall runoff (Mankin et al. 2015). Additionally, snowmelt-driven freshet is a major driver of floods across cold regions (Buttle et al. 2016; Tarasova et al. 2019).

From the water management perspective, snowpack storage provides a means of estimating or predicting available freshet water supply, as well as extreme events (i.e., floods and drought) in snowmelt-dominated regions. Studies have demonstrated strong dependencies of streamflow on snowpack in the western Canadian watersheds, including high predictive capability of winter and early spring discharge from snowpack volume (Dyer 2008) and high predictability of annual...
maximum flow and mean spring flow from maximum snowpack storage (Curry and Zwiers 2018; Dibike et al. 2021). Given such relationships, the memory of winter snowpack storage adds value to short-range streamflow prediction as an influential initial condition, e.g., for flood forecasting (e.g., Zahmatkesh et al. 2019; Vionnet et al. 2020), reservoir inflow forecasting (e.g., Samuel et al. 2019; Cohen et al. 2020), and seasonal streamflow prediction (e.g., DeChant and Moradkhani 2011; Shrestha et al. 2015; Li et al. 2019). Furthermore, the streamflow prediction skill attributable to the snowpack initial conditions is often higher than that of the meteorological forecasts (Wood and Schaeke 2008), with the highest predictability at seasonal scales arising in basins with large winter snow accumulation (van Dijk et al. 2013; Wood et al. 2016).

A warming climate is affecting the seasonal snowpack storage in cold regions, causing decreases in the duration and extent of snow cover (e.g., Derksen et al. 2015; Mudryk et al. 2018) and reductions in the mean and maximum snow depths (e.g., Fye et al. 2017; Mote et al. 2018). The multidecadal declining snowpack trends in these regions are projected to continue and intensify in the future (Mudryk et al. 2020). The snow storage loss could make future streamflow prediction more dependent on meteorological forecasts, which are generally less reliable, and less dependent on snowpack initial conditions, which are more reliable. For instance, future high-flow events caused less frequently by snowmelt and more frequently by rainfall events (e.g., Chegwidden et al. 2020) could become less predictable. Furthermore, precipitation changes, for instance, increasing magnitude and intensity of extreme precipitation in the future climate (e.g., IPCC 2018; Zhang et al. 2019), could make hydrologic predictions less reliable. Less predictable mean and extreme spring/summer streamflow conditions will make water management more challenging, including reservoir storage management (Harpold et al. 2017) and flood and drought preparedness.

Previous studies indicated that snowpack-driven change in streamflow predictability varies in different regions/periods, for instance, increase in the decadalscale historical summer streamflow predictability in the Sierra Nevada watersheds (He et al. 2016), decline in the future drought predictability over most of the western United States (Livneh and Badger 2020), and unchanged future summer runoff predictability over the Columbia River headwaters (Tsuruta and Schnorbus 2021). Such spatially heterogeneous streamflow predictability response may be influenced by a number of factors, including the magnitude of snowpack change (e.g., Shrestha et al. 2021), the magnitude and timing of streamflow change (e.g., Tan and Gan 2015; Dudley et al. 2017), precipitation bias (e.g., Pechlivanidis et al. 2020), and their interactions. Furthermore, streamflow predictability response may vary over different future periods and flow conditions (e.g., maximum and mean flow). Thus, besides snowpack change, the spatial and temporal variability of future streamflow predictability need to be considered in relation to various hydroclimatic controls and their interactions.

It is a common practice to evaluate streamflow prediction skills using retrospective forecasts or hindcasts over historical periods from process-based hydrologic models (e.g., Shrestha et al. 2015; Wood et al. 2016; Li et al. 2019; Vionnet et al. 2020). However, for future periods, assessing change in predictability is constrained by the lack of observational data, with model-to-model comparison providing the only feasible option. A framework for such comparison could include treating projections from a robust hydrologic model as pseudo observations and using a flexible and efficient statistical emulation as a prediction model. An emulation model provides a means of approximating a subcomponent of a large and complex simulation model in a flexible and computationally efficient framework (Grow and Hilton 2018), thereby allowing an evaluation of the cause–effect relationship among specific variables of interest (e.g., snowpack and streamflow in this study). Given that emulators allow a consideration of a limited number of variables, they provide dimension reduction (relative to process-based models) and facilitate computationally efficient evaluations, such as sensitivity and variable importance analysis and large ensemble streamflow prediction. Statistical emulation has been widely used for hydroclimatic assessments, such as to provide computationally efficient future hydrologic projections (Vano and Lettenmaier 2013; Schnorbus and Cannon 2014), and to analyze controls and sensitivities of climatic variables on hydrologic responses (Shrestha et al. 2019; Chegwidden et al. 2020). Additionally, statistical regression approaches have been used to detect changes in streamflow predictability in the historical climate (He et al. 2016) and to evaluate future warm-season streamflow predictability under declining snowpack (Livneh and Badger 2020). A potential improvement in this respect is the application of machine learning (ML) methods, which are increasingly being used to extract pattern and insights from geospatial data and provide powerful and flexible tools for analyzing complex and multifaceted problems in climate change science by utilizing exiting data and model simulations (Reichstein et al. 2019; Huntingford et al. 2019).

This study advances the understanding of potential changes in streamflow predictability in response to future snowpack change under anthropogenic climate change scenarios. We evaluate streamflow predictability in two river basins in western Canada, with differing historical and future snowpack responses, by focusing on two key research questions: (i) how will future changes in precipitation, temperature and snowpack storage shift the climatic controls on warm-season mean and peak flows, and (ii) how will the shifts in climatic controls affect future streamflow predictability? To address these questions, we employ an artificial neural network (ANN) ML model as an emulator of historical and projected hydrologic model simulations. We use the ANN emulator for evaluating relative controls of snowpack and meteorological drivers (seasonal precipitation and temperature) on streamflow responses. Further, we apply an ensemble streamflow prediction framework, by treating hydrologic model simulated snow water equivalent (SWE) values as “known” predictor, and precipitation and temperature from a suite of GCMs as ensemble forecasts, and we evaluate changes in future prediction skills.

2. Study basins

This study focuses on two large river basins in western Canada, Liard (275 000 km²) and Athabasca (133 000 km²)
that originate in the western Cordillera and flow into the Mackenzie River (Fig. 1). The subarctic Liard basin is colder (mean annual temperature \(\sim-2.0^\circ C\), 1981–2010) than the boreal Athabasca (\(\sim-3.9^\circ C\)), with both basins receiving similar mean annual precipitation amount (\(\sim570\) mm) (Shrestha et al. 2019; Eum et al. 2017). Most of the two basins remain under freezing conditions during October–March and runoff regime is nival, with over 60% of the annual flow occurring from snowmelt-driven runoff (Shrestha et al. 2021). The cooler temperatures in Liard lead to about 3 times higher 1 April snow accumulation (SWE\(_{A1}\)) than in Athabasca (Fig. 2), while September–March precipitation in Liard is only slightly higher than Athabasca (not shown). Coniferous forest and mixed forest are the dominant vegetation types in the two basins (>50% area), with limited resource development in the Liard basin and major oil-sands development in the Athabasca basin (Bonsal et al. 2020).

We selected two hydrometric stations at the upper and lower reaches of each river, consisting of 1) Liard River at the upper crossing (Liard-UC), 2) Liard River near the mouth (Liard-M), 3) and Athabasca River at Hinton (Athabasca-H), and 4) Athabasca River below Fort McMurray (Athabasca-FM) (Fig. 1). This provides a representative sample of streamflow predictability results, including potential differences due to hydroclimatic variability and change (both historical and future) and basin physiography.

3. Methods and data

a. Base model: VIC

We use historical and projected future streamflow and SWE simulations from previous studies of the Liard (Shrestha et al. 2019) and Athabasca (Eum et al. 2017; Dibike et al. 2018) basins using the Variable Infiltration Capacity (VIC) hydrologic model (Liang et al. 1994; Hamman et al. 2018) at 1/16° resolution. These studies used an ensemble of seven and six GCMs, respectively, from phase 5 of the Coupled Model Intercomparison Project (CMIP5) (Taylor et al. 2012) under the representative concentration pathways (RCP) 4.5 and 8.5 (Table S1 in the online supplemental material). Two different downscaling methods were used, consisting of multivariate bias correction/downscaling (Cannon 2018) for Liard, and bias-corrected spatial disaggregation (Wood et al. 2004) for Athabasca to force the VIC models at a daily resolution. Since this study uses spatially averaged and seasonally aggregated precipitation and temperature values, and monthly/seasonal precipitation and temperature fields from different downscaling methods have been found to have comparable skills (Maurer and Hidalgo 2008), uncertainties due to the differences in downscaling methods is not considered to be critical. Readers are referred to the original studies on Liard (Shrestha et al. 2019) and Athabasca (Eum et al. 2017; Dibike et al. 2018) basins for a detailed description of the VIC model setup,
calibration/validation, GCM selection and future climate projections.

b. Emulator model: Artificial neural networks

While the VIC model can provide a process-based approach for evaluating the future change in streamflow predictability, the attribution of future change in streamflow predictability to snowpack storage would be complicated by the large number of state variables in the VIC model (e.g., soil moisture and snowpack storage, thermal condition, etc.). In contrast, an emulator model provides a means of replicating external behavior of the complex VIC model in a flexible and computationally efficient framework, based on the input–output relationship of specific variables of interest. This also allows an attribution of future change in a predictand variable to selected predictor variables (e.g., snowpack and streamflow in this study). We used ANN-based ML method for emulation of the VIC model simulated flow, because of its ability to effectively capture nonlinear relationships between predictor and response variables through well-established relationships.

Fig. 2. Basin-averaged historical and projected future warm-season (April–August) mean temperature and total precipitation, and VIC model simulated SWE$_{A1}$ over Liard-M and Athabasca-FM basins. The points for each scenario/period represent the 7 (6) GCMs for Liard-M (Athabasca-FM), and the filled points denote median values of the 7 (6) GCMs.
training algorithms. ANNs also have a wide range of hydrological applications, including in climate change research (e.g., Maier et al. 2010; Abrahart et al. 2012; Huntingford et al. 2019; Zounemat-Kermani et al. 2020). The structure of ANN used in this study consisted of a two-layer feedforward network with a hyperbolic tangent activation function. Bayesian regularization (hereafter referred to as the BRNN) was used to adjust weights and biases, because of its robustness in model training, i.e., use of an objective stopping criteria for more parsimonious network that is less prone to overfitting and has a reduced need for cross validation (Burden and Winkler 2009). Previous studies have successfully demonstrated the application of BRNNs for streamflow simulation (e.g., Shrestha and Nestmann 2009; Rasouli et al. 2020). We also set up multiple linear regression (MLR) models for the study basins and compared the model performances with the BRNN.

We trained separate BRNN models for two VIC-simulated flow variables: (i) mean April–August flow ($Q_{\text{mean}}$) and (ii) annual maximum flow ($Q_{\text{max}}$) that mostly occurs in June–July in the two basins (Fig. 3). The $Q_{\text{mean}}$ and $Q_{\text{max}}$ are important variables for water management as they represent the warm-season water availability and peak flow with potential of floods, respectively. Input variables for the BRNN to predict both $Q_{\text{mean}}$ and $Q_{\text{max}}$ consisted of VIC-simulated SWEA$_1$ and downscaled seasonal precipitation ($P$) and temperature ($T$) from GCMs, expressed as $Q = f(\text{SWEA}_1, P, T)$. SWEA$_1$ was used instead of maximum SWE because it is a convenient variable for data assimilation and model initialization for streamflow forecasting (e.g., Huang et al. 2017; Gichamo and Tarboton 2019) that has been previously used for the assessment of climate-driven changes in streamflow predictability (He et al. 2016; Livneh and Badger 2020; Tsuruta and Schnorbus 2021). Since SWEA$_1$ represents response to cold-season $P$ and $T$, only warm-season $P$ and $T$ consisting of April–August mean values (AMJJA$_P$ and AMJJA$_T$) for $Q_{\text{mean}}$ and April–July mean values (AMJJ$_P$ and AMJJ$_T$) for $Q_{\text{max}}$ were used as inputs (cold-season $P$ and $T$ are not included to avoid collinearity). Given the predominant role of snowpack on streamflow in this region, and our focus to evaluate streamflow prediction skill with respect to SWEA$_1$ as an initial condition and warm-season $P$ and $T$ as ensemble forecasts (as discussed in section 3c), other state variables, such as soil moisture, were not included. Given the non-stationary changes in the driving and response variables, we trained a single and independent BRNN for each of the three periods (1981–2010, 2041–70, and 2071–2100), and scenarios (historical, RCP4.5, and RCP8.5) by combining all GCM data (7 for Liard and 6 for Athabasca), which are hereafter referred to as Hist 81–10, RCP4.5 41–70, RCP8.5 41–70, RCP4.5 71–00, and RCP8.5 71–00. This avoids having to use the model trained for one period for an evaluation of different period. Hence, for each period, the training data for the Liard and Athabasca BRNN models consisted of 7 GCMs × 30 years and 6 GCMs × 30 years, respectively (see Table S1 for a list of GCMs used).

The BRNN models were setup using the R “brnn” package (Rodriguez and Gianola 2021) and trained with 2–10 neurons in the hidden layer. The optimal number of neurons—with the lowest root-mean-square error criteria (RMSE)—were selected by using the fivefold cross-validation repeated 10 times (each time leaving out 20% of the data for validation) using the R “caret” interface (Kuhn et al. 2021). The repeated $k$-fold cross-validation reshuffles and restratifies the data during each iteration and improves the reliability of the performance estimation (Refaelzadeh et al. 2016). The model selection (number of neurons) is based on the performance of the data left out during cross validation, and similar procedure for selecting the optimal network configuration has been applied in previous studies (e.g., Li et al. 2019; Tyralis et al. 2021). Finally, the BRNNs with selected optimal number of neurons were retrained to take advantage of all available data for each period and improve generalization. The retrained models were used to simulate $Q_{\text{max}}$ and $Q_{\text{mean}}$ responses for each period, which were used for evaluating streamflow prediction skills.

The BRNN emulator also provides a means for analyzing variable importance (VI) of the input variables in relation to the output variable that is consistent with the prediction model. VI also improves the transparency of ML models. We used the R “vip” package (Greenwell et al. 2020) for the VI analysis, by applying a model-agnostic permutation-based method that was originally developed for random forests machine learning (Breiman 2001).

c. Ensemble streamflow prediction

We designed a testbed experiment as a means to extract insights on the change in streamflow predictive skill due to future changes in snowpack storage component. We followed a standard procedure for retrospective ensemble streamflow prediction (ESP) system that combines the initial hydrologic conditions with a resample of climate data drawn from a specific period (e.g., Wood and Lettenmaier 2006; Wood et al. 2016; Shrestha et al. 2015). Specifically, for each 30-yr period, predictor variables consisted of the VIC-simulated SWEA$_1$ values for each year and GCM, and an ensemble of seasonal precipitation and temperature traces from the respective GCM. In other words, we considered SWEA$_1$ as a “known” predictor and precipitation and temperature as “unknown” forecasts. This approach is analogous to an ESP system, with SWEA$_1$ treated as initial condition and seasonal $P$ and $T$ traces excluding the forecast year and the following year from GCMs treated as ensemble forecasts. For example, the 2045 $Q_{\text{mean}}$ ensemble forecast for each GCM uses VIC-simulated 2045 SWEA$_1$, and each of the 28-member AMJJA$_P$ and AMJJA$_T$ consisting of values from 2041 to 2044 and from 2047 to 2070 as predictor variables. The year following the forecast (2046 in the example) was excluded to discount possible autocorrelation in AMJJA$_P$ and AMJJA$_T$. Hence, by using $P$ and $T$ from the same 30 years (except for the forecast year and the following year) as the ensemble forecasts, we conditioned the ESP experiment to be consistent with the 30-yr climate over each period. This method was repeated for each GCM, with each forecast year consisting of 28 × 7 member ensemble forcings for Liard and 28 × 6 member ensemble for Athabasca, thus allowing a comparison of inter-GCM variability of the prediction skills. For the evaluation of model fits and skills, the results for each year were averaged over the
28-member ensembles for each GCM. Thus, the intra-GCM variability of temperature and precipitation over the 28-member ensemble is averaged out from the predicted streamflow, while the effect of SWE$_{A1}$, held constant across the ensemble, remains unchanged.

d. Methods of evaluation

We first evaluated the performance of selected BRNN model—obtained from the fivefold repeated cross-validation method—by comparing with the base VIC model output as “observations,” using three skill metrics: (i) RMSE, (ii) Nash Sutcliffe coefficient of efficiency (NSE) (Nash and Sutcliffe 1970), and (iii) Kling–Gupta efficiency (KGE) (Gupta et al. 2009). In the second step, predictive skills of BRNN in the simulation and ESP modes—hereafter referred to as BRNN-sim and BRNN-ESP, respectively—were analyzed. Two deterministic skill metrics were considered consisting of (i) ratio of RMSE to standard deviation (SD) of observation (RSR) and (ii) KGE. Since SD measures the dispersion from mean value, RSR is equivalent to RMSE skill with climatological mean as the reference. Alternatively, RSR is a measure of predictive skill relative to the naïve climatological mean forecast, with better skill than climatological mean for RSR $\to 1$ and higher skill for RSR $\to 0$. KGE is based on decomposition of model performance into correlation, variability bias, and mean bias,
with KGE = 1 denoting perfect agreement between simulations and observations, and KGE ≈ −0.41 when benchmarked against the climatological mean (Knoben et al. 2019).

In conjunction with the deterministic metrics, we analyzed two commonly used categorical metrics in the ESP studies (WMO 2011; Lavaysse et al. 2015; Shrestha et al. 2015): (i) percentage correct statistics (PCS) and (ii) true skill statistics (TSS). These metrics were used to assess the model’s skill to predict frequencies within discrete boundaries and above/below thresholds, respectively. Hence, the predictability of streamflow distributions and extreme categories—both with implications on water management—are explicitly analyzed. The PCS was evaluated relative to the tercile boundaries of the observations, i.e., values below the 33rd percentile are below normal, values between the 33rd and 67th percentiles are near normal, and values above the 67th percentile are above normal. The PCS, which gives a combined total of predictions falling in the three categories divided by the total number of years (Table 1), varies between 0% and 100%, and the higher the PCS value, the better the model performance. Since PCS is based on the discrete tercile boundaries, a 33% skill could be achieved by chance alone (Shrestha et al. 2015). Thus, PCS skill > 33% is a useful metric to evaluate if the model performance is better than chance. The TSS—also referred to as Hanssen and Kuipers’ skill score (Hanssen and Kuipers 1965)—measures the ability to distinguish “yes” event and “no” event forecasts. It is expressed as the difference between hit rate and false alarm rate based on a contingency table (Table 2). TSS ranges between −1 and 1, denoting worst and best predictions, respectively. In this study, PCS was used to assess the overall skill for all three categories, while TSS was best predictions, respectively. In this study, PCS was used to assess the overall skill for all three categories, while TSS was best predictions, respectively. In this study, PCS was used to as-

\[
\text{PCS} = \frac{A + B + C}{N}, \quad (1)
\]

where \(N\) = total number of simulations:

\[
\text{TSS} = \frac{ad - bc}{(a + c)(b + d)}. \quad (2)
\]

4. Results and discussion

a. Projected snowpack and streamflow changes

First, we summarize the historical and projected future AMJJA_P, AMJJA_T, and SWEA1 for the two basins in terms of 30-yr means of each GCM, averaged over the basins (Fig. 2). The projected future temperature and precipitation generally increase progressively over RCP4.5 41–70, RCP8.5 41–70, RCP4.5 71–00, and RCP8.5 71–00, compared to Hist 81–10 period. The projected future SWEA1 values differ between the two basins, with much higher values in Liard-M compared to Athabasca-FM. Specifically, the fraction of SWEA1 remaining under RCP8.5 71–00 relative to Hist 81–10 is ~80% for Liard and <50% for Athabasca, with the values for RCP4.5 41–70, RCP8.5 41–70 and RCP4.5 71–00 also showing higher proportional declines for Athabasca-FM. These differences mainly stem from the proximity of the basin’s winter temperature state to the freeze/melt threshold, which lead to smaller snowfall fraction and SWEA1 accumulation for the warmer Athabasca-FM compared to the colder Liard-M (Shrestha et al. 2021). The SWEA1 responses for the Liard headwater subbasin (Liard-UC) show higher median values for RCP4.5 41–70, RCP8.5 41–70, and RCP4.5 71–00 than Hist 81–00, which is in contrast to the response of the entire basin (Liard-M) (Fig. S1). The slight increase in SWEA1 is likely because projected precipitation increase is able to compensate the temperature-driven SWEA1 decline in the warmer headwaters. However, for the Athabasca-H headwater subbasin, SWEA1 declines for all future scenarios/periods

---

**Table 1. Comparison table used for calculation of percentage correct skill statistics (PCS).**

<table>
<thead>
<tr>
<th>Observed event</th>
<th>&lt;33rd percentile</th>
<th>≥33rd percentile and ≤67th percentile</th>
<th>&gt;67th percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Hit (a)</td>
<td>False alarm (c)</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Miss (b)</td>
<td>Correct rejection (d)</td>
<td></td>
</tr>
</tbody>
</table>

**Table 2. Contingency table used for the calculation of Hanssen and Kuipers (1965) true skill statistics (TSS).**

<table>
<thead>
<tr>
<th>Observed event</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted event</td>
<td>Hit (a)</td>
<td>False alarm (c)</td>
</tr>
<tr>
<td>Yes</td>
<td>Miss (b)</td>
<td>Correct rejection (d)</td>
</tr>
</tbody>
</table>

---

**Fig. 2.** The projected future temperature and precipitation generally increase progressively over RCP4.5 41–70, RCP8.5 41–70, RCP4.5 71–00, and RCP8.5 71–00, compared to Hist 81–10 period. The projected future SWEA1 values differ between the two basins, with much higher values in Liard-M compared to Athabasca-FM.)

---

**Fig. S1.** The slight increase in SWEA1 is likely because projected precipitation increase is able to compensate the temperature-driven SWEA1 decline in the warmer headwaters. However, for the Athabasca-H headwater subbasin, SWEA1 declines for all future scenarios/periods.
because of small or no change in precipitation in conjunction with warming temperature.

Results for $Q_{\text{mean}}$ and $Q_{\text{max}}$ generally follow similar trajectories for the two basins (Fig. 3), with the highest median $Q_{\text{mean}}$ and $Q_{\text{max}}$ values in RCP8.5 41–70 and RCP4.5 71–00 for Liard-M and Athabasca-FM, respectively. The median values of all four future scenarios/periods are higher than Hist 81–10, with the increases ranging between ~1% and ~10%. The $Q_{\text{max}}$ timing also becomes progressively earlier and the range across the GCMs is generally wider over the latter periods/warmer scenarios. In the case of the two headwater subbasins (Fig. S2), there are more pronounced increases in $Q_{\text{mean}}$ for both Liard-UC and Athabasca-H, while $Q_{\text{max}}$ for Athabasca-H show some decreases with warmer scenarios/latter periods. This suggests differences in precipitation, temperature, and SWE interactions on flow response within the basins. However, the trajectories of $Q_{\text{max}}$ timing over the headwaters are generally similar to the entire basins.

b. BRNN performance and variable importance

The selected best BRNN model results obtained from the repeated fivefold cross-validation show varying model performances, with differences between the periods, scenarios, and prediction variables, as well as the basins (Table 3). In general, the $Q_{\text{mean}}$ results are superior to $Q_{\text{max}}$ in terms of NSE and KGE criteria, suggesting better predictability of the mean streamflow condition than extreme condition when using SWEA1 and seasonal temperature and precipitation as input variables. The RMSE values for the projected future periods/RCPs tend to be higher than Hist 81–10, as they are influenced by higher projected streamflow compared to the historical values (Fig. 3). Compared to the MLR models (Table S2), the BRNN models showed slightly better performance for $Q_{\text{mean}}$ and larger improvements for $Q_{\text{max}}$ in terms of the three performance metrics. This is because the three input variables explain a larger fraction of variance using MLR, and thus there is a smaller room for improvement when using the more sophisticated BRNN. In contrast, larger improvements in performance of $Q_{\text{max}}$ were achieved by using BRNN because the three input variables explain a smaller fraction of variance using MLR.

The scatterplot of BRNN-sim and BRNN-ESP against VIC results illustrate the ability of the trained ANN to reproduce the low and high $Q_{\text{mean}}$ and $Q_{\text{max}}$ values for the two basins, as well as scenarios/future periods (Fig. 4). Usually the higher values show larger scatter from the 1:1 line than the lower values. Similar to RMSE, the scatter—as shown by SD—tend to be higher for projected future periods/RCPs than the historical period. This is because of increasing range of $Q_{\text{mean}}$ and $Q_{\text{max}}$ values with RCP scenarios and future periods compared to Hist 81–10. As expected, the BRNN-ESP (prediction mode) models have larger spreads (i.e., larger SDs) compared to the BRNN-sim (simulation mode).

The VI percentage scores obtained from BRNN models—in terms of relative controls of predictor variables on $Q_{\text{mean}}$ and $Q_{\text{max}}$ responses—indicate dominant controls of SWEA1 and seasonal precipitation, especially for $Q_{\text{mean}}$ (Fig. 5). Consistent with the previous study of the Liard basin (Shrestha et al. 2019), seasonal temperature has a minor influence on $Q_{\text{mean}}$ and $Q_{\text{max}}$ while the influence of precipitation is lower in this study than the previous study due to the inclusion of SWEA1. The relative controls vary considerably across the two basins and periods as well as historical and future periods. Specifically, while SWEA1 exerts larger controls on $Q_{\text{mean}}$ for both basins over Hist 81–10 period, the influence of SWEA1 declines and AMJJA_P increases successively from 41–70 to 71–00 and RCP4.5 to RCP8.5. In the case of $Q_{\text{max}}$, however, the VI changes are in opposite directions for the two basins, with the influence of SWEA1 increasing for Liard-M and decreasing for Athabasca-FM along the two RCP scenarios and future periods. Such increasing controls of SWEA1 for Liard-M $Q_{\text{max}}$ despite the decline in SWEA1 (Fig. 2) appear counterintuitive. However, as stated earlier, the percent declines in SWEA1 for Liard-M are relatively small compared to the total SWEA1 (except RCP8.5 71–00). Additionally, progressively earlier $Q_{\text{max}}$ timing (Fig. 3) shortened the forecast horizon relative to 1 April and likely contributed to increasing influence of SWEA1 on $Q_{\text{max}}$. In contrast, although $Q_{\text{max}}$ are also projected to occur earlier for Athabasca-FM, steeper declines in SWEA1 combined with increasing contribution of AMJJA_P likely played a more influential role in reducing the control of SWEA1 on $Q_{\text{max}}$.

<table>
<thead>
<tr>
<th>Period</th>
<th>Neurons</th>
<th>RMSE</th>
<th>NSE</th>
<th>KGE</th>
<th>Neurons</th>
<th>RMSE</th>
<th>NSE</th>
<th>KGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hist 1981–2010</td>
<td>3</td>
<td>293</td>
<td>0.83</td>
<td>0.87</td>
<td>2</td>
<td>1415</td>
<td>0.55</td>
<td>0.63</td>
</tr>
<tr>
<td>RCP4.5 2041–70</td>
<td>4</td>
<td>321</td>
<td>0.85</td>
<td>0.89</td>
<td>2</td>
<td>1726</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td>RCP8.5 2041–70</td>
<td>2</td>
<td>335</td>
<td>0.84</td>
<td>0.88</td>
<td>2</td>
<td>1679</td>
<td>0.63</td>
<td>0.70</td>
</tr>
<tr>
<td>RCP4.5 2071–2100</td>
<td>9</td>
<td>345</td>
<td>0.86</td>
<td>0.89</td>
<td>9</td>
<td>1615</td>
<td>0.65</td>
<td>0.72</td>
</tr>
<tr>
<td>RCP8.5 2071–2100</td>
<td>5</td>
<td>399</td>
<td>0.79</td>
<td>0.84</td>
<td>3</td>
<td>1777</td>
<td>0.64</td>
<td>0.71</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Period</th>
<th>Neurons</th>
<th>RMSE</th>
<th>NSE</th>
<th>KGE</th>
<th>Neurons</th>
<th>RMSE</th>
<th>NSE</th>
<th>KGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hist 1981–2010</td>
<td>6</td>
<td>120</td>
<td>0.74</td>
<td>0.80</td>
<td>7</td>
<td>320</td>
<td>0.72</td>
<td>0.78</td>
</tr>
<tr>
<td>RCP4.5 2041–70</td>
<td>8</td>
<td>116</td>
<td>0.73</td>
<td>0.79</td>
<td>6</td>
<td>421</td>
<td>0.53</td>
<td>0.60</td>
</tr>
<tr>
<td>RCP8.5 2041–70</td>
<td>3</td>
<td>143</td>
<td>0.77</td>
<td>0.82</td>
<td>7</td>
<td>450</td>
<td>0.68</td>
<td>0.74</td>
</tr>
<tr>
<td>RCP4.5 2071–2100</td>
<td>6</td>
<td>122</td>
<td>0.81</td>
<td>0.85</td>
<td>3</td>
<td>439</td>
<td>0.67</td>
<td>0.74</td>
</tr>
<tr>
<td>RCP8.5 2071–2100</td>
<td>4</td>
<td>128</td>
<td>0.85</td>
<td>0.88</td>
<td>6</td>
<td>534</td>
<td>0.61</td>
<td>0.69</td>
</tr>
</tbody>
</table>

TABLE 3. Number of neurons corresponding to the selected best performing BRNN. The model performance is based on the data left out during fivefold repeated cross validation.
The contribution of SWE$_{A1}$ on VI percent scores are higher in case of the two headwater subbasins than the entire basins (Fig. 5 and Fig. S3). Furthermore, in contrast to the entire basins, VI scores for $Q_{\text{mean}}$ show smaller or no change. These differences are likely due to smaller changes in SWE$_{A1}$ in the colder headwater basins (Fig. S1). The influences of smaller SWE$_{A1}$ change along with the earlier $Q_{\text{max}}$ timing (Fig. S2) are also reflected in terms of increases in VI scores of SWE$_{A1}$ with warmer scenarios/latter periods, for both Liard-UC and Athabasca-H $Q_{\text{max}}$ simulations. The $Q_{\text{max}}$ VI changes for Athabasca-H contrast with the Athabasca-FM, indicating differences in controls between the headwater subbasin and entire basin. These differences in model performance and VI scores can be expected to influence the future streamflow predictability.

c. Changes in deterministic streamflow predictability

We present comparisons of BRNN streamflow prediction skills relative to VIC simulations for the historical and future climates (Figs. 6 and 7). The range of BRNN results represents the inter-GCM variability of the driving precipitation and temperature, and VIC-simulated SWE$_{A1}$ for a given period/scenario.

The results obtained from the GCM ensemble generally depict consistent patterns, e.g., higher RSR values (thus lower skill) for BRNN-ESP compared to BRNN-sim (Fig. 6). The BRNN-sim RSR skills for the two basins generally show small increases or decreases under RCP4.5/RCP8.5 scenarios and 41–70/71–00 periods compared to Hist 81–10 period, suggesting that the RSR limits remain mostly unaffected by the future changes in seasonal precipitation, temperature and SWE$_{A1}$. Exceptions are the Athabasca-FM $Q_{\text{max}}$ along with the Liard-M $Q_{\text{mean}}$ and $Q_{\text{max}}$ projections for RCP8.5 71–00, for which higher RSR values imply reduced predictability due to future changes in the three input variables. As expected, the BRNN-ESP RSR skills for $Q_{\text{mean}}$ are inferior to BRNN-sim, and the skills decline successively with the RCP scenarios and two future periods. These results align with the declining SWE$_{A1}$ magnitudes (Fig. 2) and decrease in the SWE$_{A1}$ VI (Fig. 5). In the
case of $Q_{\text{max}}$ for Liard-M, the BRNN-ESP RSR skills improve with the two scenarios/future periods, which is consistent with the increasing VI of SWE$_{A1}$, and related to relatively smaller SWE$_{A1}$ loss (Fig. 2e) and earlier $Q_{\text{max}}$ timing (Fig. 3e). Thus, the reduced forecast horizon of $Q_{\text{max}}$ relative to 1 April compensate the small SWE$_{A1}$ loss, and lead to improved RSR skill metric. This contrasts with the decline in RSR skill for $Q_{\text{mean}}$, because $Q_{\text{mean}}$ is an average condition for April–August and not affected by the change in timing. In the case of Athabasca-FM $Q_{\text{max}}$, the BRNN-ESP RSR skill for Hist 81–10 as well as, future scenarios/periods are close to 1, indicating that the model has very little predictive skill beyond the naïve forecast of the climatological mean.

Similarly, the ensemble range generally depicts consistent patterns of reduced KGE skills, when BRNN-ESP results are compared with BRNN-sim (Fig. 7). The BRNN-ESP KGE skills also have wider ranges compared to BRNN-sim, which is indicative of the larger inter-GCM precipitation and temperature variability in the BRNN-ESP mode. The BRNN-sim KGE skills for the two basins are similar across the historical/future periods and RCPs, except slight improvement for Liard-M $Q_{\text{max}}$ future simulations. Thus, similar to RSR, the KGE skills also remain mostly unaffected by the future changes in SWE$_{A1}$, precipitation and temperature. Further, the general trajectories of the KGE skills for BRNN-ESP, i.e., decline of Liard-M $Q_{\text{mean}}$ and Athabasca-FM $Q_{\text{mean}}$ and $Q_{\text{max}}$ along with the improvements in Liard-M $Q_{\text{max}}$ simulations, under RCP scenarios and future periods are consistent for both skill metrics. As stated earlier, the improvement in skills for Liard-M $Q_{\text{max}}$ ESP simulations can be attributed to the increasing VI of SWE$_{A1}$, due to the reduced forecast horizon. In contrast, the decreased KGE skill for the Athabasca-FM despite the reduced forecast horizon for $Q_{\text{max}}$ can be linked to the steep SWE$_{A1}$ reduction (Fig. 2b), coupled with the increased contribution of precipitation forecast (Fig. 5d). The change in RSR skill metrics for the Liard-UC and Athabasca-H subbasins are less pronounced (Fig. S4) than the entire basins (Fig. 6), although there are similar tendencies of decline in skills for $Q_{\text{mean}}$ with warmer scenarios/latter periods. In the case of KGE skills, the results indicate improvement or decline for $Q_{\text{mean}}$, depending on the period/scenario (Fig. S5). These milder changes in predictive skills correspond to the smaller changes in VI (Fig. S3) in the headwater subbasins as a result of smaller SWE$_{A1}$ changes (Fig. S2). In the case of $Q_{\text{max}}$ the results indicate an overall improvement for both Liard-UC and Athabasca-H, which are similar to Liard-M, again suggesting the combined
influence of smaller SWE$_{A1}$ decline coupled with earlier $Q_{\text{max}}$ timing. The contrasting responses in $Q_{\text{max}}$ skills for the Athabasca-H and Athabasca-FM indicate that the future changes in streamflow predictability could be highly heterogeneous over a basin, in addition to across different basins, i.e., Liard-M versus Athabasca-FM.

d. Changes in categorical streamflow predictability

The range of categorical skills obtained from the GCM ensembles depict expected patterns, such as generally higher PCS and TSS metrics for BRNN-sim compared to BRNN-ESP (Figs. 8 and 9). Additionally, the PCS for BRNN-ESP are over the threshold of 33rd percentile, thus, the residual skills due to snowpack remain above the naïve forecast. However, the change in skill values among different periods/scenarios are generally small or nonexistent, especially for PCS (Fig. 8). An exception is the notable improvements in PCS skill for Liard BRN-ESP $Q_{\text{max}}$ simulation, which is consistent with the deterministic results. The TSS values for upper tercile of $Q_{\text{mean}}$ (high flow) and lower tercile of $Q_{\text{max}}$ (warm-season low flows) are highly variable, especially for BRNN-ESP (Fig. 9). In this case, BRNN-sim results for all five scenarios/time periods for Liard-M are very similar, implying little influence of change in SWE$_{A1}$, precipitation, and temperature. The TSS values for Athabasca-FM BRNN-ESP are lower than those for Liard-M for all periods/scenarios, which is likely due to lower VI of SWE$_{A1}$ for the former, suggesting smaller SWE$_{A1}$ values also affect the warm-season high and low flow predictability. There is no consistent pattern of change in the TSS skills for Athabasca-FM over different periods/scenarios, probably because Hist 81–10 have low skill (<0.3, no skill = 0), thus reduced contributions of SWE$_{A1}$ in future periods/scenarios play less of a role in the TSS skills. Again, the TSS skills for Liard-M $Q_{\text{max}}$ improve over the future periods/scenarios compared to Hist 81–10, thus, the aforementioned improvement due to shortened forecast horizon also apply to the upper tercile of the $Q_{\text{max}}$ distributions.

The categorical skills for the headwater subbasins show slightly better performance than the entire basins (Figs. S6 and S7). Both PCS and TSS values are generally higher for Liard-UC and Athabasca-H compared to Liard-M and Athabasca-FM, attributable to higher snowpack storage and VI scores for the headwater basins. However, there are no discernable change in skills with future scenarios/periods, with the exception of a
slight improvement for Liard-UC $Q_{\text{mean}}$ and a slight decline for Athabasca-H $Q_{\text{mean}}$ for PCS and TSS. Overall, while the direction of change for the categorical skills are generally consistent with the deterministic skills, the magnitude of changes for both headwaters and entire basins are less pronounced for the categorical skills.

e. Discussion

This study found contrasting patterns of change in streamflow predictability skills among the basins/subbasins and streamflow variables. The prominent changes in BRNN-ESP skills, which are conditioned on the future SWEA1 change, are the decline in $Q_{\text{mean}}$ predictability, for Liard-M and Athabasca-FM basins. This result is generally consistent with the overall decline in the predictability of warm-season (April–July) streamflow under future snowpack loss in the western United States (Livneh and Badger 2020). We found the change in VI as a main determinant of the change in predictability, with reduction in VI of SWEA1 leading to lower predictability and vice versa. Furthermore, increase in VI of precipitation will make future streamflow prediction more dependent on unknown precipitation forecasts. The connection between streamflow predictability and SWEA1 change is further exemplified by the results for colder headwater subbasins (Figs. S4 and S5), where the SWEA1 VII for $Q_{\text{mean}}$ simulations show small increases/decreases (Fig. S3), in response to relatively smaller increase or decrease in SWEA1 (Fig. S1). Consequently, the $Q_{\text{mean}}$ RSR and KGE skills for the headwater subbasins increase/decrease marginally. These results are also generally consistent with no change or small improvement in the $Q_{\text{mean}}$ prediction skills over the northernmost headwater basin of the Columbia River (Tsuruta and Schnorbus 2021).

The relationship between VI and predictability scores is further reinforced by the $Q_{\text{max}}$ results, i.e., the improvement and decline of prediction skills with the increase or decrease of VI scores for Liard-M and Athabasca-FM, respectively. The $Q_{\text{max}}$ RSR and KGE skills for the colder Liard-UC and Athabasca-H headwaters also improve with the increase in SWEA1 VII (Figs. S4 and S5). As stated earlier, the improved predictability of $Q_{\text{max}}$ can be explained in terms of the reduced forecast horizon due to an earlier $Q_{\text{max}}$ timing, offsetting the effect of relatively small SWEA1 loss. However, over the entire Athabasca basin, the larger proportion of SWEA1 loss—despite the reduced forecast horizon—leads to the decline in $Q_{\text{max}}$ prediction skill. Overall, the results suggest larger potential declines in the $Q_{\text{mean}}$ and $Q_{\text{max}}$ predictability in the warmer southern basins of western Canada, such as Fraser and Columbia, which are projected to experience steep SWE decline (Shrestha et al. 2021).
The changes in categorical skills are less pronounced than the deterministic skills for both headwaters and entire basins. The categorical skills, being based on the distributions among tercile boundaries, are more tolerant to errors than the deterministic skills. This also suggests that the errors generally lie within the respective tercile boundary. In other words, SWEₐ change has a larger effect on the magnitude than the distribution of streamflow. Overall, the control of SWE on streamflow predictability is highly complex and while the reduction in SWE does affect the streamflow prediction skill, it is possible to have an improvement in streamflow predictability with snowpack loss. The direction of change in the skill metric depends on the streamflow variable considered and magnitude of SWE loss, as well as interactions with primary driving variables (precipitation and temperature). Given that the interactions may influence or even counter the effect of SWE loss on streamflow predictability, the loss of SWE alone is not a sufficient condition for the reduction in streamflow predictability.

5. Conclusions

This study provided an assessment of future changes in streamflow predictability with respect to potential SWE changes in a warmer climate, over two river basins (Liard and Athabasca) in western Canada. Both basins have a snowmelt-driven streamflow regime, but higher fractional projected SWE decline in the relatively warmer Athabasca basin. We developed a Bayesian regularized neural network (BRNN) based testbed ensemble streamflow prediction (ESP) system by considering historical and future projections from VIC hydrologic model simulations as “observations.” We conditioned BRNN-ESP with the VIC simulated 1 April SWE (SWEₐ) as known predictors and ensemble traces of seasonal precipitation and temperature as unknown forecasts. We analyzed a range of deterministic and categorical skill metrics for Q₁ and Qₘ₉₉ simulations, which correspond to the warm-season mean and extreme flow conditions.

The study also demonstrated the flexibility and computational efficiency of the BRNN emulator for evaluating streamflow predictability. Specifically, by isolating SWEₐ and warm-season P and T as the drivers of streamflow response from the large-scale VIC model, an ESP system could be conditioned on SWEₐ as a known predictor and seasonal P and T as unknown forecasts. The emulator model also allowed two levels of skill evaluation, with the simulation mode providing the ability of the three variables to capture Q₁ and Qₘ₉₉ response, and the ESP mode providing future change in

![Figure 8](image-url)
SWE\textsubscript{A1}-driven $Q_{\text{mean}}$ and $Q_{\text{max}}$ predictability. Additionally, BRNN provided a consistent means of evaluating both variable importance and streamflow predictability. In regions where other variables such as soil moisture have a larger influence, the methodology could be adopted to assess their effects on streamflow predictability. However, the overall skills are limited by the ability of BRNN to emulate the VIC simulated streamflow, which is provided by BRNN-sim results. The study also assumed the VIC model future projections as pseudo observations, which are affected by a number of uncertainties (e.g., Hattermann et al. 2018).

Our results indicated that changing variable importance of SWE\textsubscript{A1} is a key determinant of the change in future $Q_{\text{mean}}$ and $Q_{\text{max}}$ predictability over the two RCP scenarios and future periods (RCP4.5 and RCP8.5 scenarios over 2041–70 and 2071–2100) compared to the historical period (1980–2010). The decline in SWE\textsubscript{A1} generally led to a decline in its variable importance and predictability of $Q_{\text{mean}}$ and $Q_{\text{max}}$ with the warmer scenarios/latter periods, while the increase in variable importance of precipitation will make future streamflow prediction more dependent on unknown precipitation forecasts. An exception is the increasing future $Q_{\text{max}}$ predictability for the colder Liard and headwater subbasins, likely due to relatively smaller declines in SWE\textsubscript{A1} and progressively earlier $Q_{\text{max}}$ timing under future scenarios that shortened the forecast horizon relative to the 1 April forecast period. The changes in the categorical skill metrics for the two basins are less pronounced than the deterministic metrics, likely because the categorical skills—which are defined on the tercile boundaries—have higher tolerance to errors than the deterministic metrics. However, the directions of change between the two tend to agree.

Overall, our study has shown that the influence of future snowpack change on streamflow predictability is highly complex and depends on the streamflow variable and magnitude of SWE change, as well as interactions with the climatic drivers. Furthermore, the changes in predictability can be highly variable across a basin, as well as over different basins. The decline in the predictive skills for the Athabasca basin suggests larger reductions in the predictability in warmer basins with large SWE loss, such as in Fraser and Columbia basins. However, the loss of SWE alone is not a sufficient condition for the change in streamflow predictability. Nevertheless, the change in future streamflow predictability will bring
new challenges for water resources managers, especially in regions/subbasins with potentially large snowpack loss.

Acknowledgments. This study was conducted with internal funding from Environment and Climate Change Canada (ECCC). The authors thank Jennifer Peskelevits (ECCC) for GIS assistance. We thank three anonymous reviewers for their comments, which led to an improved version of the manuscript.

Data availability statement. The hydrologic model projections on which this study is based are too large to publicly archive within available resources. Portions of the data and/or code used in this study will be made available upon request by contacting the corresponding author.

REFERENCES


DeChant, C. M., and H. Moradkhani, 2011: Improving the character-


