Khoshkalam, Y., Rousseau, A. N., Rahmani, F., Shen, C., & Abbasnezhadi, K. (2023). Applying transfer learning techniques to enhance the accuracy of streamflow prediction produced by long short-term memory networks with data integration. *Journal of Hydrology*, *622*, 129682.

https://doi.org/10.1016/j.jhydrol.2023.129682

Applying Transfer Learning Techniques to Enhance the Accuracy of Streamflow Prediction Produced by Long Short-term Memory Networks with Data Integration

Yegane Khoshkalam[1*], Alain N. Rousseau[1], Farshid Rahmani[2], Chaopeng Shen[2], Kian Abbasnezhadi[3]

(1) Institut National de la Recherche Scientifique - Eau Terre Environnement (INRS-ETE), Québec City, QC, Canada.

(2) Pennsylvania State University-Main Campus, Department of Civil and Environmental Engineering, University Park, PA, United States

(3) Climate Data Analysis, Science and Technology Branch, Environment and Climate Change Canada, Toronto, ON, Canada

*Corresponding author

Abstract

Recently, it has been demonstrated that the use of lagged discharge in long short-term memory (LSTM) networks represents an effective method for streamflow prediction, so-called, data integration (DI). However, it is uncertain if a transfer learning (TL) model, which did not include recent discharge when trained on the source region, can reap the benefits of including recent discharge data. This study investigated the ability of TL to provide daily streamflow predictions for a few watersheds in a snow-dominated region (target) while transferring the knowledge acquired from the conterminous United States (source) based on an LSTM architecture. The performance of the TL model was compared with that of a physically based model (PBM), HYDROTEL. Additionally, testing the source model on the target region demonstrated the performance of the model used in TL. The approaches applied to improve the accuracy of TL included use of: (i) DI of recent observed flows and simulated flows (HYDROTEL) to improve predictions, (ii) different meteorological and physiographic variables from the source and target datasets, (iii) incremental

numbers of training watersheds from the target region. Testing the source model with DI at the target region produced test-period median Kling-Gupta-efficiency (KGE) and Nash-Sutcliffe-log-model-efficiency (Nash-log) values of 0.837 and 0.870, respectively. The best performance was achieved with the TL model using DI with a maximum number of watersheds (median values of 0.953 and 0.942 for Nash-log and KGE, respectively). Moreover, including PBM simulated flows improved predictions, reducing the variability of performance metrics. Our findings show that TL and additional new procedures could significantly enhance streamflow predictions even when DI models are available.

Keywords

Streamflow, Prediction, Transfer Learning, Deep Learning, LSTM, Data Integration

1. Introduction

Climate change has been affecting water resources availability and management worldwide. In regions such as northeastern North America, where winter precipitation is dominated by snow, climate variability and global warming are expected to affect hydrologic processes. The seasonality of river flows may change with earlier spring peak flows, in addition to increased winter flows, and decreased summer flows (Barnett et al., 2005; Dibike & Coulibaly, 2005). Providing these regions with reliable streamflow predictions is of high importance.

Currently, streamflow is generally modeled using either physically based models (PBMs) or data-driven models such as deep-learning (DL) or machine learning (ML) models. The former models depend on governing equations considering complex mechanisms of the hydrologic cycle, while the latter models rely on existing relationships among data. The fact that key processes and human activities are inadequately represented in PBMs limits their use; not to mention their high computational costs (Shen et al., 2018). On the other end, DL models have shown outstanding achievements in predicting and understanding land-atmospheric interactions at low computational

cost without knowing the complex mechanisms considered in PBMs (Shen et al., 2021; Shen & Lawson, 2021).

Recently, there has been significant interest in long short-term models (LSTMs) in hydrology, which are the focus of this study. They are specially designed to process sequential data, and they have been broadly used for time series prediction of hydrologic variables (Alizadeh et al., 2021; Fang & Shen, 2020; Kratzert et al., 2018; Ouyang et al., 2021; Rahmani, Lawson, et al., 2021; Rahmani, Shen, et al., 2021). Furthermore, considering three gates in LSTM design led to significant improvements in the prediction accuracy of time series compared to traditional recurrent neural networks (RNN). Several LSTM applications have been explored to improve hydrological prediction and forecasting. Cho and Kim (2022) used the LSTM to predict the residual errors of WRF-Hydro (Weather Research and Forecasting hydrological modeling system) simulations when compared to observations. Hunt et al. (2022) developed a medium-range streamflow forecast system using LSTM in a hybrid system that was initially trained with historical ERA5 reanalyses and observations, and under operational forecasting conditions, the LSTM ingested variables from the European Centre for Medium-Range Weather Forecasts (ECMWF) Integrated Forecasting System (IFS). Girihagama et al. (2022) demonstrated that coupling an attention mechanism to an LSTM model could improve model performance during prediction using unseen non-overlapping data compared to the standard LSTM. They concluded that this attention-based LSTM model could be used with a higher level of confidence to develop real-time streamflow forecasting systems for Canadian watersheds. On the other hand, Granata et al. (2022) compared the performance of a novel model based on stacking Random Forest and Multilayer Perceptron algorithms during forecasting with that of a bi-directional LSTM model (a complicated model). Both models had equivalent forecasting capabilities; however, the stacked model performed better at predicting peak flows, but much worse at predicting low flows.

Meanwhile, Feng et al. (2020) showed that they could further improve the accuracy of streamflow predictions using lagged flow observations as inputs to an LSTM. They called the procedure of leveraging recent flow observations to improve predictions, data integration (DI). However, such exceptional performance of DL models in hydrologic sciences (here, e.g. LSTMs) comes from training these models on large datasets such as CAMELS (Addor et al., 2017) (Catchment Attributes for Large Sample Studies) to mitigate overfitting and further improve the prediction accuracy (Shin et al., 2016; Zhuang et al., 2020), as pooling data together becomes beneficial for deep networks (Fang et al., 2022). In most parts of the world, there is a lack of accessibility to such datasets in terms of quality and quantity. For example, in many parts of the world, including Québec, Canada, streamflow observations are quite sparse, with records often available only for a few years (Alipour & Kibler, 2018; Bitew & Gebremichael, 2011; Feng et al., 2021). Therefore, can the most relevant DL models reliably achieve significant performance for streamflow prediction? If not, which methods could be coupled to help them attain more accurate predictions?

TL is a common technique in artificial intelligence (AI) that mainly uses the knowledge contained in a task, i.e., the source domain, to improve the performance of a model on another similar task, i.e., the target domain where data collection is limited. Although TL was first introduced by Pratt and Jennings (1996), where a controversial argument was made based on a survey of transfer between connectionists, it did not receive growing interest until 2009 (Pan & Yang, 2009). Due to the increase in dataset size and complexity as well as the desire for faster and more efficient model training, TL has been more and more popular in recent years. Based on differences between the source and target, TL can be divided into three categories (Pan & Yang, 2010; Redko et al., 2019). It is known as transductive TL, also referred to as domain adaptation in some studies, when the source and target share the same task but have different domains. When the source task and the target task are different within the same domain, inductive TL takes place, and the learning is done via inductive bias. Finally, if the source and target have different domains and

different tasks, this type of learning is known as unsupervised TL whereby common representations are learned despite the absence of no direct similarities between the source and target. One of the key advantages is that it can drastically minimize the amount of labelled data required to train a model. This is particularly helpful in domains where labelled data are difficult to come by or expensive to acquire (Blitzer et al., 2007). TL has been widely used to transfer models from image classification and recognition datasets to similar tasks, such as in remote sensing (Guo et al., 2020; Liu et al., 2022; Nowakowski et al., 2021). However, only some studies have focused on performing TL, particularly to predict the hydrological time series (Bedi, 2022; Kimura et al., 2019). This further underscores the importance of performing TL in hydrological modeling. For example, Ma et al. (2021) attempted to transfer a pretrained LSTM with CAMELS dataset across continents to three regions with different densities of data, namely, Britain, central Chile, and China. They showed TL could substantially improve streamflow prediction in the data-scarce areas of Chile and China, but it could not elevate the results for Britain. The benefits gradually lessened if the target region had long-term records. Despite wide uses of TL, it also has some limitations. Indeed, one challenge remains, the source domain and the target domain must be sufficiently similar. Also, the pre-trained model and TL technique being used can have a big impact on the performance (Zhao et al., 2021), thus it behooves to carefully experiment to identify the best strategy for a task to avoid negative transfer (Pan & Yang, 2010; Torrey & Shavlik, 2010).

It is undeniable that previous studies have enhanced our knowledge of employing TL in hydrological modelling. In streamflow prediction particularly, there are still a lot of unanswered questions surrounding this subject. Here, we applied TL using LSTM as the modeling platform pretrained over the CAMELS dataset, for streamflow predictions of eight watersheds in Southern Québec, Canada. This study answers several research questions: (i) How does the TL model compare against an operational physically based streamflow model? (ii) Can TL improve the

performance of DI models? (iii) Can PBM simulations provide value as inputs for a TL model? (iv) How does the seasonality performance of TL compare to that of a PBM?

2. Data and Methods:

2.1. Datasets

This study used two types of datasets to transfer knowledge from highly monitored regions in the USA (source dataset) to a moderately monitored region in Southern Québec, Canada (target dataset). CAMELS was selected as the source dataset (Addor et al., 2017). This dataset is freely accessible and provides data for 671 watersheds with minimum human disturbances across CONUS (conterminous United States) (shown in Figure 1). CAMELS has various sub datasets covering meteorological or forcing data, observed flows, and physiographic data or static attributes. All the data are aggregated and reported at the watershed scale. The period for most of the watersheds goes back to 1980. Daymet (Thornton et al., 2012), which had the highest spatial resolution (1 km) among the three gridded data sources, was used to generate the meteorological data. All watersheds in CAMELS (671) were selected to train the source model (training period from January 1, 1989, to January 1, 2006, and testing period from January 1, 2006, to October 1, 2013).

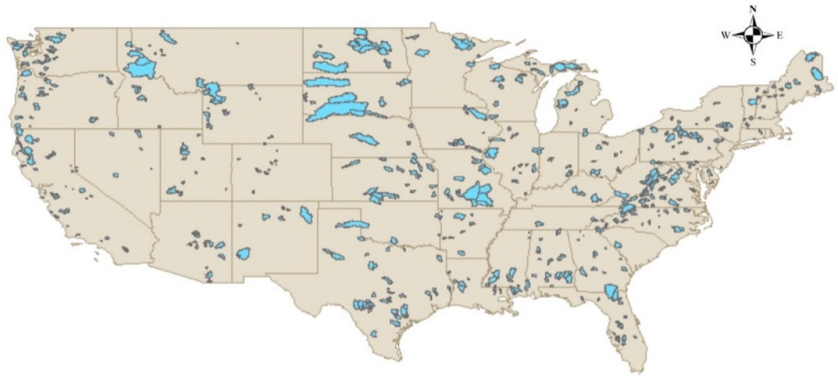


Figure 1: Location of 671 CAMELS catchments.

Southern Québec was selected to conduct the TL with just a few watersheds, eight (see Table 1, along with their corresponding static attributes data). These watersheds are located over five hydrographic regions of the St. Lawrence River (see Figure 2). Their corresponding drainage areas vary from medium watersheds (591 km^2) to large ones (6768 km^2). Foulon and Rousseau (2018) reported that the region surrounding the St. Lawrence River depicted in Figure 2 has a mean annual temperature of 1.8 ºC and a mean annual total precipitation of 940 mm based on the meteorological data (1981–2002, 1995 and 1996 being unavailable) from Natural Resources Canada. All watersheds are snow-dominated, while the annual peak flow generally occurs during the spring freshet. Table 2 and Table 3 (Foulon & Rousseau, 2018) provide a summary of the hydroclimatic characteristics for the two hydrologic seasons of summer (1 June to 30 November) and winter (1 December to 31 May). The mean winter rainfall is more varied, with a mean of 214 mm and a standard deviation of 64 mm, meanwhile, the mean summer rainfall is 548 mm and more uniform among the watersheds (standard deviation of 32 mm). The mean snowfall is 273 mm, while the standard deviation of the mean snowfall is 55 mm. In the summer and winter, the mean temperature is 11.2 ºC and -4.2 ºC, respectively. In accordance with the seasons, the mean temperature during summer and winter has a standard deviation of 1.7 ºC and 2.6 ºC, respectively. The mean daily flows are 1.2 and 1.9 mm during summer and winter, with standard deviations of 0.4 and 0.23, respectively. Summer rainfalls and convective storms, which are more variable than snowfalls, are responsible for the higher variability of summer flows.

161 *Table 1: Static attribute data of the eight study watersheds in Southern Québec*

Hydrographic regions	Watershed	Average slope [m/km]	Average elevation [m]	Drainage area [km^2]	Forest fraction [%][1]	Dominant land cover	Fraction of dominant land cover [%][2]
Saint-Laurent northwest	Batiscan	4.7	379.0	4505	94.2	Deciduous Broadleaf Forest	52.3
Saint-Laurent southeast	Chaudière	2.5	392.0	5820	94.9	Deciduous Broadleaf Forest	73.4
	Bécancour	4.9	368.0	914	69.2	Mixed Forests	69.2
Saguenay and Lake Saint Jean	Chamouchouane (sub-watershed)	0.7	160.0	591	92.5	Evergreen needleleaf Forest	48.1
Saint-Laurent southwest	Châteauguay	1.1	193.0	2492	82.5	Deciduous Broadleaf Forest	77.4
	Yamaska (sub-watershed)	1.6	132.0	1505	39.8	Mixed Forests	39.8
Outaouais and Montreal	Rouge	4.2	384.0	5480	94.8	Deciduous Broadleaf Forest	69.2
	Gatineau	4.33	443	6768	71	Evergreen needleleaf Forest	40

162

[1] Where forest fraction refers to the summation of three land covers, namely, Mixed Forests, Deciduous Broadleaf Forest, and Evergreen needleleaf Forest.

[2] Where fraction of dominant land cover refers to the land cover with the highest fraction among all the land covers of the watershed of interest.

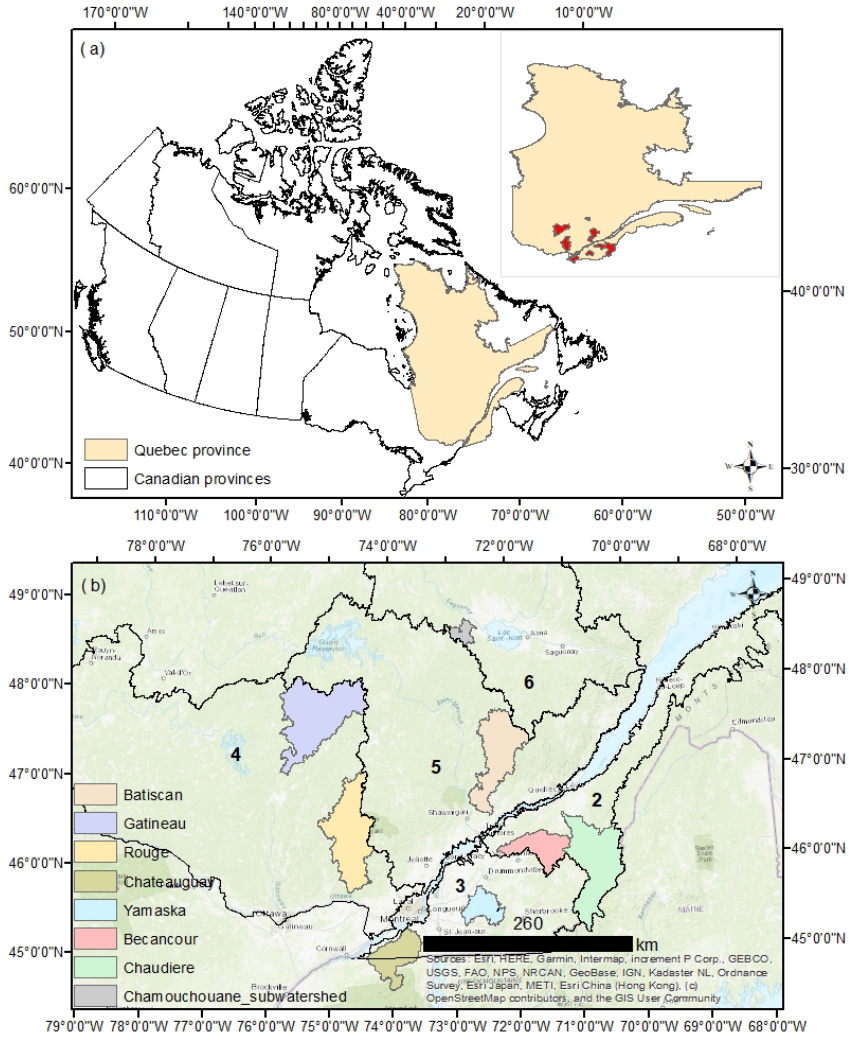


Figure 2: Location of the study watersheds, Southern Québec, Canada. The horizontal and vertical axes correspond to longitude and latitude, respectively. Numbers on the lower map indicate the five hydrographic regions of Québec associated with the study watersheds.

Table 2:Climate characteristics of the study watersheds (1982-2002).

Watershed	Rain(mm)						Snow(mm)			Mean temperature (ºC)					
	Summer			Winter			Annual			Summer			Winter		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
Batiscan	337	558	645	97	180	403	286	356	416	8.4	10.3	11.5	-7.7	-5.4	-3.1
Bécancour	392	585	809	129	260	490	169	260	372	10.3	11.9	13.3	-5	-2.9	-1
Chamouchouane	293	518	690	85	131	248	219	290	383	6.4	8.6	10	-11.4	-8.7	-5.6
Châteauguay	402	512	620	174	269	429	137	193	252	12.4	13.8	15.2	-3.8	-1.1	0.9
Chaudière	421	590	794	179	253	392	216	266	316	9.5	11.2	12.5	-5.4	-3.3	-1.5
Rouge	372	529	613	100	175	333	248	327	368	9.2	10.8	11.9	-6.9	-4.4	-2
Yamaska	476	577	743	180	305	526	122	204	294	11.6	13.1	14.5	-3.9	-1.5	0.6
Gatineau	324	519	671	86	145	242	224	290	350	7.9	9.7	11.4	-8.8	-6.4	-3.6

Table 3: Summary of daily flows of the study watersheds (1982-2002)

Watershed	Q(mm/day)					
	Summer			Winter		
	Min	Mean	Max	Min	Mean	Max
Batiscan	0.9	1.6	2.4	1.4	2.1	3.1
Becancour	0.6	1.1	1.9	1.5	2.1	2.9
Chamouchouane	1.1	1.8	2.6	1.2	1.5	1.9
Chateaguay	0.3	0.6	1.3	1.1	1.8	2.6
Chaudiere	0.5	1.1	2.1	1.6	2.2	3
Rouge	0.8	1.2	1.7	1.3	2	2.9
Yamaska	0.4	0.8	1.7	1.2	2	2.7
Gatineau	1	1.5	2.4	1.1	1.7	2.5

Most of the required variables of the target dataset came from HYSETS (hydrometeorological Sandbox-École de Technologie Supérieure) (Arsenault et al., 2020), a gathering of various datasets. Among the available meteorological datasets, the daily ERA5 reanalysis data (Hersbach et al., 2020) was chosen as the primary source of forcing data for the target dataset. The reliability of this dataset over the target domain has been demonstrated by several studies (Tarek et al., 2020a, 2020b). The training period ranged from December 1, 1982, to December 1, 1991, while the testing period was from December 1, 1991, to December 1, 2002. The required input variables for near-optimum daily streamflow prediction using an LSTM model have been reported previously by Feng et al. (2020) (Table S1). Considering that some of the input variables may not be accessible for Southern Québec, the common input variables are introduced in Table 4 along with their corresponding dataset from which they were extracted. Another model used along the HYSETS database to extract some physiographic data was that produced using PHYSITEL, a semi-automatic GIS-based framework (Noël et al., 2014; Turcotte et al., 2001). PHYSITEL allows easy watershed segmentation and parametrization of relatively homogeneous hydrological

units (RHHUs or hillslopes) and interconnected river segments (RSs). To extract physiographic data (e.g., elevation), the value of the variable of interest at the centroid of each RHHU was multiplied by the area of the RHHU. After the summation of all values, the resulting product was divided by the area of the whole watershed. The hydrometric data were provided by the Québec Ministry of the Environment and the Fight against Climate Change (MELCC, 2022).

Table 4: Summary of the common forcing and static attributes data used from source and target dataset.

	Variable name	Description	Data source	Unit
Forcing	PRCP	Precipitation	ERA5(HYSETS)	mm/day
	Tmax	Minimum temperature	ERA5(HYSETS)	ºC
	Tmin	Maximum temperature	ERA5(HYSETS)	ºC
Attributes	Elev_mean	Mean elevation at watershed scale	PHYSITEL	m
	Slope_mean	The mean slope at the watershed scale	PHYSITEL	m/km
	Area_gages	Drainage area	HYSETS	km^2
	Frac_forest	Forest fraction	PHYSITEL	-
	Dom_land_cover_frac	Fraction of the dominant land cover class	PHYSITEL	-
	Dom_land_cover	Dominant land cover type	PHYSITEL	-
	Soil_porosity	Volumetric soil porosity	PHYSITEL	-
	Soil_conductivity	Saturated hydraulic conductivity	PHYSITEL	cm/hr
	Geol_class_1st	Most common geologic class in the watershed	GLiM(HYSETS)	-
	Geol_class_2nd	Second most common geologic class in the watershed	GliM(HYSTES)	-
	Geol_porosity	Subsurface porosity	GLHYMPS(HYSTES)	-
	Geol_permeability	Subsurface permeability	GLHYMPS(HYSTES)	m^2

2.2. Long Short-Term memory (LSTM) architecture

The modelling framework in this study was long short-term memory networks (LSTM). Recurrent Neural Networks (RNN), in contrast to typical Artificial Neural Networks (ANN), can strongly connect data from previous time series with that from a current time series. However, RNNs are unable to preserve the long-term dependencies in the input sequences when the distance

between the relevant information and the point where it is required increases. To solve the vanishing or exploding gradient issues in the RNN computational process, LSTM was developed.

LSTM has four functions to operate with three of them acting as governing gates (Hochreiter & Schmidhuber, 1997) controlling the information flow through cells. Additionally, another cell state so-called c_t has been added in the LSTM structure to save the learned information; then the stored information is transferred to cell h_t. The structure of the four functions is illustrated in Figure 3. The cell state, located in the upper left corner of the diagram illustrates the transition of c_{t-1} to c_t and represents the core of LSTM. Meanwhile, a vector of zeros of a user-defined length is initialized to represent the hidden state *h*.

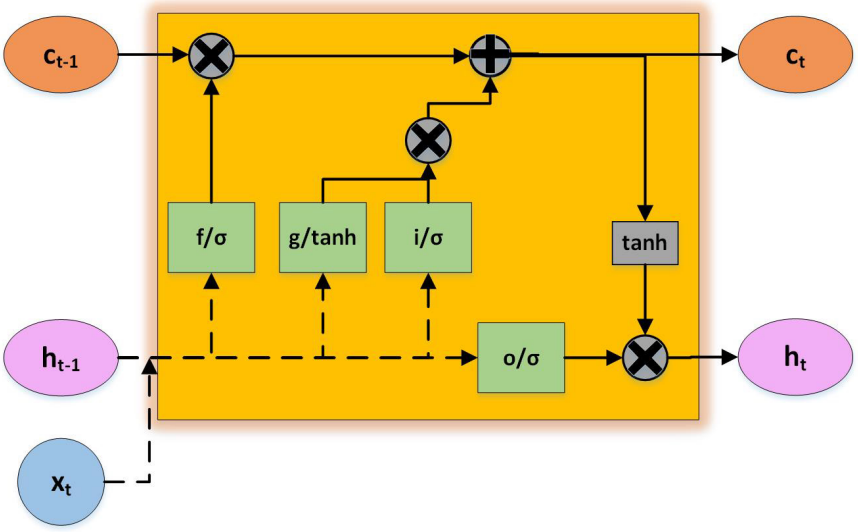


Figure 3: The illustration of a Long Short-term Memory (LSTM) unit.

Then, LSTM uses a sigmoid function (*f*) acting as a decision gate in the cell that decides which information to be forgotten and to what degree. This gate which is called the forget gate was initially introduced by (Gers et al., 2000). The f_t function is introduced in Eq (1):

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \qquad \text{Eq (1)}$$

Where f_t is the resulting vector ranging between 0 and 1, while σ is the sigmoid function ($\sigma(x) = \frac{1}{1+e^{-x}}$). The two adjustable weight matrices and the bias vector are W_f, U_f, and b_f, respectively.

Then, two additional functions are used for deciding the information contained in the cell state. More specifically, function g_t hyperbolic tangent function ($\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$) creates a vector of new candidate values for the cell state in the range (-1, 1), function *i*, another sigmoid function in the range (0, 1) assigns what information can be added to the cell state and plays the role of input gate. Finally, these two functions are combined with the forget gate information (*f* function) to update the previous step of cell state c_{t-1} to the current c_t as follows:

$$g_t = \tanh(W_g x_t + U_g h_{t-1} + b_g) \qquad \text{Eq (2)}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \qquad \text{Eq (3)}$$

$$c_t = f_t \times c_{t-1} + i_t \times g_t \qquad \text{Eq (4)}$$

Where the × operator implies element-wise multiplication and ensures the information will be preserved in LSTM (both f_t and i_t range between 0 and 1, in addition to assigning values of 0 for the forget gate and values of 1 for i_t for keeping the valuable information). The final gate is the output gate or *o* function responsible for which information should be transferred out of the cell state and updates the hidden state using a hyperbolic tangent function.

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \qquad \text{Eq (5)}$$

$$h_t = o_t \times \tanh c_t \qquad \text{Eq (6)}$$

Thus, the long-term dependencies can be learned through these equations in LSTM in a way that, first, no information can get lost, because the cell state is defined as a linear function, unlike RNNs. On top of that, the introduction of the new three gates in LSTM allows this neural network to preserve the information for long time steps. This feature of LSTM can prevent exploding and vanishing of the back-propagation step (Hochreiter & Schmidhuber, 1997) which can occur in

RNNs. The output of the final LSTM layer is connected to a dense layer at the last time step (h_n) as follows:

$$y = W_d h_n + b_d \qquad \text{Eq (7)}$$

Where y is the final predicted value of streamflow and W_d and b_d are the weight matrix of the dense layer and the bias term, respectively. It should be mentioned that applying these structures would significantly increase the number of adjustable parameters for the learning process, which can result in some degree of overfitting for cases such as for small datasets. The LSTM model used for pretraining using all observations of 671 watersheds has shown very good performance for snow-dominated regions in the USA (Feng et al., 2020) although the CAMELS dataset includes information from various regions.

2.3. Transfer Learning

Transfer learning was used to predict streamflow for a few watersheds in Southern Québec. During the training process of deep networks such as LSTM, some weights are assigned to hidden layers in addition to nonlinear activation functions. The information during training is saved through weights, with some parts of the network self-organizing to perform specific functionalities as dictated by the architecture of the LSTM.

TL was performed in two steps. Step one, the source model was pretrained on the source dataset, also known as weight initialization. Step two, after the pretraining, the stored model along with the initialized weights was (entirely or partially) trained once more on the target dataset. The weights could be modified during training on the target dataset (weight initialization) or kept constant (weight freezing). In regions with limited access to data, without performing TL, the model cannot find the relationships that exist among the inputs (e.g., forcing, and static attributes) and the target (here, observed flows). Final training modified and fine-tuned all or parts of the weights to better

fit the model over the target dataset while the model converged faster. The advantage of weight freezing over pretraining is the reduction of the size of the training dataset required for calibration and the possibility of overfitting (Ma et al., 2021). The TL process is illustrated in Figure 4. TL model was constructed based on training on all watersheds in Southern Québec (in contrast to watershed-by-watershed modeling).

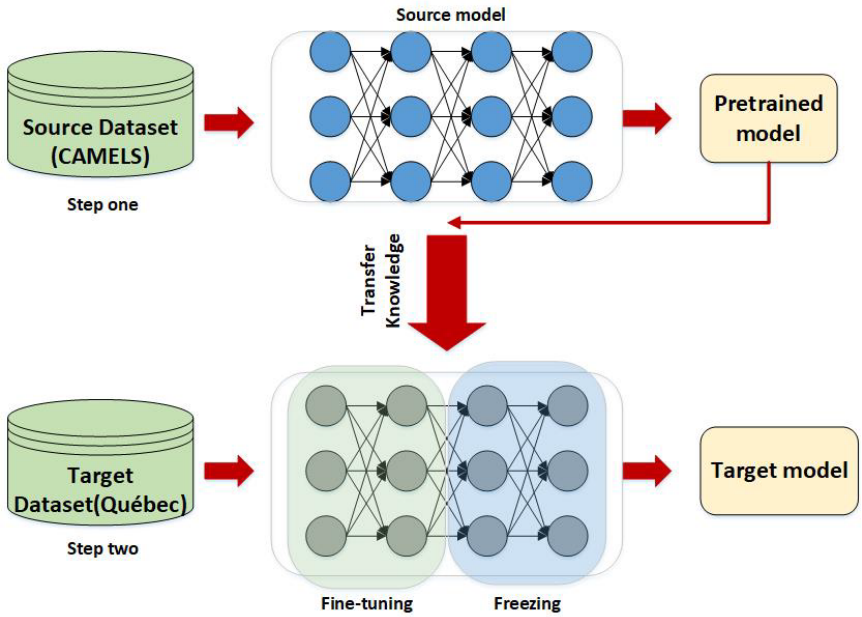


Figure 4: Structure of the transfer learning (TL) process.

2.4. Physically based model

The PBM chosen for this study is HYDROTEL, a semi-distributed hydrological model compatible with remote sensing and GIS data (Bouda et al., 2014; Bouda et al., 2012; Fortin et al., 2001; Fossey et al., 2015; Turcotte et al., 2007). The model was used to provide simulated flows to the LSTM to develop a physics-informed TL model and a benchmark as well. The Québec Ministry of the Environment and the Fight against Climate Change has been applying HYDROTEL for hydrological forecasting (Turcotte et al., 2004) and hydroclimatic projections (DEFACC, 2018). The hydrological processes modeled by HYDROTEL include spatial distribution of precipitation and temperature, snow accumulation and melt, soil temperature/freezing, vertical water balance

(i.e., infiltration, evapotranspiration, soil water movement), wetland hydrology, overland flow routing, and river routing. Hydrological modeling through HYDROTEL is based on dividing the watershed into hillslopes (a.k.a, relatively homogeneous hydrological units, RHHU), subdivided into three soil layers, and interconnected river segments (RSs) draining the units. The first soil layer is considered a thin layer. Interflow and baseflow are controlled through the second and third layers, respectively. The water contents of the three soil layers are calculated as state variables at each time step (either daily or sub-daily). Readers are referred to (Fortin et al., 2001), (Turcotte et al., 2007), and (Fossey et al., 2015) for more details on HYDROTEL. The model was automatically calibrated using a global optimization algorithm, DDS (see Tolson and Shoemaker (2007) and Foulon and Rousseau (2018) for more details) and a combination of the Kling-Gupta efficiency criterion (Kling et al., 2012) and the Nash-Sutcliffe efficiency criterion (Nash & Sutcliffe, 1970). The latter was calculated using log-transformed flows (Oudin et al., 2006). This approach can be beneficial for predicting both high and low flows (Foulon & Rousseau, 2018). Notably, LSTM and HYDROTEL used the data time series differently during training. While LSTM was trained on all the available data, observed flows were excluded from November to March (under ice cover) during the calibration of HYDROTEL and a one-year spin-up period was considered.

2.5. Data pre-processing

Data pre-processing and normalization are crucial tasks for ML and DL studies because the approach used can influence how accurate predictions are made as well as how long training takes (Nawi et al., 2013). Each problem requires specific methods and algorithms for data pre-processing which depends on the types of variables. Common data pre-processing includes data cleaning, dealing with missing data, data transformation, which changes the scale, type, and probability distributions of variables, and applying techniques such as feature selection and dimensionality reduction to reduce the number of input variables (Maharana et al., 2022). The goal is to obtain a mean close to zero for data distribution to speed up the learning process and

the convergence. According to Feng et al. (2020), if the watersheds are batched together during streamflow modeling, the lack of data pre-processing would result in considering wetter watersheds more in the calculation of default loss, while with data normalization, the model errors are identically distributed among watersheds within the same batch. Normalization of inputs and output was done according to the data pre-processing suggested by Feng et al. (2020). Normalization of the streamflow variable was marginally different from other variables. Dimensionless streamflow was obtained by dividing the streamflow by the watershed area and mean annual precipitation. This way, The LSTM model can better realize the relationships between the dimensionless target (streamflow) and input variables. Then, another transformation was applied to achieve a Gaussian distribution as follows:

$$v* = log_{10}(\sqrt{v} + 0.1) \quad \text{Eq (8)}$$

where v^* and v are the streamflow variables after and before transformation, respectively. Then, a final transformation was applied to all the variables including the forcing data, streamflow, and static attributes to normalize them as follows:

$$x_{i,new} = \frac{x_i - \bar{x}}{\sigma} \quad \text{Eq (9)}$$

where, $x_{i,new}$, x_i, $\bar{x}$, and σ are the standardized, raw, mean, and standard deviation values for the variable, respectively. The results, shown in Section 3, were converted to the original scale before standardization.

2.6. Numerical experiments

2.6.1 Testing the source model on target watersheds

To understand the performance of the source model on the target region, the model trained using all CONUS basins, was tested on watersheds of Southern Québec in three ways: (i) the forward

model with forcing and static attributes given as inputs during training and testing (the term forward is used to ease the comparison with DI models); (ii) the DI model trained on CONUS and tested on target watersheds (the inputs were forcing data, static attributes, and observed streamflows from the previous day which for simplicity is called Q_{t-1} afterwards); and (iii) finally, the DI model trained on CONUS, while during testing, the observed streamflows (prediction target) were replaced with HYDROTEL simulated flows, and then used to replace Q_{t-1}. The actual observed streamflows were used to evaluate the performance of this specific case during the calculation of evaluation metrics. More details about the process of testing the source model on the target region can be found in Figure S2. The training period over CAMELS was from January 1, 1989, to January 1, 2006, while the testing period of the target dataset was from December 1, 1991, to December 1, 2002, so the results can be compared with the results introduced in Section 2.6.2.

2.6.2 Improving TL approaches

The baseline TL model that was compared with HYDROTEL simulated flows and observed flows, was the forward model incorporating only forcing and static attribute data (Results are shown in Section 3.2). The following approaches were proposed to train and improve the prediction accuracy of the baseline TL model against HYDROTEL:

(i) The TL model was pretrained on all available forcing and static attribute data in CAMELS (input variables of source dataset introduced in Table S1). In contrast, the input data used during training on the target region was more limited than CAMELS (input variables introduced in Table 4). For simplicity, performing TL while different forcing and static attributes variables are used from source and target datasets is called different input data (DD) as opposed to performing TL while the same forcing and static attributes variables are used from source and target datasets, namely, same input data (SD). Figure 5 depicts how forcing and static attribute variables differ between the source and target datasets.

The variables used for model training in TL with SD are displayed as the input layer (target) in Figure 5 (apart from *Qsim*, which represents HYDROTEL simulated flows and is not included in the source dataset). Yet there remains the question of whether to use DD or SD to do TL modeling for the other approaches of this study. Values of 1 and 0 in the names of TL models (see Table 6) indicate performing TL with DD and SD, respectively.

(ii) The TL model was trained with Q_{t-1} as well as forcing and static attribute data (called data integration, DI). In the name of the models, DI appears before TL to distinguish the DI models from the forward ones. All DI-TL models were developed and used while DI was applied during pretraining and training on the source and target datasets, respectively. There was a particular model whereby DI was applied only in training on the target region and not during pretraining to show if TL can improve the performance of DI models. Thus, the name of this model is indicated as TL-DI-1(8) (see Table 6).

(iii) The TL model was trained with simulated flows of HYDROTEL along with forcing and static attribute data. It has been shown that HYDROTEL simulated flows could provide satisfactory predictions for the selected watersheds of Southern Québec (Foulon & Rousseau, 2018). Two TL models of DI and forward were constructed in this part. In the names of these models, S denotes the use of simulated flows in model training.

(iv) The TL model that had the number of training watersheds increased from 8 to 34 in Southern Québec (locations of the 34 watersheds are displayed in Figure S1). In the name of this series of TL models, the number of watersheds involved in the training distinguishes this approach from the rest.

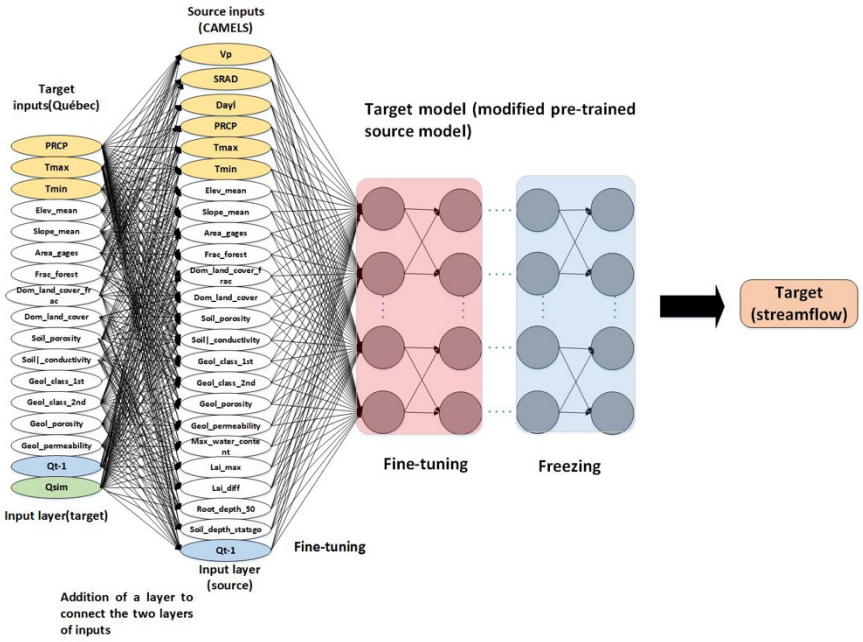


Figure 5: Schematic of performing TL while different forcing and static attribute variables from source and target datasets were used in model training. Table 2 and Table S1 contain detailed descriptions of the variables presented in the target input layer and source input layer. The input variables for forcing, static attribute variables, recent discharge observations, and simulated discharge are represented, respectively, by the yellow, white, blue, and green cells in both target and source input layers.

To complement the comparison, local models (referred to the LSTM models trained on the target region without performing TL) were trained for each approach to understand the benefit of using TL compared to local model training. More detailed descriptions of all models (TL, and local models) and the flowchart of the model building process are provided in Figures S3 and S4.

2.7. Model evaluation criteria and hyper-parameters

2.7.1. Performance metrics

In the current study, the following metrics were deployed to assess model performance.

- Kling-Gupta efficiency (KGE) (Kling et al., 2012): it breaks up Nash-Sutcliffe efficiency (NSE) into its constitutive parts (correlation, variability bias, and mean bias). Like NSE, KGE ranges from -∞ and 1 while KGE of 1 shows perfect agreement between simulated and observed flows. It has been used for model calibration and validation since it

addresses some of the shortcomings of using only NSE as the performance metric and is given by:

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad (10)$$

Where $\beta = \frac{\overline{S}}{\overline{O}}$ and $\gamma = \frac{\sigma_s/\overline{S}}{\sigma_o/\overline{O}}$ and r = Pearson Correlation Coefficient

- Log Nash-Sutcliffe efficiency (Nash-log) (Oudin et al., 2006): Nash-log is the NSE (Nash & Sutcliffe, 1970) calculated on log-transformed flows. It has greater efficiency in terms of favoring both high and low flows during streamflow simulation. Nash-log ranges from -∞ to 1 and a value of 1 shows the optimal value of Nash-log. It is calculated as follows on log-transformed flows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(O_i - S_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \quad (11)$$

- Root Mean Squared Error (RMSE) (Singh et al., 2005): It measures the spread of error residuals and it ranges from 0 to ∞ while RMSE of 0 shows no residual errors. It is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(S_i - O_i)^2}{n}} \quad (12)$$

- Pearson Correlation Coefficient (Corr) (Pearson, 1920): Pearson Correlation is a measure of linearity between two variables, for this case study between simulated and observed flows. A value between -1 and 1 is assigned to Corr where -1 shows total negative correlation, 1 shows total positive correlation, and 0 illustrates no correlation among the variables. Corr or r is the root square of r^2 that according to Krause et al. (2005) explains

the degree of dispersion of observed values by predicted values. If a model always results in underestimation or overestimation, it could still have a high value for r^2 and as a result, in Corr, even if all the predicted values by that model are wrong. Thus, it is suggested to use additional evaluation metrics along with using Corr to interpret the modeled values. The equation to calculate Corr is given by:

$$Corr = \frac{\sum_{i=1}^{n}(O_i - \overline{O})(S_i - \overline{S})}{\sqrt{\sum_{i=1}^{n}(O_i - \overline{O})^2 \sum_{i=1}^{n}(S_i - \overline{S})^2}} \quad (13)$$

- Percent Bias (PBIAS) (Yapo et al., 1996): It measures how much the average of simulated values are inclined to be higher or lower than those observed. It ranges between -∞ and ∞ while 0 shows the optimal value. It is given by:

$$PBIAS = \frac{\sum_{i=1}^{n}(S_i - O_i)}{\sum_{i=1}^{n} O_i} \times 100 \quad (14)$$

Where in all the above equations, $\overline{O}$ and $\overline{S}$ refer to the means of observed *(O)* and simulated (*S*) flows, respectively. σ_s and σ_o stand for standard deviations of simulated and observed flows, respectively, where *n* is the number of observed flows. All metrics were calculated for the testing period. For sake of comparison, the period from January 1, 1995, to May 31, 1997, was excluded due to the absence of simulated flows of HYDROTEL in this period.

2.7.2 Hyperparameters

The parameters defining the architecture of DL models are termed hyperparameters. Finding their optimal values of represents a critical task as it directly affects the accuracy and reliability of the learning models. A proper selection of hyperparameters can distinguish a state-of-the-art performance from an average performance (Hutter et al., 2015). The RMSE between predicted and observed streamflows was selected as the objective function to be minimized, or loss function, and the Adadelta algorithm was adopted as the optimization method (Zeiler, 2012). In comparison to pure stochastic gradient descent, the Adadelta algorithm has a very low

computational overhead and dynamically self-adjusts over time using only first order information. The method appears to be robust to noisy gradient information, various model architectural choices, diverse data modalities, and selected hyperparameters while not requiring manual optimization of the learning rate. The optimized parameters for the models used in this study were batch size, hidden layer size, epoch, parameters of dropout algorithm, and length of training instances. Batch size controls the number of training examples given to the model for calculating the loss function before updating the matrices of weights. Hidden layer size determines the number of layers between input and output layers which also characterizes the degree of complexity of the network. One epoch is defined as one forward run for all training examples. The length of training instances determines the length of the time series for each training sample. Dropout, which randomly drops network nodes along with all the other connections, is a regularization technique specifically for neural networks. The degree of regularization is controlled through a parameter called dropout rate. The dropout rate is the proportion of units that are dropped out during training. For instance, a dropout rate of 0.5 means that half of the network units are randomly dropped out during each training iteration. The source LSTM model and the TL models have their hyperparameter combinations manually tested and selected. Table 5 contains information about the tested and optimal values of hyperparameters. The length of training instances was the same for both models (365 days). The snow and subsurface storage processes can be represented by training instances of 365 days. However, the length of the selected training instance does not restrict the LSTM memory to 365 days.

Table 5: Hyperparameters Description

		Number of epochs	Batch size	Hidden layer size	Length of training instances	Dropout rate
Pretraining	Tested	100,200, 300, 500	100, 200, 671	64, 128, 256	100, 200,365	0, 0.2, 0.5
	Selected	300	671	256	365	0.5
Training	Tested	100, 200, 300	8	64, 128, 256	365	0.5
	Selected	300	8	256	365	0.5

3. Results

This section is organized as follows. First, the results of testing the source model on target watersheds are presented. Second, the overall performance of the baseline TL model is discussed. Third, the added value of approaches to boost the performance of the baseline TL model are introduced. And fourth, the performance of HYDROTEL and the best TL model for different seasons are presented. Table 6 shows the performance of all models presented in this study.

Table 6: Summary of the results achieved by all the models

Model	Nash-log $_{median}$	Nash-log $_{mean}$	KGE_{median}	KGE_{mean}
HYDROTEL	0.679	0.642	0.845	0.842
Local[1] (8[2])	0.685	0.670	0.750	0.731
TL-0[3](8)	0.762	0.751	0.839	0.813
TL-1[4](8)	0.745	0.724	0.787	0.753
TL-S[5]-1(8)	0.822	0.823	0.840	0.827
TL-0(34)	0.765	0.765	0.851	0.815
DI[6]-Local (8)	0.927	0.905	0.894	0.887
DI-TL-0(8)	0.923	0.922	0.904	0.906
DI-TL-1(8)	0.933	0.928	0.949	0.936
DI-TL-S-1(8)	0.806	0.803	0.838	0.830
DI-TL-0(34)	0.953	0.953	0.942	0.927
TL-DI-1(8)	0.883	0.872	0.887	0.874
Local-S (8)	0.795	0.778	0.768	0.767
DI	0.870	0.774	0.837	0.781
DI-S	0.570	0.526	0.745	0.704

[1] Where local model refers to training the source model with the target dataset.
[2] Where the number in the parentheses indicates the number of training watersheds.
[3] Where 0 indicates: performing TL while the same forcing and static attribute variables from source and target datasets are used in model training (SD).
[4] Where 1 indicates: performing TL while different forcing and static attribute variables from source and target datasets are used in model training (DD).
[5] Where S refers to the use of simulated flows in the model training.
[6] Where DI refers to the use of Q_{t-1}(recent observed flows) in training.

3.1 Performance of source model tested on target watersheds

Testing the forward model on target watersheds achieved a negative value for Nash-log median test period. Thus, further assessment was not pursued. Figure 6 shows the results of testing the DI model on target watersheds. DI model using observed flows could provide satisfactory results during testing (KGE and Nash-log values of 0.837 and 0.870, respectively) on the target region. Testing the DI model on the target watersheds with limited access to observation records (shown as DI-S in Figure 6) could achieve decent median KGE and Nash-log values of 0.745 and 0.569. Moriasi et al. (2007) indicated that Nash-log values larger than 0.5 could provide a good fit of the hydrograph and a satisfactory fit with the 1:1 line on the plot of observed versus simulated flows.

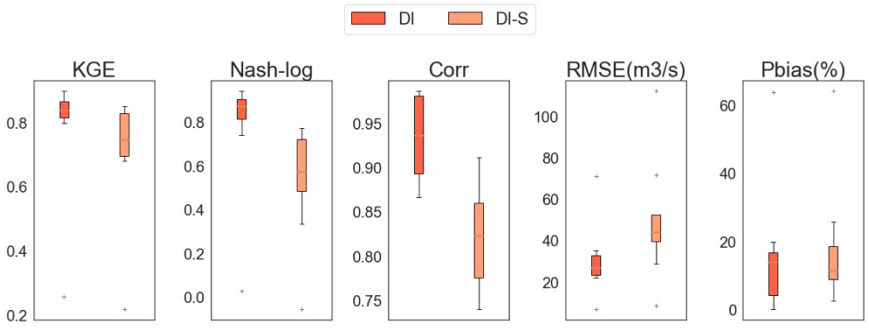


Figure 6: Results from testing the source model with DI applied on the target watersheds when trained on all data available in CAMELS (671 watersheds with training period between 1989 and 2006). Each boxplot contains the metric for eight target watersheds used during testing (between 1991 and 2002). DI-S shows the results of testing the source model trained on the CAMELS dataset while DI has been applied. In this case, Qt and Qt-1 were replaced with HYDROTEL simulated flows during testing on the target region. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker delineate 5, 25, 50, 75, and 95 percentiles, respectively. Outliers are shown by cross markers.

3.2 Performance of baseline TL model

HYDROTEL and TL performed better than the local model during the test period (Figure 7). HYDROTEL and TL had similar median KGE values (0.845 and 0.839, respectively). TL offered solid performance for predicting baseflows (as quantified by Nash-log) compared to HYDROTEL, with a higher median and smaller variabilities in predicted values (median values of 0.762 and 0.679 for TL and HYDROTEL, respectively). Even the local model performed slightly better than HYDROTEL in capturing the baseflows (median value of 0.685). The median Corr of the test

period was slightly larger for TL compared to HYDROTEL (0.884 and 0.871, respectively), while the RMSE of TL was relatively smaller than that of HYDROTEL (37.515 m^3/s and 39.360 m^3/s, respectively). However, the PBIAS values of HYDROTEL were better than those of TL, with a closer median value to zero (value of -1.007 %), while the median PBIAS of TL was -7.842 %.

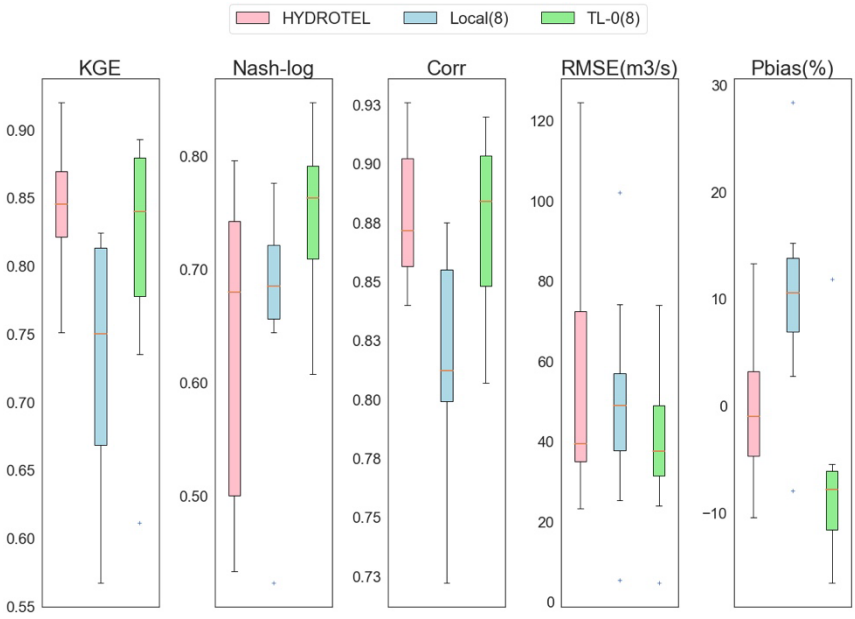


Figure 7: Evaluation metrics of baseline TL model versus HYDROTEL and the local model during the test period for eight watersheds in Southern Québec. TL-0(8), and Local(8) represent the baseline TL (forward) model and the local model while the number of training watersheds is eight (shown by 8 in the legend) and the same forcing and static attribute variables from source and target datasets were used during TL (indicated by 0). The lower whisker, lower box edge, center bar, upper box edge, and upper whisker delineate 5, 25, 50, 75, and 95 percentiles, respectively. Outliers are shown by cross markers. Each boxplot contains the metric for the eight (8) watersheds from the target region trained on all data available between 1982 and 1991. For HYDROTEL, the first available year for each watershed was used as a spin-up period.

Figure 8 shows the performance of the above models based on their hydrographs for one of the watersheds, the Batiscan watershed along with the 10^{th} and 90^{th} percentiles of observed data and the observed mean annual hydrographs (For the selected watersheds of Southern Québec, the general shapes of the mean annual hydrograph are almost similar. Hence, the hydrograph of only one of them is presented. Hydrographs of the remaining watersheds can be found in the supporting material). TL and HYDROTEL (cornflower blue and green curves in Figure 8, respectively) showed good agreement with the observed flows and are well within the 10^{th} to 90^{th} percentile band of the observed flows (shaded area in Figure 8). In some places, the local model

outreached the 10th to 90th percentile band. While HYDROTEL underestimated the peak flows, the TL model was inclined to overestimate the occurrence of peak flows in spring.

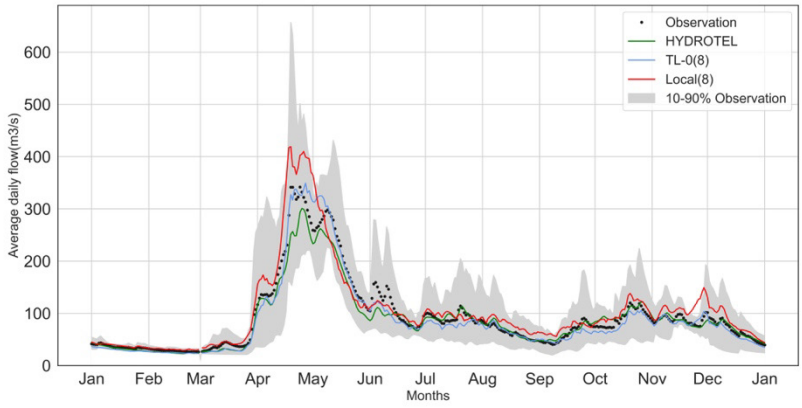


Figure 8: Comparison of simulated (baseline TL model, HYDROTEL, and the local model) and observed mean annual hydrographs for the Batiscan watershed (8-year mean) for the test period (12/1991-12/2002). TL-0(8), and Local (8) represent the baseline TL (forward) model and the local model while the number of training watersheds is eight (shown by 8 in the legend) and the same forcing and static attribute variables from source and target datasets were used during TL (indicated by 0). The lower and upper limits of the shaded area represent the 10th and 90th percentiles of observed data.

3.3 Added value of applying different approaches to improve the baseline TL model

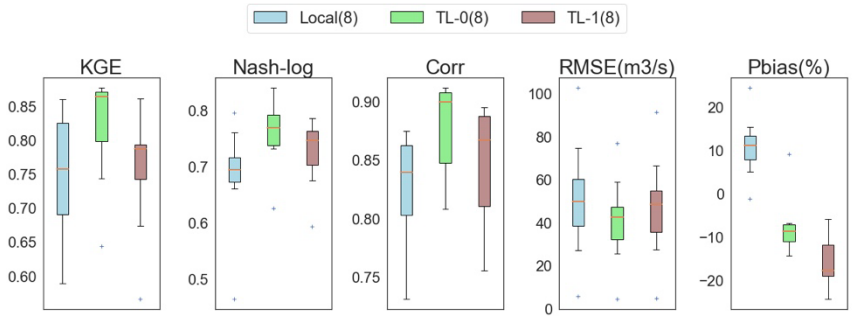
This section introduces the results of applying different approaches to improve the efficiency of the baseline TL model against HYDROTEL.

3.3.1 Performance of the TL models using different forcing and static attribute variables from source and target datasets

The results of performing TL (forward model) using DD and SD are illustrated in Figure 9(a). Using the SD approach resulted in higher evaluation metrics (median values of 0.864, 0.768, 0.899, 42.347 m^3/s, and -8.73 % for KGE, Nash-log, Corr, RMSE, and PBIAS, respectively) compared to using DD (median values of 0.787, 0.745, 0.867, 48.374 m^3/s, and -17.72 % for KGE, Nash-log, Corr, RMSE, and PBIAS, respectively) to perform TL. In DI-TL models (see Figure 9(b)), the use of DD could achieve a slightly higher median KGE value than SD in developing DI-TL models.

514 The median Nash-log and RMSE values of DI-TL models using SD and DD were almost of the
515 same: median Nash-log values of 0.919 and 0.933 for DI-TL-0(8) and DI-TL-1(8), respectively.
516 Median RMSE values were 17.511 m^3/s and 17.987 m^3/s for the DI-TL models using SD and DD
517 during training, respectively. The median Corr value of the case when SD was used, was slightly
518 higher (value of 0.971 compared to 0.964). In contrast, the median PBIAS of using DD in DI-TL
519 models was lower (value of -1.193 % compared to -5.748 %) with smaller variabilities around the
520 median. Figure 10 shows the mean observed annual hydrograph compared to the mean annual
521 hydrographs of simulated flows by TL and DI-TL models using DD and SD for the Batiscan
522 watershed. In the hydrograph of TL models, using SD could improve the accuracy of simulated
523 flows compared to DD (comparison of blue and medium spring green curves in Figure 10). At the
524 same time, slight improvements were provided in DI-TL models using DD when compared to SD
525 (brown curve versus the yellow curve in Figure 10).

526 In summary, SD seemed to be a more effective approach and the rest of the models were built
527 using SD in TL except for cases where HYDROTEL simulated outputs were added as inputs to
528 the TL model.



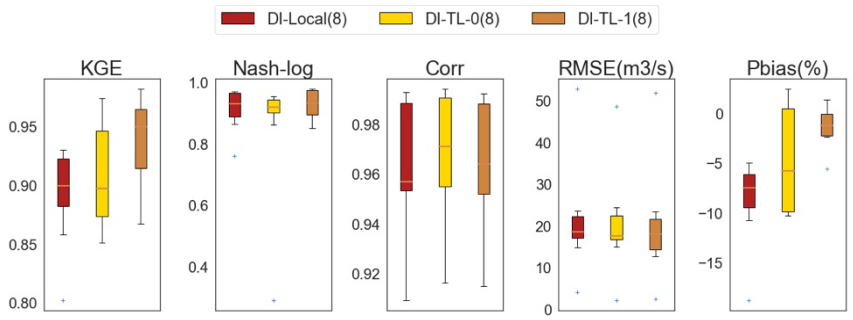


Figure 9: Comparison of the results of TL and DI-TL models using DD (performing TL while different forcing and static attributes were used from the source and target datasets -indicated by 1 in the legend) and SD (performing TL while the same forcing and static attributes were used from the source and target datasets - indicated by 0 in the legend) with the corresponding local model. The number of training watersheds is eight (illustrated by 8 in the legend). (a) forward models, and (b) DI models. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker delineate 5%, 25%, 50%, 75%, and 95% of evaluation metrics, respectively. Outliers are shown by cross markers. Each boxplot contains the metric for the eight (8) watersheds from the target region trained on all data available between 1982 and 1991. For HYDROTEL, the first available year for each watershed was used as a spin-up period.

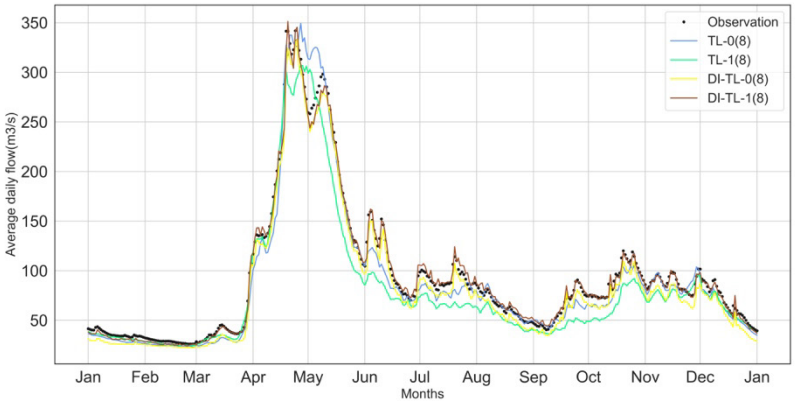


Figure 10: Comparison of observed and simulated 8-year mean annual hydrographs for the Batiscan watershed during the test period (12/1991-12/2002). Simulated flows were generated by TL (forward) and DI-TL models using DD (performing TL while different forcing and static attributes were used from the source and target datasets - indicated by 1 in the legend) and SD (performing TL while the same forcing and static attributes were used from the source and

target datasets -indicated by 0 in the legend) while the number of training watersheds is eight (indicated by 8 in the legend).

3.3.2 Performance of DI-TL models

The performance metrics reported for DI models are shown in Figure 11. Providing recent observed flows as an input to the TL model generally improved model accuracy. The values of the metrics of DI models showed that the increments provided by TL (weight initialization and freezing) were minor compared to applying TL with forward models (the sequence of increments can be realized by comparing the TL models with the local model).

Even for this series of models with minor increments, TL could elevate the results from 0.894 (median KGE value of DI-Local (8)) to 0.942 for the median KGE of DI-TL-0(34) which showed the best performance among DI models. Furthermore, focusing on DI-TL-0(8) and DI-TL-0(34), we see that their performances are very similar (for DI-TL-0(34) median values of 0.942, 0.953, 15.425 m^3/s, 0.970, and 0.066 % for KGE, Nash-log, RMSE, Corr, and PBIAS, respectively; and for DI-TL-0(8) median values of 0.904, 0.923, 16.025 m^3/s, 0.900, and -5.800 % for KGE, Nash-log, RMSE, Corr, and PBIAS, respectively). The performance of TL-DI-1(8) confirms the power of TL to improve the DI models even though the DI model was not used during pretraining (the values of the metrics reported in Figure 11 are too close to clearly see the difference. Indeed, without loss of continuity, if we now look at NSE (Nash-Sutcliffe Efficiency) values, the reported values for DI-Local (8) (0.908) and TL-DI-1 (8) (0.914) have a 0.06 NSE difference. Figure 12 shows the mean annual hydrograph of the Batiscan watershed for the simulated flows generated by DI models and the observed flows. The performance of TL-DI-1(8) in capturing the peak flows during the spring was similar to that of DI-TL-0(34) (dark blue curve versus orange curve) while having some problems in the summer and the fall to capture the peak flows. Although TL-DI-1(8) could achieve very good results given their performance metrics and hydrographs, when

compared to those of the DI-TL models (DI-TL-0(8) and DI-TL-0(34)), not including DI during pretraining caused larger variabilities in the predictions around the median.

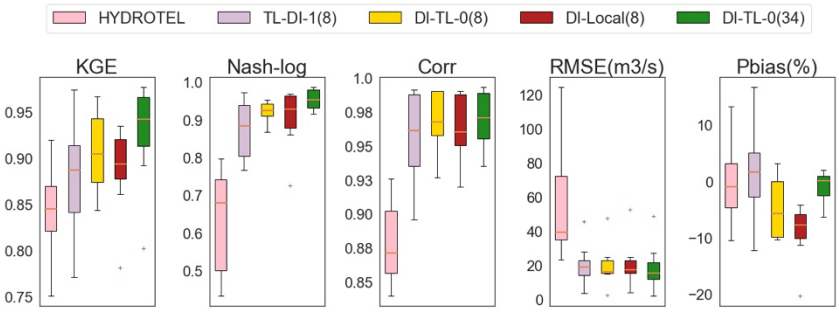


Figure 11: Comparison of evaluation metrics for DI-TL models during the test period with HYDROTEL. Numbers in the parentheses indicate the number of watersheds employed from the target dataset in model training. TL-DI-1(8) shows the performance of the model that the DI model was not used during pretraining. 1 and 0 in the legend are the indications of performing TL with DD (performing TL while different forcing and static attributes were used from the source and target datasets) and SD (performing TL while the same forcing and static attributes were used from the source and target datasets), respectively. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker delineate 5%, 25%, 50%, 75%, and 95% of evaluation metrics, respectively. Outliers are shown by cross markers. Each boxplot contains the metric for the eight (8) watersheds from the target region trained on all data available between 1982 and 1991. For HYDROTEL, the first available year for each watershed was used as a spin-up period.

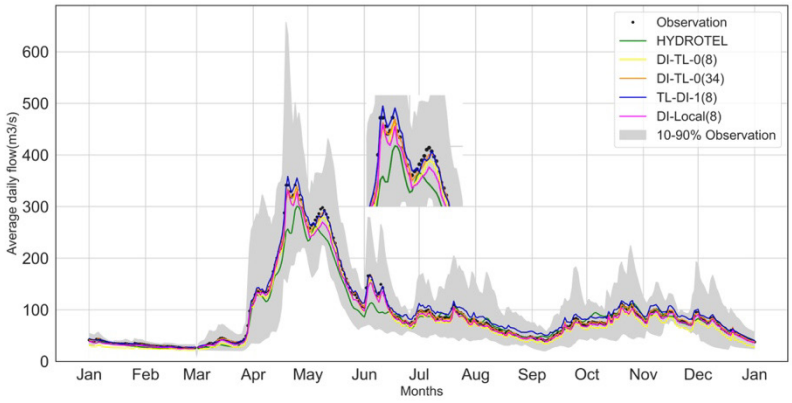


Figure 12: Comparison of simulated (DI-TL models, HYDROTEL, and the local model) and observed mean annual hydrographs for the Batiscan watershed (8-year mean) in the test period (12/1991-12/2002). The number of watersheds used from the target dataset in model training is indicated by numbers in parentheses. TL-DI-1(8) shows the performance of the model that the DI model was not used during pretraining. TL with DD (performing TL while different forcing and static attributes were used from the source and target datasets) and SD (performing TL while the same forcing and static attributes were used from the source and target datasets) are indicated by the numbers 1 and 0, respectively, in the legend. The lower and upper limits of the shaded area represent the 10^{th} and 90^{th} percentiles of observed data.

3.3.3 Performance of TL models incorporating simulated flows as inputs

Figure 13 shows the results of TL models including HYDROTEL simulated flows as inputs. Their median KGE values (TL-S-1(8) and DI-TL-S-1(8)) were nearly identical (0.84 and 0.838, respectively) to that of HYDROTEL (0.845) alone. However, lower variabilities in predictions were observed around the median value for TL-S-1(8) compared to HYDROTEL. Nash-log values of all physics-informed models (values of 0.795, 0.806, and 0.822 for Local-S(8), DI-TL-S-1(8), and TL-S-1(8), respectively) built in this study were higher than the median value of HYDROTEL (0.679). Once more, the variabilities of predictions around the median of Nash-log values were smaller for physics-informed models compared to HYDROTEL. The median values of Corr for the physics-informed models were greater than the median of HYDROTEL with superior performance for DI-TL-S-1(8) (value of 0.90). The median values of RMSE for the physics-informed models (values of 36.910, 35.335, and 35.90 m^3/s for Local-S(8), DI-TL-S-1(8), and TL-S-1(8), respectively) were slightly samller than that of HYDROTEL(39.36 m^3/s) alone. Finally, the PBIAS of HYDROTEL was closer to zero than those of the other models. Overall, TL-S-1(8) had the best performance among this series of models.

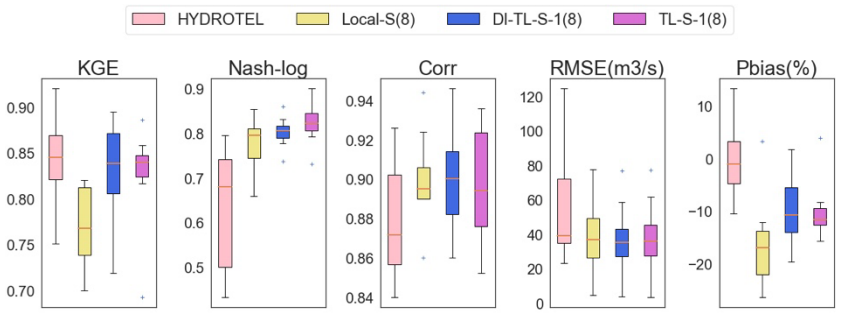


Figure 13: Results from the prediction in target watersheds using the TL models incorporating HYDROTEL simulated flows as inputs alongside prediction results of HYDROTEL. S illustrates including simulated flows in model training. 1 is the indication of performing TL while different forcing and static attributes were used from the source and target datasets. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker delineate 5%, 25%, 50%, 75%, and 95% of evaluation metrics, respectively. Outliers are shown by cross markers. Each boxplot contains the metric for the eight (8) watersheds from the target region trained on all data available between 1982 and 1991. For HYDROTEL, the first available year for each watershed was used as a spin-up period.

3.3.4 Performance of TL models when increasing the number of watersheds

Figure 14 shows the results of increasing the number of watersheds with similar climatological characteristics. In both series of models (forward and DI models), an increase in the number of watersheds enhanced the performance but of a different increment. While for TL models (TL-0(8) and TL-0(34)), slight improvements are observed for all performance metrics, for DI-TL models (DI-TL-0(8) and DI-TL-0(34)), including more watersheds during training using the target region resulted in greater improvements compared to TL models, especially in terms of KGE (values of 0.904 and 0.942 for DI-TL-0(8) and DI-TL-0(34)), Nash-log (values of 0.923 and 0.953 for DI-TL-0(8) and DI-TL-0(34)), and PBIAS(-5.80 % and 0.066 % for DI-TL-0(8) and DI-TL-0(34)).

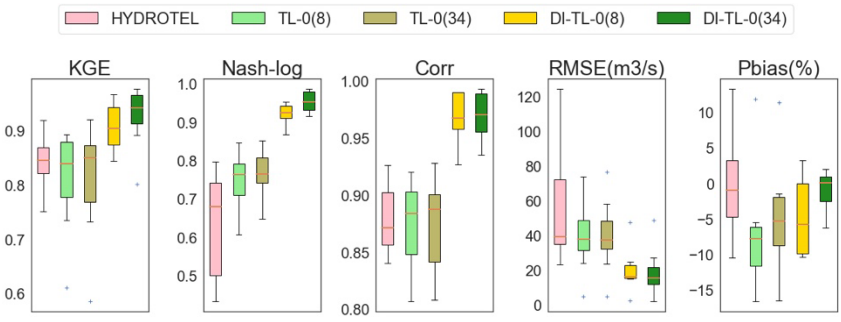


Figure 14: Performance of TL models with the number of watersheds increased (34 watersheds used in training) as opposed to the HYDROTEL and the TL models without increasing the number of watersheds (8 watersheds used in training). 0 is the indication of performing TL while the same forcing and static attributes were used from the source and target datasets. The lower whisker, lower box edge, center bar, upper box edge, and upper whisker delineate 5%, 25%, 50%, 75%, and 95% of evaluation metrics, respectively. Outliers are shown by cross markers. Each boxplot contains the metric for the eight (8) watersheds from the target region trained on all data available between 1982 and 1991. For HYDROTEL, the first available year for each watershed was used as a spin-up period.

3.4 Seasonal performance

Figure 15 displays the results of seasonal flow predictions by HYDROTEL and DI-TL-0(34) (this model represents the best TL model in this study). TL had better performance metrics throughout the year than HYDROTEL; having a better median test period during all seasons. In the winter, spring, summer, and fall, for instance, the median KGE values of DI-TL-0(34) were 0.911, 0.946,

0.892, and 0.904, respectively, and for HYDROTEL, the corresponding median KGE values were 0.761, 0.832, 0.659, and 0.711.

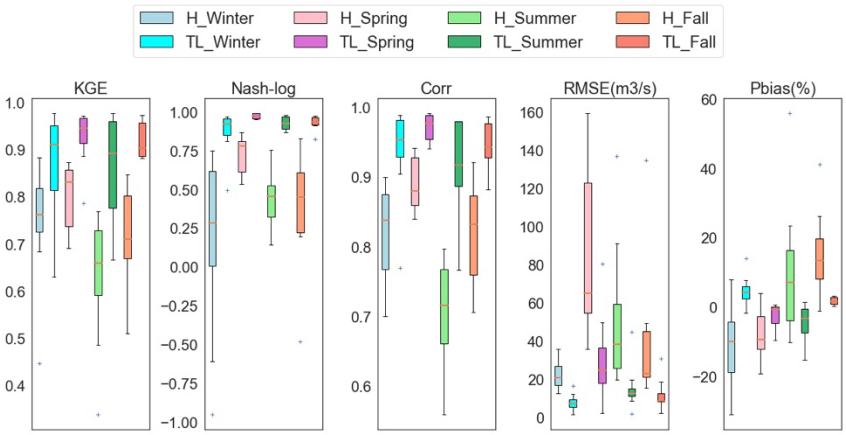


Figure 15: Comparison of seasonality performance of HYDROTEL with the TL model. H stands for HYDROTEL, while TL indicates the performance of the selected TL model for this analysis, DI-TL-0(34). The lower whisker, the lower box edge, the center bar, the upper box edge, and the upper whisker represent 5%, 25%, 50%, 75%, and 95% of the data, respectively. Each boxplot contains the metric for the eight (8) watersheds from the target region trained on all data available between 1982 and 1991. For HYDROTEL, the first available year for each watershed was used as a spin-up period.

4. Discussion

In this section, first, we discuss the performance of the baseline TL model and the results of testing the source model on target watersheds. Second, the incremental benefits of applying different approaches to improve TL for streamflow prediction are discussed. Finally, the seasonal performances of the best TL model and HYDROTEL are thoroughly discussed.

4.1. Testing the source model and baseline TL model

Results of testing the source model on target watersheds means that if we intend to test the forward model on a region where we only have meteorological data and static attributes, that would not be a practical test to take, and meteorological data are not impactful enough during testing. Comparison of the results of testing the DI model with the results of testing the forward model on the target region singles out the power of DI models to provide streamflow predictions

with acceptable accuracy even during testing while having access to some simulated flows. When compared to HYDROTEL, the baseline TL model and the local model both outperformed the latter in capturing the baseflows. That is due to the remarkable ability of LSTM models to capture low flow patterns. This suggests the current version of HYDROTEL cannot not reproduce satisfactorily baseflow processes. When calibrated, PBMs like HYDROTEL often focus more heavily on the peaks and sacrifice low flows, but LSTM is less sensitive to such tradeoffs. Analysis of the performance metrics and the mean annual hydrographs of the baseline TL model and of HYDROTEL indicates that the former requires some improvements in flow prediction during spring and summer, two seasons with a higher frequency of peak flows.

4.2. TL improving approaches

The analysis of the results of using DD (different forcing and static attribute variables for source and target datasets) and SD (same forcing and static attribute variables for source and target datasets) in TL and DI-TL models indicates that using SD seems preferable in this study. Although using DD provided slight improvements in DI-TL models; the gains were minor compared to improvements provided by SD in forward TL models (row(a) in Figure 9), where the use of SD showed superior performance over DD. However, this does not mean that using DD in TL modeling will not significantly improve with shorter data records or other regions since these models are data-driven, and results may differ from region to region based on the data involved in training. Moreover, the use of DD in TL development shows the power of DL to adapt easily to new conditions. However, it may not be as easy for PBMs. For PBMs, running the model without having the exact inputs required is relatively impossible. Although the performance of DI-TL models was improved compared to HYDROTEL, it should be noted that in this series of models, Q_{t-1} was included in TL modeling while data assimilation was not provided for HYDROTEL. However, it is much easier to integrate data (here, Q_{t-1}) in this way than performing standard data assimilation (e.g., ensemble Kalman filtering) with PBMs (e.g., (Samuel et al., 2019)). Comparing

673 increments provided by DI-TL models and TL models (sequence of increments were higher for
674 TL models), it can be realized that TL is mainly more effective for cases dealing with regional data
675 scarcity. Although the target dataset used in this study had access to fewer variables than those
676 available through CAMELS, we could access an almost complete flow record for the selected
677 watersheds of Southern Québec. This is in contradiction with the concept of data scarcity. But
678 these DI-TL models have the potential to be used for providing predictions for nearby watersheds
679 dealing with data sparsity by fine-tuning the DI-TL model. Comparable performances of DI-TL-
680 0(8) and DI-TL-0(34) (Figure 11) suggests that although increasing the number of watersheds
681 could decrease the prediction errors, we could get almost the same results with fewer watersheds.
682 Although the performance of TL-DI-1(8) was comparable to that of other models in the series of
683 DI-TL models, the analysis of the TL-DI-1(8) model revealed that larger variabilities made in the
684 predictions of the TL-DI-1(8), demonstrate the value of pretraining with DI over the source model
685 and how well knowledge gained from the source dataset can be used to improve predictions in
686 the target region. Analysis of TL models incorporating simulated flows of HYDROTEL showed
687 that DI-TL-S-1(8) could not elevate the results significantly, although the DI model was used
688 during pretraining. One of the reasons for this outcome could be associated with the errors
689 embedded in the HYDROTEL simulated flows that lowered the accuracy of predictions generated
690 by the TL model. Overall, analysis of the results of this approach shows using HYDROTEL
691 simulated flows as inputs were beneficial since they could improve Nash-log, Corr, and RMSE
692 while achieving similar KGE values. Moreover, they could reduce the variabilities of predictions
693 for all watersheds compared to HYDROTEL (especially based on the Nash-log and PBIAS
694 values). Although combining TL technique with DI showed significant enhancement in predictions
695 (Figure 11), the superior performance of TL-S-1(8) demonstrates the importance of physics-
696 informed TL and DL models in real-world applications. This is due to the restricted availability of
697 streamflow data in some areas and the often-higher cost of even obtaining them compared to
698 using PBM simulated flows as an alternative. Finally, TL models where the number of watersheds

had been increased, using a more extensive dataset at the target region achieved better performance in both forward and DI models. Small improvements provided by increasing the number of watersheds (consistent with the results reported for DI-TL models) indicate the importance of grouping the watersheds before training rather than just increasing the number of watersheds which can be further investigated in the future.

The prediction results indicate that TL models achieve better prediction performance compared to applying the LSTM model on the target region without TL (local model training), which is consistent with the findings previously reported by (Bedi, 2022; Li et al., 2021). In summary, among all TL models, the DI-TL model with the number of watersheds increased from 8 to 34 in the target region obtained the best performance (readers are referred to Figure S12 and Figure S13 to see the boxplots and hydrographs of all models). Compared to the work of Ma et al. (2021), although they used a higher density of basins with a shorter training period and a longer pretraining period, the highest median NSE value was 0.86 for CAMELS-CL and CAMELS-GB. In this study, the median NSE value of the baseline without any improvement was 0.803 using eight watersheds (TL-0(8)). As mentioned before by Feng et al. (2020), the source model developed over CAMELS reached exceptional performance, especially in mountainous and snow-dominated watersheds. The eight (8) watersheds selected for projecting TL modeling are classified as snow-dominated watersheds, demonstrating the potential of performing TL modeling for this region, achieving significant performance after applying the proposed TL improving approaches (median values of Nash-log of 0.953, KGE of 0.942, and NSE of 0.932 for DI-TL-0(34)). The final TL models were obtained by a partially frozen model probably due to having a smaller dataset in the target region. Moreover, Southern Québec watersheds can be qualified as large watersheds (the median drainage area for the watersheds at the target region was 3498 km^2 while that for CAMELS watersheds was 329.62 km^2); highlighting the fact that achieving the

optimal performance of the TL model depends on the sufficient change of weights during training. This is consistent with the results previously reported by Ma et al. (2021).

4.3. Seasonal performance of TL and HYDROTEL models

Results of TL and HYDROTEL models indicated better performance of both models in the spring and then during fall, the two seasons with large flow variability. Although HYDROTEL considered summer flows during calibration, its performance was deemed poor. The HYDROTEL version used for this study does not account for water uptake in the hydrographic network. Preliminary results with a newer version of the model have illustrated that accounting for water withdrawals can improve the results of simulated flows during the summer, albeit not shown here. Contrastingly, the performance of the DI-TL-0(34) model was relatively lower in summer (lower median, e.g., in KGE and Pearson Correlation) compared to its performance during other seasons and that may be because of the presence of more extreme flows during summer compared to winter. As a reminder during the calibration process of HYDROTEL, winter flows were not accounted for, yet the model performed relatively well compared to other seasons according to values of KGE, Corr, RMSE, and PBIAS. In contrast, DI-TL-0(34) model had the smallest median RMSE of 7.545 m^3/s and a high median KGE, Nash-log, and correlation (0.911, 0.919, and 0.954, respectively) during winter. This indicates the inherent advantage of LSTM models in tracking the hydrologic processes during the winter. Generally, both models had difficulties capturing flows during summer, and some actions should be considered to improve their results accordingly. Although the superior performance of HYDROTEL was observed during the spring in general, there is a probability of large bias involved in simulations during the spring due to higher values of Corr and RMSE compared to the performance during the winter.

5. Conclusions

Lagged discharge in long short-term memory (LSTM) networks has been shown to be a successful technique for streamflow prediction, or so-called data integration (DI). While the use of DI has proven useful with deep learning (DL), no study has previously shown the ability of a transfer learning (TL) model to reap the significant benefits of adding recent discharge data. In this study, an LSTM model trained on CAMELS (Addor et al., 2017) dataset over the CONUS (conterminous United States) was transferred to a snow-dominated region (utilizing HYSETS (Arsenault et al., 2020) as the target dataset), in Quebec, Canada, and multiple improvement strategies were implemented to enhance the performance of the TL model. Performances of TL models were compared to a physically based model (PBM), HYDROTEL. How well the source model would function in TL was determined by testing the source model on the target region. Results of testing the source model on the target region did not yield meaningful results when only meteorological and physiographic data from CAMELS were used during training. However, once the DI model had been used to produce predictions for the target region through testing, the accuracy of the predictions varied depending on the availability of recent observations of discharge in the target region.

In the TL scheme, the baseline TL model which was trained with meteorological and physiographic data performed well when compared to HYDROTEL, especially when it came to capturing low flows. However, it needed to be enhanced. Additionally, it was discovered that performing TL with the same meteorological and physiographic variables from the source and target datasets provided a more acceptable performance than doing so with different meteorological and physiographic variables. Among the approaches suggested for improving the performance of the baseline TL model, employing DI with a maximum number of watersheds from the target region achieved the highest prediction accuracy. The DI-TL model that was trained using only eight watersheds, nevertheless, was likewise highly effective. The performance of the

DI-TL models when DI was not used during pretraining showed that TL could even increase the performance of DI models, the models requiring less improvement as they employed recent discharge observations in the model training. When streamflow observations did not exist, including HYDROTEL simulated flows could improve the streamflow predictions by reducing the prediction errors.

The results of this study highlight the significance of the TL scheme to transfer knowledge from data-rich regions to snow-dominated regions and improve streamflow prediction for those regions due to the ability of the LSTM model to extract information from large datasets and tracking of long-term hydrological processes. The TL models used in this study can also be applied to nearby watersheds not having sufficient streamflow records and provide predictions with good accuracy. Furthermore, this study shows that large hydrological datasets like CAMELS contain more information that can possibly be modeled by a PBM. Improving the prediction accuracy of extreme flows to elevate the performance of the TL model during seasons like summer certainly represents a valuable avenue of research. Slight improvements provided by increasing the number of training watersheds in TL demonstrate the necessity of exploring some criteria to group the watersheds before training to achieve exceptional predictive skills in future works rather than just increasing the number of watersheds. Furthermore, it is suggested to investigate the performance of physics-informed TL models (Shen et al., 2023) during the spatial test (training in one site and testing in another site for the same period).

Acknowledgments

The authors would like to thank Richard Arsenault of ETS (École de Technologie Supérieure, Montreal, QC, Canada) for providing access to the HYSETS database. We also would like to thank the Québec Ministry of the Environment and the Fight against Climate Change (Ministère de l'Environnement et de la Lutte contre les changements climatiques) for allowing us to use parts

of their modeling platform of PHYSITEL; especially Charles Malenfant for his technical support. The authors acknowledge Stéphane Savary and Étienne Foulon from INRS (Centre Eau Terre Environement) for their assistance. This research was supported by the Natural Sciences and Engineering Research of Canada Discovery Program (A.N. Rousseau, principal investigator).

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293-5313.

Alipour, M., & Kibler, K. M. (2018). A framework for streamflow prediction in the world's most severely data-limited regions: Test of applicability and performance in a poorly-gauged region of China. *Journal of hydrology*, *557*, 41-54.

Alizadeh, B., Bafti, A. G., Kamangir, H., Zhang, Y., Wright, D. B., & Franz, K. J. (2021). A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. *Journal of hydrology*, *601*, 126526.

Arsenault, R., Brissette, F., Martel, J.-L., Troin, M., Lévesque, G., Davidson-Chaput, J., Gonzalez, M. C., Ameli, A., & Poulin, A. (2020). A comprehensive, multisource database for hydrometeorological modeling of 14,425 North American watersheds. *Scientific Data*, *7*(1), 1-12.

Barnett, T. P., Adam, J. C., & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, *438*(7066), 303-309.

Bedi, J. (2022). Transfer learning augmented enhanced memory network models for reference evapotranspiration estimation. *Knowledge-Based Systems*, *237*, 107717.

Bitew, M. M., & Gebremichael, M. (2011). Assessment of satellite rainfall products for streamflow simulation in medium watersheds of the Ethiopian highlands. *Hydrology and Earth System Sciences*, *15*(4), 1147-1155.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Proceedings of the 45th annual meeting of the association of computational linguistics,

Bouda, M., Rousseau, A. N., Gumiere, S. J., Gagnon, P., Konan, B., & Moussa, R. (2014). Implementation of an automatic calibration procedure for HYDROTEL based on prior OAT sensitivity and complementary identifiability analysis. *Hydrological Processes*, *28*(12), 3947-3961.

Bouda, M., Rousseau, A. N., Konan, B., Gagnon, P., & Gumiere, S. J. (2012). Bayesian uncertainty analysis of the distributed hydrological model HYDROTEL. *Journal of Hydrologic Engineering*, *17*(9), 1021-1032.

Cho, K., & Kim, Y. (2022). Improving streamflow prediction in the WRF-Hydro model with LSTM networks. *Journal of Hydrology*, *605*, 127297. https://doi.org/10.1016/j.jhydrol.2021.127297

DEFACC. (2018). Document d'accompagnement de l'Atlas hydroclimatique du Quebec méridional. *Ministère de l'Environnement et de la Lutte contre les Changements Climatiques, Quebec.*

Dibike, Y. B., & Coulibaly, P. (2005). Hydrologic impact of climate change in the Saguenay watershed: comparison of downscaling methods and hydrologic models. *Journal of Hydrology*, *307*(1-4), 145-163.

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The Data Synergy Effects of Time-Series Deep Learning Models in Hydrology. *Water Resources Research*, *58*(4), e2021WR029583.

Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *Journal of Hydrometeorology*, *21*(3), 399-413.

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793.

Feng, D., Lawson, K., & Shen, C. (2021). Mitigating prediction error of deep learning streamflow models in large data-sparse regions with ensemble modeling and soft data. *Geophysical Research Letters*, *48*(14), e2021GL092999.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., & Villeneuve, J.-P. (2001). Distributed watershed model compatible with remote sensing and GIS data. I: Description of model. *Journal of hydrologic engineering*, *6*(2), 91-99.

Fossey, M., Rousseau, A. N., Bensalma, F., Savary, S., & Royer, A. (2015). Integrating isolated and riparian wetland modules in the PHYSITEL/HYDROTEL modelling platform: model performance and diagnosis. *Hydrological Processes*, *29*(22), 4683-4702.

Foulon, É., & Rousseau, A. N. (2018). Equifinality and automatic calibration: What is the impact of hypothesizing an optimal parameter set on modelled hydrological processes? *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, *43*(1), 47-67.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural computation*, *12*(10), 2451-2471. https://doi.org/10.1162/089976600300015015

Girihagama, L., Naveed Khaliq, M., Lamontagne, P., Perdikaris, J., Roy, R., Sushama, L., & Elshorbagy, A. (2022). Streamflow modelling and forecasting for Canadian watersheds using LSTM networks with attention mechanism. *Neural Computing and Applications*, *34*(22), 19995-20015. https://doi.org/10.1007/s00521-022-07523-8

Granata, F., Di Nunno, F., & de Marinis, G. (2022). Stacked machine learning algorithms and bidirectional long short-term memory networks for multi-step ahead streamflow forecasting: A comparative study. *Journal of Hydrology*, *613*, 128431. https://doi.org/10.1016/j.jhydrol.2022.128431

Guo, X., Chen, Y., Liu, X., & Zhao, Y. (2020). Extraction of snow cover from high-resolution remote sensing imagery using deep learning on a small dataset. *Remote Sensing Letters*, *11*(1), 66-75.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., & Schepers, D. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Hunt, K. M., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences*, *26*(21), 5449-5472. https://doi.org/10.5194/hess-26-5449-2022

Hutter, F., Lücke, J., & Schmidt-Thieme, L. (2015). Beyond manual tuning of hyperparameters. *KI-Künstliche Intelligenz*, *29*(4), 329-337.

Kimura, N., Yoshinaga, I., Sekijima, K., Azechi, I., & Baba, D. (2019). Convolutional neural network coupled with a transfer-learning approach for time-series flood predictions. *Water*, *12*(1), 96.

Kling, H., Fuchs, M., & Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology*, *424*, 264-277.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005-6022.

Krause, P., Boyle, D., & Bäse, F. (2005). Comparison of different efficiency criteria for hydrological model assessment. *Advances in geosciences*, *5*, 89-97.

Li, Q., Wang, Z., Shangguan, W., Li, L., Yao, Y., & Yu, F. (2021). Improved daily SMAP satellite soil moisture prediction over China using deep learning model with transfer learning. *Journal of Hydrology*, *600*, 126698.

Liu, Z., Yang, Q., Shao, J., Wang, G., Liu, H., Tang, X., Xue, Y., & Bai, L. (2022). Improving daily precipitation estimation in the data scarce area by merging rain gauge and TRMM data with a transfer learning framework. *Journal of Hydrology*, *613*, 128455.

Ma, K., Feng, D., Lawson, K., Tsai, W. P., Liang, C., Huang, X., Sharma, A., & Shen, C. (2021). Transferring hydrologic data across continents–leveraging data-rich regions to improve hydrologic prediction in data-sparse regions. *Water Resources Research*, *57*(5), e2020WR028600.

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*.

MELCC. (2022). *Daily hydrometric data*. https://www.cehq.gouv.qc.ca/atlas-hydroclimatique/stations-hydrometriques/index.htm

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885-900.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282-290.

Nawi, N. M., Atomi, W. H., & Rehman, M. Z. (2013). The effect of data pre-processing on optimized training of artificial neural networks. *Procedia Technology*, *11*, 32-39.

Noël, P., Rousseau, A. N., Paniconi, C., & Nadeau, D. F. (2014). Algorithm for delineating and extracting hillslopes and hillslope width functions from gridded elevation data. *Journal of Hydrologic Engineering*, *19*(2), 366-374.

Nowakowski, A., Mrziglod, J., Spiller, D., Bonifacio, R., Ferrari, I., Mathieu, P. P., Garcia-Herranz, M., & Kim, D.-H. (2021). Crop type mapping by using transfer learning. *International Journal of Applied Earth Observation and Geoinformation*, *98*, 102313.

Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., & Michel, C. (2006). Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations. *Water Resources Research*, *42*(7).

Ouyang, W., Lawson, K., Feng, D., Ye, L., Zhang, C., & Shen, C. (2021). Continental-scale streamflow modeling of basins with reservoirs: Towards a coherent deep-learning-based strategy. *Journal of hydrology*, *599*, 126455.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345-1359.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25-45.

Pratt, L., & Jennings, B. (1996). A survey of transfer between connectionist networks. *Connection Science*, *8*(2), 163-184.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. *Environmental Research Letters*, *16*(2), 024025. https://doi.org/10.1088/1748-9326/abd501

Rahmani, F., Shen, C., Oliver, S., Lawson, K., & Appling, A. (2021). Deep learning approaches for improving prediction of daily stream temperature in data-scarce, unmonitored, and dammed basins. *Hydrological Processes*, *35*(11), e14400. https://doi.org/10.1002/hyp.14400

Redko, I., Morvant, E., Habrard, A., Sebban, M., & Bennani, Y. (2019). *Advances in domain adaptation theory*. Elsevier. https://doi.org/10.1016/C2016-0-05108-2

Samuel, J., Rousseau, A. N., Abbasnezhadi, K., & Savary, S. (2019). Development and evaluation of a hydrologic data-assimilation scheme for short-range flow and inflow forecasts in a data-sparse high-latitude region using a distributed model and ensemble Kalman filtering. *Advances in Water Resources*, *130*, 198-220.

Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., Baity-Jesi, M., Fenicia, F., Kifer, D., & Li, L. (2023). Differentiable modeling to unify machine learning and physical models and advance Geosciences. *arXiv preprint arXiv:2301.04027*. https://doi.org/10.5194/egusphere-egu23-15968

Shen, C., Chen, X., & Laloy, E. (2021). *Broadening the use of machine learning in hydrology* (Vol. 3). Frontiers Media SA. https://doi.org/10.3389/frwa.2021.681023

Shen, C., Laloy, E., Elshorbagy, A., Albert, A., Bales, J., Chang, F.-J., Ganguly, S., Hsu, K.-L., Kifer, D., & Fang, Z. (2018). HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. *Hydrology and Earth System Sciences*, *22*(11), 5639-5656.

Shen, C., & Lawson, K. (2021). Applications of deep learning in hydrology. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 283-297. https://doi.org/10.1002/9781119646181.ch19

Shin, H.-C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, *35*(5), 1285-1298.

Singh, J., Knapp, H. V., Arnold, J., & Demissie, M. (2005). Hydrological modeling of the Iroquois river watershed using HSPF and SWAT 1. *JAWRA Journal of the American Water Resources Association*, *41*(2), 343-360.

Tarek, M., Brissette, F. P., & Arsenault, R. (2020a). Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. *Hydrology and Earth System Sciences*, *24*(5), 2527-2544.

Tarek, M., Brissette, F. P., & Arsenault, R. (2020b). Large-scale analysis of global gridded precipitation and temperature datasets for climate change impact studies. *Journal of Hydrometeorology*, *21*(11), 2623-2640.

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1).

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242-264). IGI global. https://doi.org/10.4018/978-1-60566-766-9.ch011

Turcotte, R., Fortin, J.-P., Rousseau, A. N., Massicotte, S., & Villeneuve, J.-P. (2001). Determination of the drainage structure of a watershed using a digital elevation model and a digital river and lake network. *Journal of hydrology*, *240*(3-4), 225-242.

Turcotte, R., Fortin, L.-G., Fortin, V., Fortin, J.-P., & Villeneuve, J.-P. (2007). Operational analysis of the spatial distribution and the temporal evolution of the snowpack water equivalent in southern Québec, Canada. *Hydrology research*, *38*(3), 211-234.

Turcotte, R., Lacombe, P., Dimnik, C., & Villeneuve, J. (2004). Distributed hydrological prediction for the management of Quebec's public dams. *Canadian Journal of Civil Engineering*, *31*(2), 308-320.

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. *Journal of Hydrology*, *181*(1-4), 23-48.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zhao, G., Pang, B., Xu, Z., Cui, L., Wang, J., Zuo, D., & Peng, D. (2021). Improving urban flood susceptibility mapping using transfer learning. *Journal of Hydrology*, *602*, 126777. https://doi.org/10.1016/j.jhydrol.2021.126777

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43-76.