Journal of Hydrology 623 (2023) 129777

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Improved historical reconstruction of daily flows and annual maxima in gauged and ungauged basins

Jean-Luc Martel[a,*], Richard Arsenault[a], Simon Lachance-Cloutier[b], Mariana Castaneda-Gonzalez[a], Richard Turcotte[b], Annie Poulin[a]

[a] École de technologie supérieure, Department of Construction Engineering, Hydrology, Climate and Climate Change (HC^3) Laboratory, Canada
[b] Ministère de l'Environnement et de la Lutte contre les changements climatiques (MELCC), Direction de l'expertise hydrique (DEH), Canada

ARTICLE INFO

Keywords:
Multi-model averaging technique
Optimal interpolation
Distributed hydrological model
HYDROTEL
Prediction in ungauged basins

ABSTRACT

This study presents a method to improve the reconstruction of historical flows on gauged and ungauged basins. To do so, a multi-model weighted averaging of hydrological model simulations for a large spatial domain in the province of Quebec, Canada, is used. The distributed hydrological model HYDROTEL was implemented over the region and covered 95 gauged basins. An optimal interpolation (OI) assimilation method was first implemented as a baseline to improve the HYDROTEL flow simulations over the 95 basins. Then, a series of multi-model averaging techniques were applied to an ensemble of 144 HYDROTEL simulations that were generated by modifying parameter sets, driving weather datasets and evapotranspiration modules. The averaging methods were applied in a leave-one-out cross-validation scheme, where all 94 gauged basins were pooled together to compute the weights, and those weights were applied to the 95th basin. All basins were evaluated in such a manner and compared to the OI method. Implementing a year-by-year (or shorter period) weighting scheme instead of computing weights over all available data significantly improved the results. This allowed the weights to better reflect each year's hydrological characteristics rather than compromising to improve the overall average. The Kling-Gupta Efficiency (KGE) and peak flow metrics showed that the Granger-Ramanathan variant "A" (GRA) was similar in performance to the OI method but did not have the drawbacks that OI can typically introduce. The multi-model application can also be further improved by adding more simulations from other hydrological models, whereas the OI method cannot make use of such additional information, thus hitting a performance plateau. This study shows that it is possible to improve regional hydrological model simulations, both for overall flows and peak flows, and on the historical period for both gauged and ungauged basins. This can then be used to better estimate risk in flood frequency analysis and other statistical analyses.

1. Introduction

Using flood frequency analyses for the extrapolation of rare flood events (e.g., 100-year floods) is important when dealing with floodplain mapping and the design of hydraulic structures such as bridges and culverts. Extrapolating to such recurrences entails significant levels of epistemic uncertainty, especially when dealing with relatively small time series. Considering that there is also an interest in conducting such analyses in areas where streamflow observations are not available at all, there is a need for prediction in ungauged basins (PUB) to rebuild historical streamflow pseudo-observations (Hrachowitz et al., 2013). However, PUB also leads to its share of challenges and uncertainties (Blöschl et al., 2013; Sivapalan et al., 2003). While a regional flood frequency analysis could theoretically be used instead, several shortcomings such as short time series and basins heterogeneity in terms physical properties, climatology and hydrological processes, can render this option sub-optimal, and even undesirable (Ouarda et al., 2008; Shu and Ouarda, 2007). For peak flows particularly, other methods can be used such as in Kim and Shin (2018), where peak flow is estimated using the relationship between the ungauged basin's runoff coefficient and curve number which are estimated from donor basins. These methods rely on regionalization of parameters to the ungauged sites and can be considered model-independent. These methods also only generate statistical descriptors of the flow regime (flow indicators) and do not allow generating complete time series.

Hydrological modelling is an important tool to perform PUB,

[*] Corresponding author.
E-mail address: jean-luc.martel@etsmtl.ca (J.-L. Martel).

https://doi.org/10.1016/j.jhydrol.2023.129777
Received 20 October 2022; Received in revised form 8 May 2023; Accepted 3 June 2023
Available online 8 June 2023

allowing to simulate streamflow time series that could ultimately serve as the basis for flood frequency analyses (Razavi and Coulibaly, 2013). Lumped hydrological models can be applied on ungauged basins by using different types of regionalization methods, such as spatial proximity or physical similarity (Arsenault and Brissette, 2014). On the other hand, distributed hydrological models have inherent characteristics allowing them to simulate streamflow across entire regions under study, including as many ungauged sub-basins as needed. While in theory this makes distributed models the ideal option when dealing with PUB, other difficulties, such as the larger amount of physical and climatological data required, the time needed for model setup and model calibration (Martel et al., 2020), should not be overlooked.

Even though lumped or distributed models can provide streamflow PUB, multiple sources of uncertainty remain, such as the hydrological model's structure (Arsenault and Brissette, 2016), the climatological and physiographic data used (Papacharalampous and Tyralis, 2022), the calibration parameters and their equifinality (Arsenault and Brissette, 2014) and errors in hydrometric and meteorological observations used to calibrate the models (Troin et al., 2022). Thus, in order to obtain the most accurate historical streamflow pseudo-observations possible, there is an advantage to post-process the hydrological model outputs before conducting a flood frequency analysis (Lachance-Cloutier et al., 2017).

This type of post-processing can be achieved with different methods, the optimal interpolation (OI; Lachance-Cloutier et al., 2017) having shown to be a strong contender. In essence, OI is a statistically optimal method based on known theory which evaluates the spatial structure of errors in the distributed model response compared to the available observations and interpolates this error in such a way that the ungauged sites can be corrected. This type of data assimilation technique has recently been shown to provide an immediate and significant gain in performance compared to the raw simulation from a hydrological model (Lachance-Cloutier et al., 2017; Ly et al., 2013), but has traditionally been used in the field of meteorology (Heo et al., 2018; Oke et al., 2010; Phillips, 1982).

Another way of dealing with these different sources of uncertainty is to address them through an ensemble of different hydrological simulations (Arsenault et al., 2015) from both lumped and distributed models. Then, different multi-model averaging methods aim at obtaining an optimal weighting of each individual simulation that makes it possible to combine the strengths from several raw simulations to make gains in robustness and performance (Diks and Vrugt, 2010). Arsenault et al. (2015) showed that multi-model averaging techniques generally provided better streamflow simulations than those of any individual model that is part of the ensemble on 76% of a set of 429 basins in the United States. Furthermore, results showed that multi-model averaging has the advantage of providing excellent performance without needing to identify a priori which hydrological model would be the best for the given basin. Finally, the authors identify the Granger-Ramanathan type C (GRC; Granger and Ramanathan, 1994) as the best solution for site-specific flow estimation from the nine tested methods. The same methods were also tested on 383 basins in China with the results again pointing to GRC as being the best multi-model averaging method in streamflow estimation at gauged locations (Wan et al., 2021). Arsenault and Brissette (2016) applied multi-model averaging concepts to ungauged basins but noted that the GRC method, which includes a bias correction term, could lead to problems under regionalization due to the bias term scaling and thus implemented the Granger-Ramanathan type A (GRA; Granger and Ramanathan, 1984) algorithm which does not require the bias term. They found that regionalization of lumped hydrological models to ungauged sites generated streamflow that did not preserve statistics that could be corrected with multi-model averaging and thus found that multi-model regionalization of lumped models was not recommended. However, Exbrayat et al. (2011) and Razavi and Coulibaly (2016) showed that multi-model averaging could perform well in regionalization depending on the region of interest as well as the number and type of contributing hydrological model. When dealing with PUB, the combination of simulations from a distributed model can be more interesting since it will avoid the additional uncertainty from regionalization methods needed to transpose lumped models on ungauged sites. However, while combining a small number of raw simulations will provide improvements, it may not be enough to surpass a method such as the OI, in terms of performance.

To generate the most accurate historical streamflow pseudo-observations to conduct a flood frequency analysis in ungauged basins, the question remains: what is the most optimal method to use? Our hypothesis is that using a sufficiently large sample of raw hydrological simulations will provide access to more flexibility and degrees of freedom that can lead to further performance gains with some diminishing returns. In the right context, this has the potential to outperform a method like OI which is limited to a one-time – but significant – gain in performance. Furthermore, using raw simulations from a distributed hydrological model would not be subject to the data assimilation techniques' shortcomings previously raised.

The aim of this paper is to develop a methodology allowing to combine multiple raw simulations from a distributed hydrological model that outperforms the baseline data assimilation technique for the historical reconstruction of daily streamflow pseudo-observations and annual maxima time series in ungauged basins. Section 2 of the paper presents the study site, test bench and proposed methodology to reach the study's objectives, results are then presented in Section 3 and discussed in Section 4, followed by a conclusion and future work in Section 5.

2. Methodology

2.1. Study site

This study focuses on the meridional part of the province of Quebec, covering an area of approximately 726 000 square kilometers. A selection of 95 basins (shown in Fig. 1) from the 259 gauged basins operated by the *Direction de l'expertise hydrique (DEH)* of the Ministry of the Environment and Fight Against Climate Change (MELCC) was made. This selection was designed to provide coverage of both south (32 basins) and north (63 basins) shores of the St-Lawrence River, covering about 31% of the study area, while keeping the hydrometric stations with longer observational records of good quality.

The average physical characteristics, annual climatology and distribution of land-use is provided in Table 1. Each of the basins will be used in a leave-one-out-cross-validation (LOOCV) to evaluate the proposed methodology in the context of ungauged basins.

2.2. Hydrological modeling

HYDROTEL is a semi-distributed hydrological model developed by Fortin et al. (2001a); and Fortin et al. (2001b). The DEH exploits this model for its daily hydrological forecasts over 259 gauged and 28,035 ungauged basins and river reaches across the region of interest. A strong motivation behind the use of the HYDROTEL semi-distributed model is that the DEH has shared the complete calibrated platform used operationally for the region of interest (see Fig. 1). This allowed the development of a test bench based on multiple calibration strategies involving different parameters, input meteorological datasets and potential evapotranspiration formulas that also enable to emulate the operational limits. Fortin et al. (2001a) provides a complete description of HYDROTEL's required inputs and simulation of hydrological processes, which are summarized hereafter.

In terms of input variables, HYDROTEL requires drainage structure, land-use, soil type from high-resolution remotely sensed data, as well as distributed meteorological inputs (i.e., total daily precipitation in water equivalent, daily minimum and maximum temperature). Each basin (gauged or ungauged) is split into multiple homogenous sub-basins for which all hydrological processes (e.g., snowmelt and

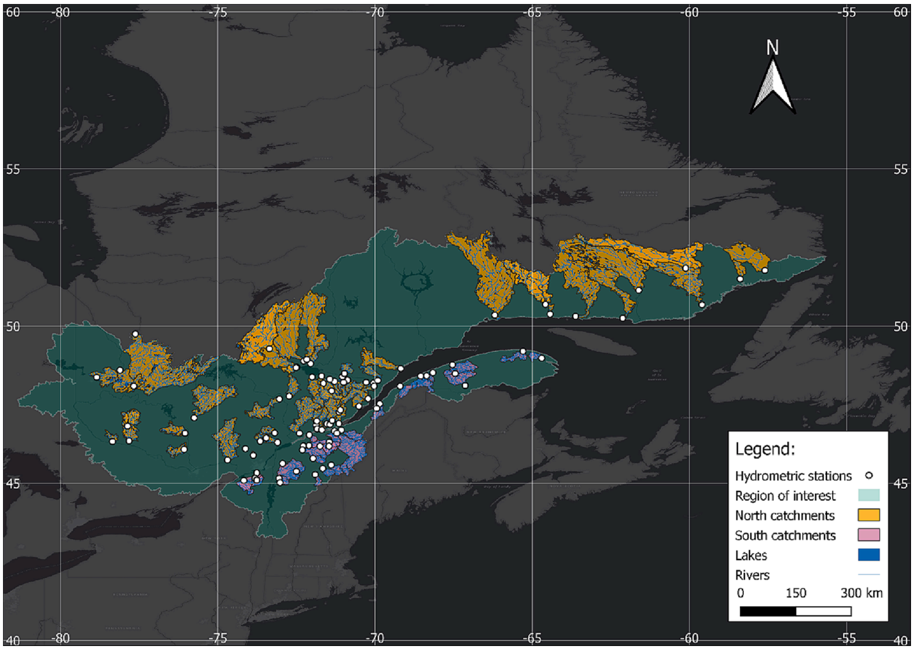

Fig. 1. Study area and the location of hydrometric stations and their basins.

Table 1
Physical characteristics, annual hydroclimatology and distribution of land-use of the 95 study basins.

Basin descriptor	Minimum	1st quartile	Median	3rd quartile	Maximum
Physical properties					
Drainage area [km^2]	44.3	468.1	991.8	2649.6	21525.0
Elevation [m]	87.7	280.4	386.6	489.7	865.2
Slope [m/m]	0.0	0.0	0.1	0.1	0.2
Climate properties					
Total precipitation [mm]	793.7	928.5	1002.7	1114.6	1341.2
Mean temperature [°C]	−1.9	0.8	2.0	3.9	6.8
Min. temperature [°C]	−42.5	−37.8	−35.1	−32.3	−29.5
Max. temperature [°C]	22.9	29.4	30.6	31.3	32.7
Land-use properties					
Agricultural pasture [%]	0.0	0.2	2.4	19.8	60.4
Bare ground [%]	0.0	2.1	4.1	7.5	21.4
Bog [%]	0.0	0.9	2.1	4.0	28.2
Coniferous forest [%]	5.4	28.3	40.3	50.4	84.5
Deciduous forest [%]	2.8	22.2	32.7	37.7	62.2
Impervious [%]	0.0	0.0	0.6	1.9	13.2
Water [%]	0.1	1.6	4.2	7.8	13.1
Wet land [%]	0.0	1.5	2.3	2.9	12.1

evapotranspiration) are computed independently.

Gridded meteorological datasets are used as inputs and corrections are also made based on each sub-basin's average elevation. Precipitation is also separated into rain and snow components on each sub-basin with a linear interpolation using both minimum and maximum daily temperature, as well as a threshold temperature for the separation between rain and snow (which is calibrated).

The evolution of the snow cover's characteristics is done with a mixed degree-day and energy balance method. However, the net absorbed solar radiation is simply estimated from a degree-day methodology. Three land-use classes are considered for the snowpack simulation with distinct melting factors for: coniferous forests, deciduous forests and open areas. The 95 selected basin's land-use types described in Table 1 are split into these three categories accordingly.

The HYDROTEL model has the option to switch between six potential evapotranspiration (PET) formulas, but only three were selected in this study: Hydro-Québec, Linacre and McGuinness. The selection of these three PET formulas was made with the goal to provide a wide range of the inter-model variability needed for this study, while limiting the number of calibrations to be performed. The first is an empirical equation developed over local basins by Hydro-Quebec (Fortin, 2000) and only requires minimum and maximum air temperature as inputs. The Linacre (1977) formula is derived from Penman and Keen (1948) and requires dew point and air temperatures, as well as both elevation and latitude of the station as inputs. McGuinness (McGuinness and Bordne, 1972) is a radiation-based formula that only requires mean air temperature as well as extra-terrestrial radiation, which can be estimated using the basin average latitude and the Julian day. It can be noted that Oudin et al. (2005) compared 27 evapotranspiration formulas (including Linacre and McGuiness) and showed that McGuinness provided the best results for hydrological modeling over 308 basins located in France, Australia and the United States.

The vertical water balance is conducted using a three layers soil model allowing to approximate the physical macro-processes involved during the infiltration and vertical redistribution of water over a soil column. The first and relatively thin layer (~10 to 20 cm) is affected by the soil evaporation and controls the surface runoff. The second layer is a transition zone between the first and third layer and produces delayed flows. The third layer is typically saturated and provides the base flow. All three layers can be affected by transpiration depending on land properties. The combined flow from the vertical water balance is then routed using a reference geomorphological hydrograph specific to each sub-basin. This geomorphological hydrograph is derived using the kinematic wave approximation with a reference flow depth and is obtained using two different land-uses (forested and open areas) for which different Manning's roughness coefficients are used.

Flow through the hydrographic network is also computed using the kinematic wave estimation. These computations are performed for each river reach based on their respective characteristics, namely the length, width, slope and Manning's roughness coefficient. However, when the reach is a lake or a reservoir, the classical continuity equation is used instead, and the flow is estimated using a flow-depth relationship depending on the width of the lake outlet.

2.3. Meteorological datasets

A selection of meteorological datasets was made with two goals in mind: 1) a reasonably long temporal coverage to conduct a flood frequency analysis and 2) diversity among the types of datasets. A total of three different meteorological datasets were thus selected to provide precipitation and temperature inputs to the HYDROTEL model: the MELCC gridded observed dataset, the ERA5 reanalysis and the SCDNA weather station product. The common period for these three datasets was from 1979 to 2018 and was kept for this study. Daily precipitation, minimum and maximum temperature were extracted for all three datasets and used as inputs for the various calibrations of the HYDROTEL model.

The daily observation gridded dataset developed by the MELCC (DCAQ; Bergeron, 2016) was selected as it is currently used as inputs in the HYDROTEL model for the DEH daily forecasting. Quality-controlled weather stations from both MELCC and Environment and Climate Change Canada (ECCC) networks were used for the interpolation, providing coverage for the period between 1961 and 2021. Ordinary kriging interpolation is used to obtain the best non-biased estimations without requiring the mean over the entire domain nor assuming its stationarity (Wackernagel, 2003). This method enables the consideration of a local average through the use of a restricted interpolation neighborhood (Bergeron et al., 2016; Li and Heap, 2014).

The ERA5 (C3S, 2019; Hersbach et al., 2020) reanalysis dataset is also used as inputs to HYDROTEL. ERA5 is the fifth generation of the European Centre for Medium-Range Weather Forecasts (ECMWF) which provides global hydrometeorological and atmospheric variables from 1950 to present. This reanalysis provides hourly outputs at a horizontal resolution of $0.25^\circ \times 0.25^\circ$. Tarek et al. (2020) demonstrated that hydrological modeling using ERA5 as inputs to hydrological modeling provides equivalent performance to observations over most of North America, including the region of interest.

The weather station product used is the station-based serially complete dataset for North America (SCDNA; Tang et al., 2020) covering the 1979–2018 period. This dataset includes 27,276 stations for which a strict quality control is performed, and the missing data is reconstructed using different strategies such as machine learning, quantile mapping and spatial interpolation. While the observed weather network used to generate the MELCC gridded dataset could also have been used, it was decided that it would only bring marginal additional information compared to its interpolated version, limiting the targeted degrees of freedom.

2.4. Development of the test bench

A test bench was developed using the platform provided by the DEH to test the work hypothesis. The HYDROTEL parameters, objective function and optimization algorithm for the calibration of the different HYDROTEL models will be covered in this section, followed by the calibration strategy to generate a total of 144 different calibrations over each of the two regions of interest, i.e., North and South domains.

2.4.1. HYDROTEL parameters

HYDROTEL has a total of 27 internal parameters that can be adjusted. Based on previous work done by Turcotte et al. (2007), 16 of these parameters can be fixed considering their low impact on the calibration objective function or due to their additive or multiplicative corrective nature applied to input data. The remaining eleven parameters to be calibrated are described in Table 2.

Along the eleven free parameters described in Table 2, two additional fixed parameters were modified to provide the desired flexibility within the HYDROTEL calibrations: the reference flow depth for the geomorphological hydrograph and a multiplicative constant for the lake outlet's width.

For the flow over the terrestrial part of the basin, two different reference flow depths were used to consider the relationship between flow depth and the Strahler stream order. Previous testing allowed selecting optimal flow depth values for the study site (0.006 and 0.010 m) and showed that a higher Strahler number performed better with smaller flow depth and vice-versa.

For the flow through the hydrographic network, the large number of lakes (especially on the North Shore of the St-Lawrence River) has shown to have a significant impact on the resulting hydrographs. Previous testing by the DEH showed that using larger width for the lake outlet allowed it to provide a more realistic hydrograph in this context. To account for this, a multiplicative constant of 1.82 applied to the lake outlet's width was found to provide the best results. In this study, both values of 1.00 and 1.82 were considered.

2.4.2. Objective function and optimization algorithm

To evaluate the performance of the HYDROTEL model during its calibration and validation processes, the modified Kling-Gupta Efficiency (KGE; Gupta et al., 2009; Kling et al., 2012) objective function was used and is defined as follows:

$$1 - KGE = \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{1}$$

where r is the linear correlation coefficient, α is the ratio of the coefficients of variation and β is the bias ratio. All these values are dimensionless and are computed between observed and simulated streamflow. The KGE ranges from a value of -∞ to 1, where 1 represents a perfect fit between observed and simulated streamflow.

As suggested by Huot (2014), the best-suited optimization method for the HYDROTEL model is the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007). A total of 500 model iterations was chosen as a compromise between good enough calibration quality and time to complete the most combinations possible. The goal here is not to obtain the best possible local optima, but rather to generate good enough calibration that will provide additional flexibility and degrees of freedom when combining all available simulations using a multi-model averaging method.

2.4.3. Calibration strategy

To obtain the desired level of flexibility from the HYDROTEL simulations, multiple calibrations of HYDROTEL parameters were conducted using different meteorological datasets. Each of the HYDROTEL calibrations follow the procedure detailed in Fig. 2. All combinations of meteorological datasets (DCAQ, ERA5, SCDNA), flow depth (0.006 and

Table 2
HYDROTEL parameters included in the calibration process.

HYDROTEL parameters	Sub-model
Threshold temperature for rain to snow	Input data
Depth of the first soil layer	Soil model
Depth of the second soil layer	Soil model
Recession coefficient	Soil model
Melting temperature threshold in a coniferous forest	Snow model
Melting temperature threshold in a deciduous forest	Snow model
Melting temperature threshold in an open environment	Snow model
Maximum melt rate in a coniferous forest	Snow model
Maximum melt rate in a deciduous forest	Snow model
Maximum melt rate in an open environment	Snow model
PET multiplicative coefficient	PET formula

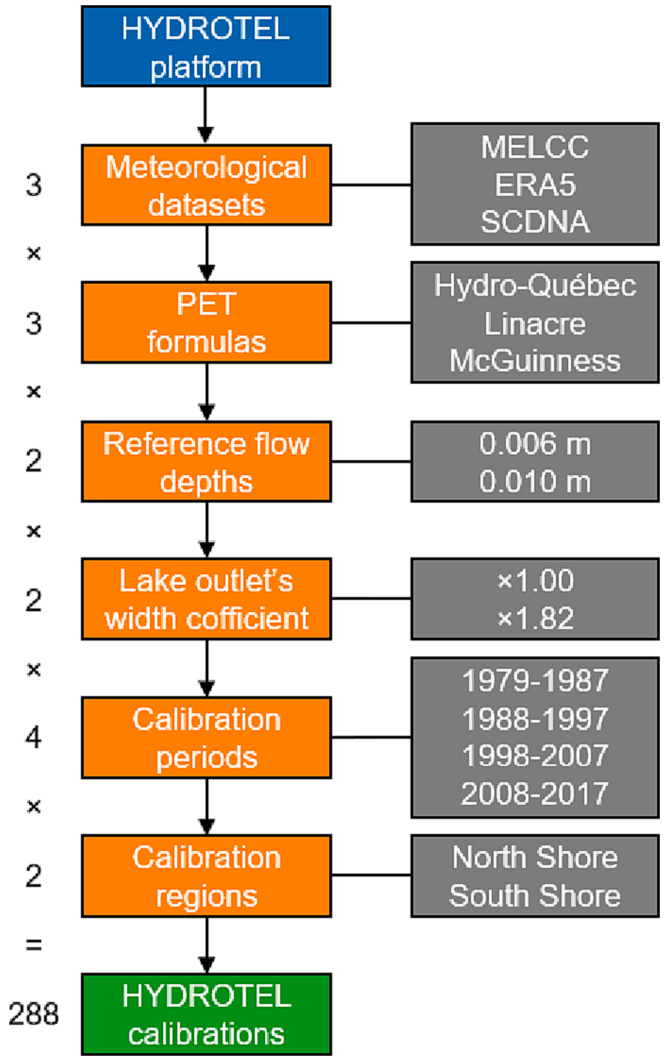

Fig. 2. HYDROTEL calibration procedure.

0.010 m), lake's outlet width multiplicative constant (1.00 and 1.82), PET formulas (Linacre, Hydro-Québec and McGuinness), calibration period (1979–1987, 1988–1997, 1998–2007 and 2008–2017), and split between the North and South shores were calibrated following the procedure described in the previous two sections, resulting in a total of 288 different HYDROTEL calibrations, with 144 calibrations for each of the North and South shores.

While the North and South shores were calibrated independently, they were ultimately combined, resulting in 144 possible simulations to be averaged. To ensure a good representation of both North and South shores of the St-Lawrence River, two thirds of the available stations in both areas were randomly selected for the calibration, and the remaining third for the validation. This random selection ensured that only hydrometric stations with at least 5 years of records over the current calibration period (e.g., 2008–2017), and was kept for all combinations of HYDROTEL models using the same period.

2.5. *Proposed methodology*

The proposed methodology is presented in the flow chart in Fig. 3 and described in the following subsections.

2.5.1. *Baseline: The optimal interpolation*

Optimal interpolation (OI) is a proven data assimilation technique with similarities to kriging (Tabios and Salas, 1985) used in the fields of

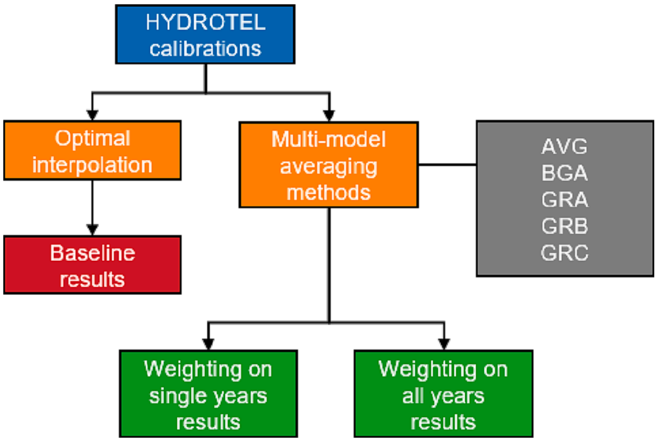
HYDROTEL calibrations
Optimal interpolation
Multi-model averaging methods
AVG
BGA
GRA
GRB
GRC
Baseline results
Weighting on single years results
Weighting on all years results

Fig. 3. Flow chart of the proposed methodology.

meteorology (Fortin et al., 2015) and hydrology (Lachance-Cloutier et al., 2017). When dealing with streamflow, the OI aims at combining available local observations with simulations from a distributed hydrological model (referred to as the background field). At each hydrometric station, the difference between the observations and the simulations is first evaluated. A log transformation is typically applied to avoid obtaining negative streamflow. The OI technique consists in exploiting the spatial correlation of the error to interpolate its value on all sites of interest (i.e., ungauged basins). For each site, corrected values are obtained by:

$$\widehat{z_e} = m_e + \sum_{i=1}^{N} w_i[z_i - m_i] \quad (2)$$

where $\widehat{z_e}$ and m_e are respectively the corrected and the background field values at an estimation site *e*, and $[z_i - m_i]$ is the difference between the observation and the background field at a given hydrometric station *i*. The weights (w_j) are obtained by finding the solution to the system of linear equations:

$$\left[\mathbf{B} + \mathbf{I}\frac{\sigma_o{}^2}{\sigma_b{}^2}\right]\mathbf{W} = \boldsymbol{b} \quad (3)$$

where **B** represents the correlation matrix of the error of the background field between the reference sites *i* and *j*, **I** is a diagonal unit matrix, σ_o^2 and σ_b^2 are the variance of the observation and background field error respectively, **W** is the vector of the weights for each reference sites and ***b*** is a vector of the correlation between the reference sites and the estimation sites «*e*». The ratio $\frac{\sigma_o^2}{\sigma_b^2}$ was fixed to 0.15 following trial and error calibration.

For the daily time series of streamflow, the OI analysis is independently conducted at each time step. The initial calibration of the distributed hydrological model provided by the DEH was post-processed using this methodology and will serve as the benchmark to evaluate the proposed methodology.

2.5.2. *Multi-model averaging methods*

A total of five multi-model average methods were tested in this study and are summarized in Table 3. These different methods allow estimating the optimal set of weights (except for the Simple Arithmetic Average; AVG method which simply provides equal weights to all simulations) for all 95 individual HYDROTEL simulations to combine them together. This multi-model averaging approach aims at combining the different simulations' strengths to outperform each individual simulation. Arsenault et al. (2015) and Wan et al. (2021) showed that weighting methods generally perform better than the individual hydrological model simulations.

The simple arithmetic average (AVG) simply consists in assigning

Table 3
Multi-model averaging methods tested in this study.

Acronym	Method description	Reference	Weights sums to unity	Negative weights possible	Bias correction
AVG	Simple Arithmetic Average	–	Yes	No	No
BGA	Bates Granger Averaging	Bates and Granger (1969)	Yes	No	No
GRA	Granger-Ramanathan A	Granger and Ramanathan (1984)	No	Yes	No
GRB	Granger-Ramanathan B	Granger and Ramanathan (1984)	Yes	Yes	No
GRC	Granger-Ramanathan C	Granger and Ramanathan (1984)	No	Yes	Yes

equal weights to each simulation. The Bates and Granger Averaging (BGA; Bates and Granger, 1969) method relies on the assumption that the simulations are unbiased and that their errors are uncorrelated. The BGA method computes weights (W_{BGA}) as follows:

$$W_{BGA} = \frac{\left(\frac{1}{\sigma_i^2}\right)}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}$$

where σ_i refers to the variance of the simulated streamflow from simulation *i*.

Finally, the Granger-Ramanathan methods A, B and C (Granger and Ramanathan, 1984) methods were employed in this study. Method A (GRA) is based on the ordinary least squares (OLS) algorithm and computes its weights (W_{GRA}) as follows:

$$W_{GRA} = \left(Qsim^T \bullet Qsim^T\right)^{-1} \bullet Qsim^T \bullet Qobs$$

where *Qobs* and *Qsim* are the observed vector and simulated matrix of streamflow, respectively. Variant B (GRB) is similar to the GRA, but the OLS algorithm is constrained to ensure that the sums of the weights are equal to unity. Variant C (GRC) is unconstrained, but it incorporates a bias-correction of the average streamflow using a constant term.

Initially, the Bayesian Model Averaging (BMA; Neuman, 2003) was considered in this study. However, it was found that it was unable to converge to an acceptable set of weights when using all 144 HYDROTEL simulations. Additionally, due to its iterative nature, the BMA's long computing time made it impractical to use throughout the analyses.

Two different metrics were used to evaluate the performance of the various multi-model averaging methods: the KGE (see Equation (1), and the normalized root-mean-squared-error (NRMSE) of the daily annual maxima streamflow (Qx1day). While the KGE provides a measure of performance across the whole hydrograph, the NRMSE Qx1day metric aims at larger values which are more closely related to the values typically used in a flood frequency analysis. However, the NRMSE Qx1day does not take the timing of the events into consideration.

3. Results

3.1. Calibration results

The first step was to generate an ensemble of HYDROTEL simulations with as much variability as possible, to maximize the flexibility of the multi-model averaging approaches. In total, 144 simulations were generated by combining various meteorological data sources and model physiographic parameters, as illustrated in Fig. 2. The calibration results for each basin (y-axis) and each model setup (x-axis) are presented in Fig. 4. Most cases generate acceptable and generally good-quality KGE scores, meaning that the calibrations performed well on those cases. Calibration KGE scores are acceptable for most basins considering the regional calibration method employed, which sacrifices basin-specific skill for overall robustness over the entire domain. A clear pattern is also seen, wherein every 16 simulations, a significant change in performance is observed. This structure reflects changes in the input meteorological data, where the DCAQ, ERA5 and SCDNA data appear in

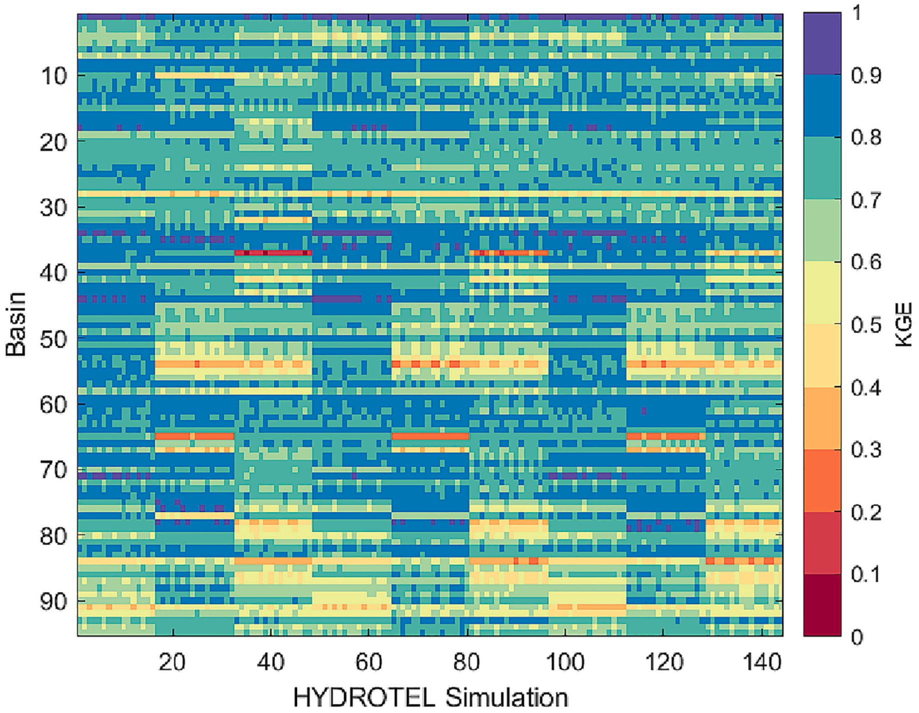

Fig. 4. Calibration KGE results for the 144 simulations over the 95 basins over the study domain.

succession. The selection of a meteorological dataset for calibration is the element that produced the most variability within the results, as can be seen by this repeating cycle of 16 simulations. The parameters related to the HYDROTEL physiographic setup were less impactful and only generated small differences within the simulated hydrographs.

3.2. *Multi-model averaging using all basins*

Once the 144 calibrations were produced, the simulated hydrographs were generated and combined using a variety of multi-model averaging algorithms. Using a Leave-One-Out Cross-Validation (LOOCV), each of the 95 basins was considered pseudo-ungauged and the simulated flows were combined for the pseudo-ungauged basin. The KGE and NRMSE Qx1day were computed and reported in Fig. 5. First, the optimal interpolation (OI) allows a significant gain in performance compared to the raw HYDROTEL simulations (HYD), both in terms of higher KGE scores and lower NRMSE values for Qx1day. Furthermore, the multi-model combinations were not able to improve upon the raw simulations as significantly as the OI, but some small improvements could be noted over the raw simulations for all but the GRC method. However, either no gains or only marginal gains were obtained, no matter what multi-model averaging method was used. Results are also presented spatially for the OI, simple average (AVG) and the GRA weighting method (which was considered the best multi-model averaging method here) in Fig. 6. However, no striking pattern emerges from these results: all multi-model averaging methods seem to perform similarly over the study domain.

3.3. *Multi-model averaging using a year-by-year approach*

The next step was to attempt to extract more information from the data by performing multi-model weightings for each year instead of on the entire period of the available datasets. Results are shown in Fig. 7, and spatially distributed results for methods OI, AVG and GRA again are shown in Fig. 8. It is clear from Fig. 7 that there is a significant additional gain in performance when computing weights year-by-year, giving more flexibility to the multi-model averaging algorithms to fit the optimal weights for the given year and not forcing it to compromise to be adequate for all years.

In the year-by-year approach, some multi-model averaging methods clearly respond better than when all data was pooled together, notably the GRA and GRB methods. GRA was slightly better in this respect and was used for the rest of the study. It can be seen in Fig. 8 that GRA performed better than the raw simulations for 67 and 63 out of 95 basins for the KGE and NRMSE Qx1day, respectively, while the OI performed better than the raw simulations for 70 and 66 basins out of 95 for the KGE and NRMSE Qx1day, respectively. Therefore, the GRA seems to perform at least as well as the OI for this study area and hydrological model simulations, both in terms of overall flow simulation (KGE) and peak flows (NRMSE Qx1day).

4. Discussion

4.1. *Improvement using the optimal interpolation*

Optimal interpolation (OI) aims to improve the simulation results overall, without regard to discontinuities in the generated hydrographs. Since the domain is composed of basins of all sizes, peak flows occur at different dates even when driven by the same meteorological conditions due to flow routing in larger basins. OI aims to apply spatial corrections at each time step, disregarding the state of specific hydrographs. This means that a compromise must be made during the interpolation, whereby some peak flows are artificially reduced, and others are increased, simply due to the timing of nearby basins that are utilized in the optimal interpolation. Nonetheless, the method provides good results on average, which is not surprising given that for this purpose, OI is a statistically optimal method. The OI simulations were consistently better than the raw simulations for a majority of basins (70% to 75% of basins), both in terms of KGE and NRMSE Qx1day (as seen in the top of the boxplots of Figs. 5 and 7). This method is therefore recommended as a good approach for processing distributed model flows, as there is a significant gain in overall and peak flow estimations.

During this study, an attempt was made to combine several simulations post-processed by the OI (not shown) with the multi-model averaging methods proposed. These tests showed that there was a gain in

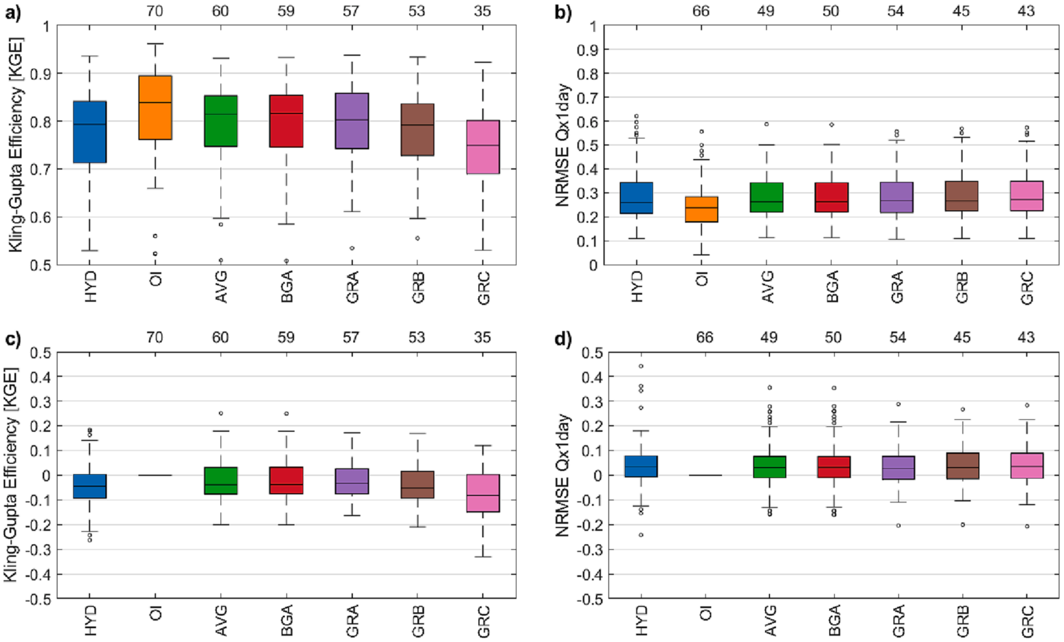

Fig. 5. Leave-One-Out Cross-Validation KGE (a, c) and NRMSE Qx1day (b, d) for the 95 basins considered as ungauged by pooling data from all years and for all basins together during multi-model weights calibration. The top panels (a, b) present boxplots of the results, while the bottom panels (c, d) present the difference between the results and the OI which serves as the baseline. The values above each boxplot show the number of basins (out of 95) for which the specific method performed better than the raw HYDROTEL (HYD) simulations.

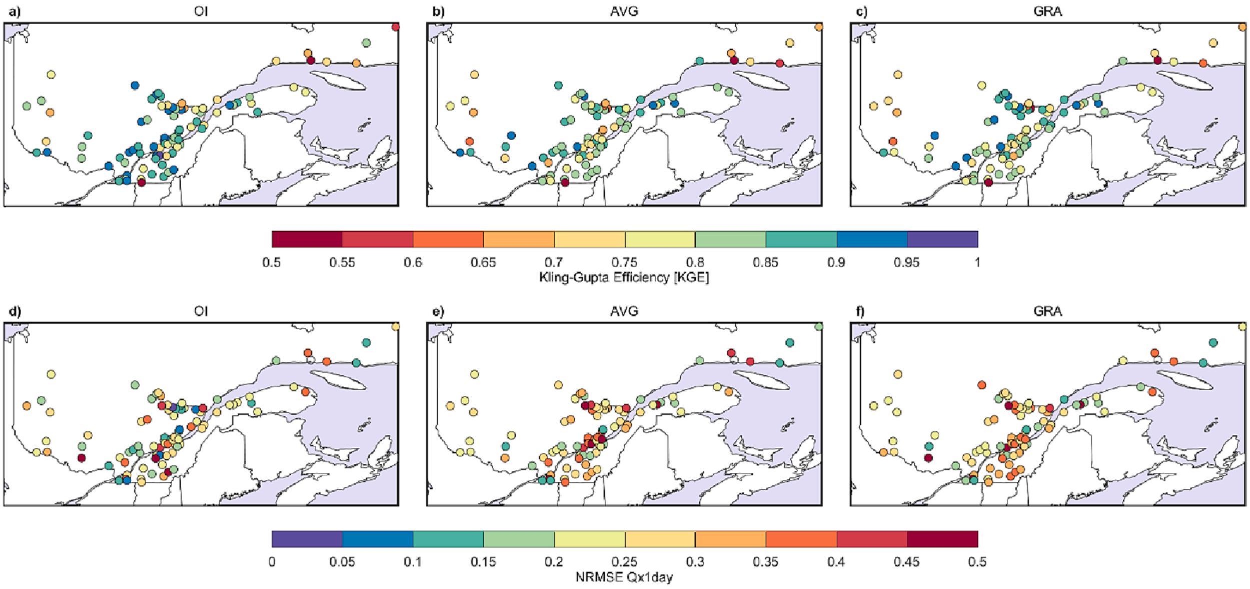

Fig. 6. Spatial distribution of the LOOCV values for KGE (a, b, c) and NRMSE Qx1day (d, e, f) for the tested methods OI (a, d), AVG (b, e) and GRA (c, f). GRA weights are computed using all available years.

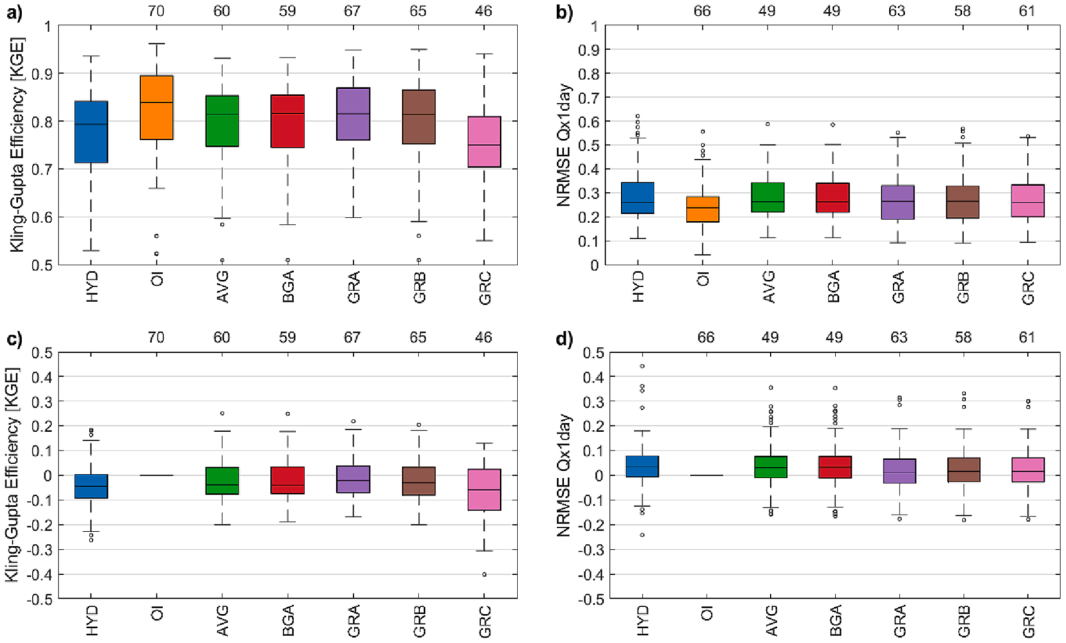

Fig. 7. Same as Fig. 5, but for all the multi-model methods use a year-by-year weighting instead of a unique one for the entire period.

robustness which was on average better than the individual simulations with OI. The robustness comes from the fact that it is not possible to select the “right simulation” from the start, when dealing with ungauged basins. The gains were only marginal, however, regardless of the weighting method used, with similar results obtained with the AVG method. One disadvantage of processing OI simulations in a multi-model framework is that multi-model weighting schemes maximize performance by using each simulation’s strengths to reduce the overall error. This requires some flexibility and variety in the ensemble of members that are used during the weighting. However, OI applies corrections in such a way that the resulting hydrographs are much more similar, leaving fewer degrees of freedom to the multi-model averaging methods to perform efficiently. Therefore, adding more OI simulations to the multi-model averaging process only produces asymptotic performance gains. In an operational setting, if multiple OI simulations are available, it is recommended to use a simple average of all these simulations to make gains in robustness, but the gains in performance are expected to be marginal. Therefore, the multi-model approach can provide better results than OI due to the fact that more information is available in the entire process for multi-model methods than what OI can use.

4.2. Comparing the different multi-model averaging methods

When using the multi-model averaging methods on the 144 simulations, it was found that significant improvements in simulation accuracy could be observed, even with the simplest averaging methods. This is in line with similar studies using lumped hydrological models (Arsenault et al., 2015). One exception is GRC, which contains a bias

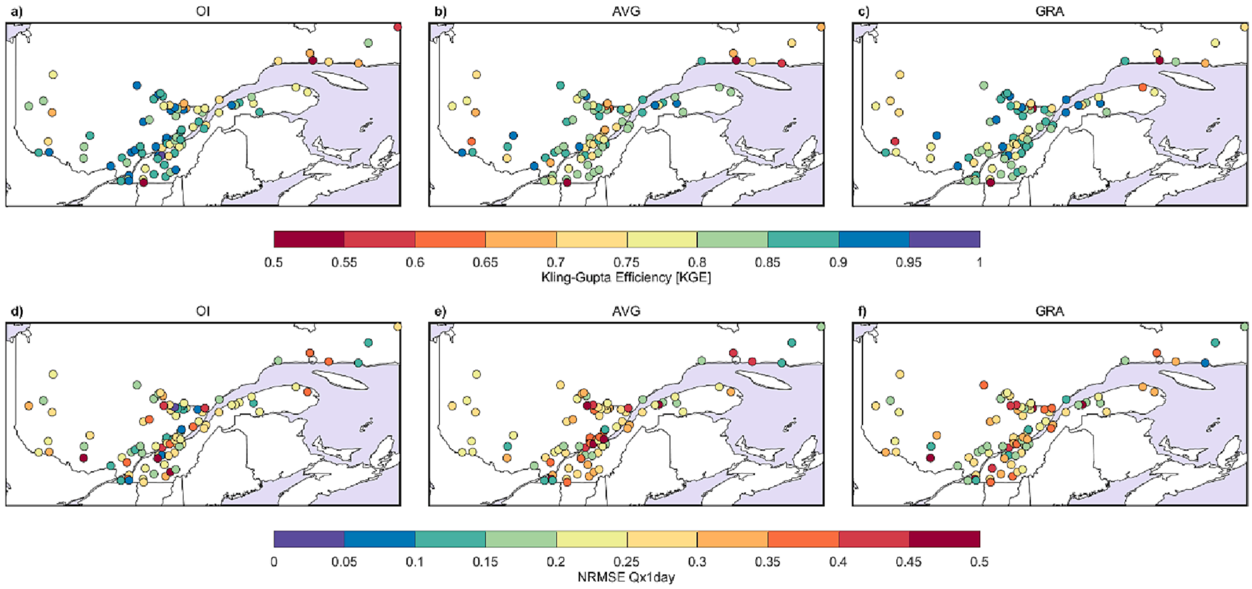

Fig. 8. Same as Fig. 6, but the GRA method uses a year-by-year weighting instead of a unique one for the entire period.

element that is not specific to each site. This constant bias ends up introducing errors to all basins, except for the basins whose flows are near the ensemble average. With the exception of GRC, the other results are in line with what was found previously in the literature, where the AVG and BGA methods perform generally worse (lower Nash-Sutcliffe values, larger biases) than with the GRA and GRB methods (Wan et al. 2021).

Furthermore, methods such as GRA and GRB performed better than the others, especially so when the weights were computed on a yearly basis. This approach of computing one set of weights per simulation year (and shorter periods, as described in the following section) allowed the weighting algorithms to adapt to specific hydrometeorological conditions over the domain and thus allow for more accurate weighted simulations. This methodology could be criticized for certain contexts as it goes against the prevailing notion that models should be as good as possible in all conditions, and the current methodology implements a form of overfitting that would not be applicable for simulation of longer time series. However, when considering the goal of performing a flood frequency analysis (especially over ungauged basins), overfitting is not necessarily an issue - we argue that it is an advantage. The goal is to have the most optimal streamflow simulation with as little uncertainty as possible, and by simulating these hydrographs on a year-by-year basis with the most precise information as possible for each year, the resulting hydrograph is of higher quality than if the weighting was performed with a single set of weights for the whole period representing the long-term compromise. This method is thus applicable in this study since the models and methods are not used for forecasting, but for historical simulations. Therefore, overfitting is not an issue in this study.

4.3. *Improvement by computing weights over shorter periods*

As demonstrated with the comparison of boxplots from Figs. 5 and 7 and maps from Figs. 6 and 8, there is a clear gain in performance by averaging weights over a shorter period. With significant gains obtained by going from the weights for the whole periods to a year-by-year basis, it was of interest to evaluate the limit of using an even shorter period, and to determine at which point overfitting would start occurring. Using the GRA method (which provided the best results in this study), six different periods were tested for the computation of the weights: year-by-year, season-by-season (i.e., every three months, starting with January, February, and March), month-by-month, 2-weeks-by-2-weeks, week-by-week, and day-by-day. Results shown in Fig. 9 indicate that by decreasing the period up to the month-by-month period, there is an increase in performance for both tested metrics. By going beyond the

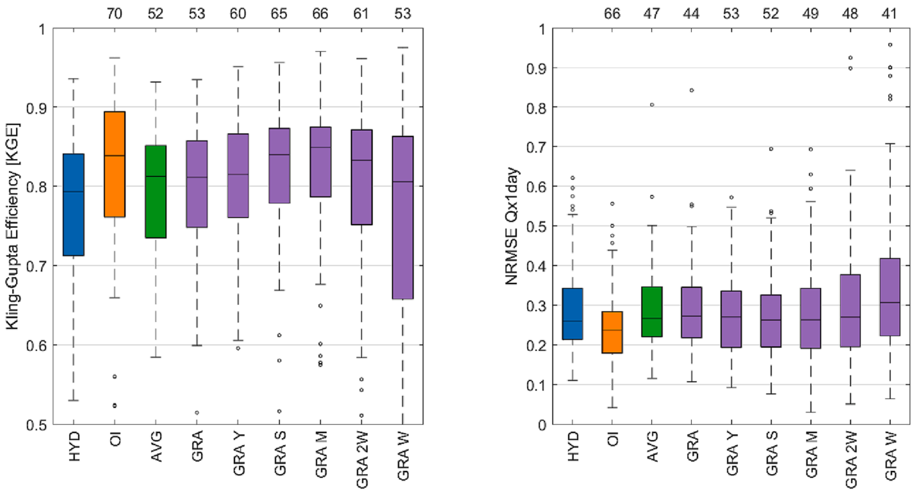

Fig. 9. Leave-One-Out Cross-Validation KGE (left panel) and NRMSE Qx1day (right panel) for the 95 basins considered as ungauged by pooling data for all basins together during multi-model weights calibration. Box plots in purple represent the GRA averaging method with the pooling of the data performed at various periods: year-by-year (Y), season-by-season (S), month-by-month (M), 2-weeks-by-2-weeks (2 W), and week-by-week (W). The values above each boxplot show the number of basins (out of 95) for which the specific method performed better than the raw HYDROTEL (HYD) simulations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

year-by-year period, the multi-model averaging method also starts to perform similarly to the OI method. However, when going beyond the month-by-month period, weights are starting to be overfitted, resulting in poorer performance when conducted in a cross-validation context like the one in this study. For instance, the day-by-day period (not shown in Fig. 9 since all the values were below the 0.5 KGE value) is strongly impacted by overfitting since there 144 weights to fit 95 values only.

To determine if a pattern exists in the improvement observed in Fig. 9, we compared these results with basin descriptors presented in Table 1. However, we found no correlation between any of the basin descriptors and the KGE or NRMSE Qx1day results obtained in this study, for all periods used to compute the weighting (i.e., year-by-year up to week-by-week). Notably, Arsenault et al. (2015) and Wan et al. (2021) conducted geographic analyses and identified low-performance basins located in arid regions with total annual precipitation below 500–600 mm/year. This could be attributed to the hydrological model being more suitable for humid climates and areas with strong snow accumulation. In our study, the driest basins had an average annual precipitation of around 800 mm/year, which is significantly above the threshold suggested by Arsenault et al. (2015) and Wan et al. (2021). This finding may explain why almost all basins exhibit improved results from the computation of weights over shorter periods.

4.4. Using more diverse hydrological models

When it comes to adding more flexibility to the sample of simulations, different formulations and models for the different hydrological processes can also be used. Unfortunately, the flexibility offered by the distributed model used in this study was limited to different potential evapotranspiration formulas in this regard. The focus has been put on different types of calibration and datasets to obtain the desired flexibility in our sample.

It is expected that different formulations and models for the hydrological processes or even different hydrological models altogether could achieve the desired results as well. To test this hypothesis, additional tests were conducted by calibrating three lumped hydrological models only using the MELCC dataset and the Oudin PET formula (Oudin et al., 2005):

- GR4J with the CemaNeige snow module (Perrin et al., 2003; Valéry, 2010)
- HMETS (Martel et al., 2017)
- MOHYSE (Fortin and Turcotte, 2007)

Even though the use of lumped models has its own shortcomings when it comes to dealing with prediction in ungauged basins (PUB), such as the need to regionalize parameters, they still allow us to glean some insights on the benefit of adding completely different hydrological model structures. The multi-model average methods were applied on all three lumped hydrological models as well as the raw HYDROTEL simulation (four simulations in total). It can be seen in Fig. 10 that greater improvements were obtained in comparison with the different simulations from HYDROTEL, exceeding the OI in terms of performance. This suggests that HYDROTEL did not provide sufficient flexibility in its structure, limiting the hydrograph variability to that of the various datasets and calibration schemes. Thus, combining different hydrological models (both lumped and distributed), or using a model with more flexibility in its structure could allow it to reach the desired gains in performance.

Hydrological models have varying strengths and weaknesses, and no single model outperforms others in all basins. In this study, three lumped hydrological models were tested, and GR4J, HMETS, and MOHYSE achieved the best KGE values respectively on 82.11 %, 16.84 % and 1.05 % of the basins. As suggested by Wan et al. (2021), combining multiple hydrological models calibrated with different objective functions can lead to more efficient and robust results, as observed in this study with significant improvement obtained using multiple HYDROTEL calibrations (Figs. 5 to 9) and different hydrological models (Fig. 10). Arsenault et al. (2015) also demonstrated that even models that yield poorer results across all basins should still be included, as they can contribute to the multi-model averaging performance. They showed this using a multi-objective optimization approach aimed at maximizing the objective function and minimizing the number of simulations in the multi-model averaging, with even the worst-performing model (also MOHYSE) making regular contributions.

It should be noted that the OI used as a baseline in this study is conducted using a regional calibration from HYDROTEL that is not specifically calibrated at each local site. On the other hand, the lumped hydrological models used in this additional analysis are calibrated locally at each site. While the results showed improved performance by using these lumped hydrological models over the OI, it can be expected that the simulation at ungauged sites from a regional model like HYDROTEL are likely to be more robust than those from lumped models that would require to be regionalized.

As demonstrated in the previous section, additional gains in performance can be obtained by computing weights over a shorter period (e.g., year-by-year, season-by-season, or month-by-month). It can be noted that improvement is obtained even up to the day-by-day period. This is likely due to the fact that there are only 4 hydrological models being combined, resulting in 4 weights for a minimum of 95 values at the day-

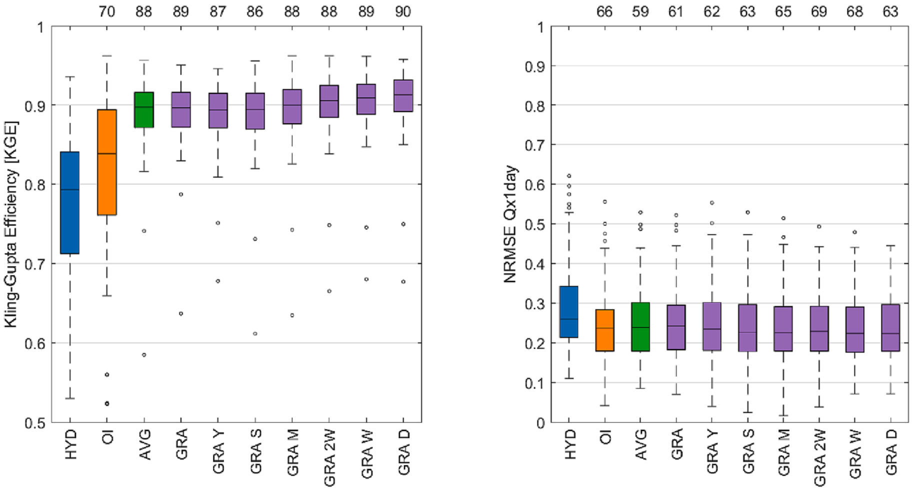

Fig. 10. Same as Fig. 9, but using the four hydrological models.

by-day period, avoiding overfitting. However, the gain in performance when combining the four hydrological models (Fig. 10) compared to the combination of all 144 HYDROTEL simulations (Fig. 9) is much smaller. This is likely because the four hydrological models are already providing most of the flexibility needed, reducing the potential gain in performance to be made by computing weights over shorter periods. This finding suggests that the flexibility needed to increase the multi-model averaging performance can be obtained in different manners, namely:

1. Increasing the number of calibrations from a given hydrological models by using alternative parameters, meteorological datasets, or model structure.
2. Increasing the number and variety of hydrological models (both lumped and distributed).
3. Reducing the period at which to compute the weights of the multi-model averaging.

4.5. *Flood frequency analyses comparison*

Considering that the main objective of this study was to generate the most accurate historical streamflow pseudo-observations for conducting a flood frequency analysis in ungauged basins, additional analyses were conducted to evaluate the proposed methodology. To carry out these analyses, four different basins were selected, including one large and one small basin located on the North shore, as well as one large and one small basin on the South shore. All four selected basins had complete time records between 1979 and 2017. The best overall multi-model method (GRA), which uses a year-by-year period (GRA Y) for computing the weights, was compared with OI and observations.

Table 4 shows the results of KGE and NRMSE Qx1day values for the four chosen basins. These four basins exhibit a similar pattern to the boxplots illustrated in Fig. 9. Improvements in performance can be observed up to the season-by-season or month-by-month period, followed by a decline. Although this trend is apparent in the KGE results, some differences are noticeable for the NRMSE Qx1day, particularly in the basin 061022.

The simulations of the annual maxima series (Qx1day) were first compared, as these are directly used by the flood frequency analysis. The results are shown in Fig. 11. However, it is important to note that not all basins provided satisfactory results for either the GRA S, the OI, or both. Overall, it was observed that both methods provide good estimations of observed streamflow in a LOOCV context. GRA Y does not typically lead to overestimations of Qx1day, while OI tends to slightly overestimate the results for the larger basins (Fig. 11-a and b). However, GRA Y underestimates the results for basin 030,282 (Fig. 11-d).

Using the Qx1day time series presented in Fig. 11, flood frequency analyses were conducted and the results are presented in Fig. 12. The Gumbel distribution was used and its parameters were adjusted using the maximum likelihood approach. In all cases, both methods are able to provide good results mostly within the 95% confidence intervals of the observations (except for the lower return periods in basin 061,022 – Fig. 12-b). With respect to the selected basins, the GRA Y method tends to provide results closer to observations while OI tends to slightly overestimate the values while remaining within the confidence interval boundaries, in line with results obtained in Fig. 11. However, this is not necessarily the case for all basins where different behaviors are observed. The aim of this analysis was simply to validate if the estimations provided by both methods led to satisfactory results for a flood frequency analysis.

As a reminder, the results obtained for the multi-model approach could be significantly improved if more structural variability was possible for HYDROTEL, as demonstrated with results and discussion in section 4.4 and Fig. 10. Thus, while already providing satisfactory results, this method can still be improved beyond what is obtained in this paper.

4.6. *What is the best method to use?*

In this study, raw simulations were improved through two approaches: the optimal interpolation (OI) as implemented by Lachance-Cloutier et al. (2017) was used as the reference case, and the multi-model averaging method. Both methods have advantages and shortcomings that need to be mentioned.

First, it is important to note that this study implements a two-step approach to estimating flows at ungauged basins on a regional scale. Typically, regionalization methods perform either (1) model parameter regionalization at ungauged sites, (2) distributed model regional simulations or (3) for hydrological indicators, statistical methods such as regression and kriging have also been proposed. Model parameter regionalization requires modelling each basin individually and simulating them with a hydrological model, which is parameterized using “donor” basin parameters. This method introduces high levels of uncertainty and introduces major spatial discontinuities between neighboring basins depending on their physical characteristics (Arsenault and Brissette, 2014; Razavi and Coulibaly, 2013). The distributed hydrological model method performs reasonably well (in this study raw HYDROTEL simulation), but it is also a compromise solution where the model is made to simulate flows adequately on all sites by allowing loss of skill on some of the basins (Kumar and Samaniego, 2013). Finally, statistical methods to regionalize peak flows have been proposed, but suffer from the same limitations as model parameter regionalization in terms of uncertainty (Perez et al., 2019). Recently, deep learning approaches have started being used to estimate streamflow at ungauged sites, but these have been applied in large-scale applications of multiple basins and not on a regional domain with a wide array of basin scales (Arsenault et al. 2023, Kratzert et al. 2019). A combined approach was implemented by Wang et al. (2023) where a distributed model had its parameters regionalized using random forest methods in China, with success. Therefore, the two methods presented in this study are post-processing steps that take the result from a distributed regional model and improve them by reducing the effects of the compromise of regional calibration.

The OI is a simple data assimilation technique that has multiple advantages: it is a statistically optimal method based on known theory, provides smooth spatial corrections by taking all spatial components into consideration, can be conducted on each day independently, and requires low computational power. However, this method also has various shortcomings that should not be overlooked: breaking down the

Table 4
KGE and NRMSE Qx1day results for the four selected basins.

	KGE				**NRMSE Qx1day**			
	041,902	061,022	023,303	030,282	041,902	061,022	023,303	030,282
GRA	0.770	0.887	0.876	0.784	0.243	0.245	0.194	0.322
GRA Y	0.775	0.867	0.865	0.807	0.295	0.279	0.166	0.239
GRA S	0.848	0.897	0.869	0.823	0.244	0.327	0.196	0.177
GRA M	0.852	0.877	0.861	0.815	0.293	0.396	0.198	0.257
GRA 2 W	0.850	0.853	0.844	0.800	0.313	0.428	0.251	0.248
GRA W	0.831	0.812	0.801	0.745	0.305	0.511	0.345	0.334

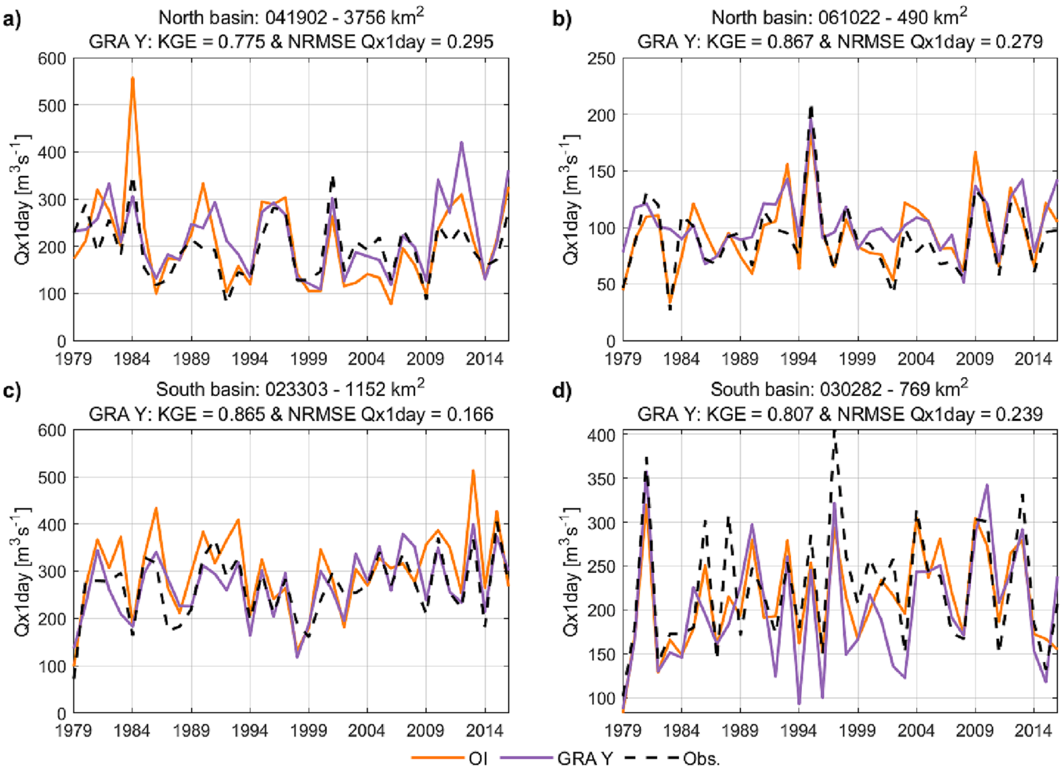

Fig. 11. Qx1day values for observations (black), OI (orange) and GRA Y (purple) over two basins on the North shore (top) and two basins on the South shore (bottom). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

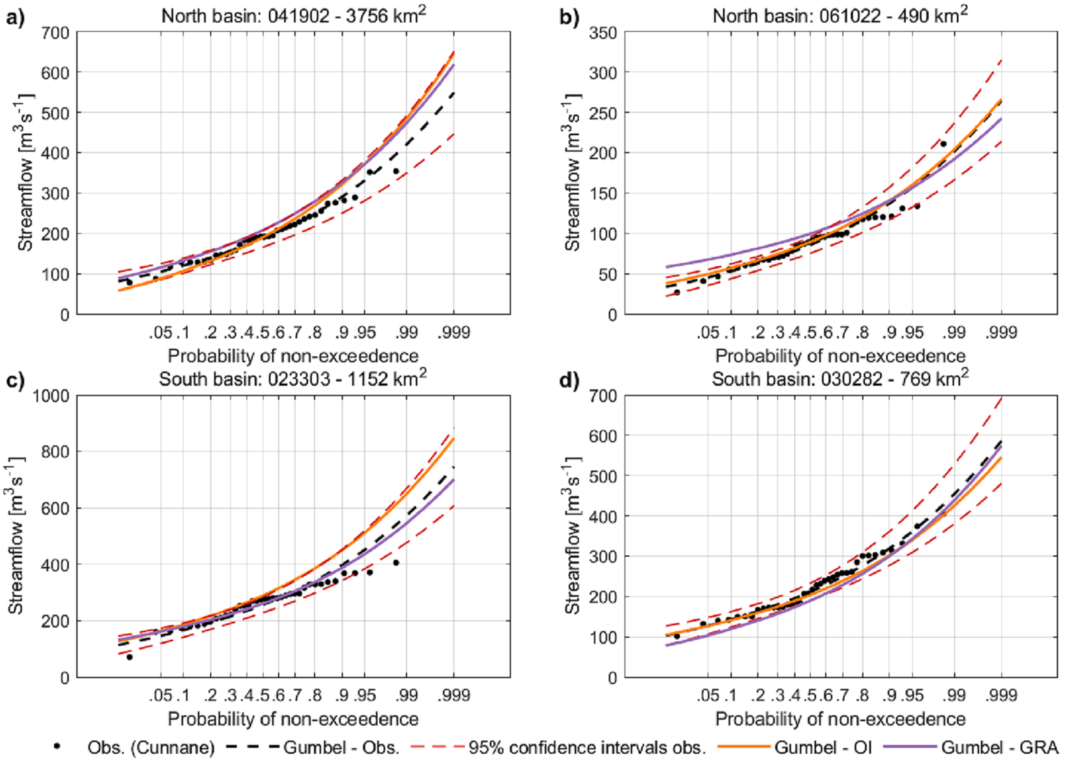

Fig. 12. Flood frequency analyses using the Gumbel distribution on the Qx1day values for observations (black), OI (orange) and GRA (purple) over two basins on the North shore (top) and two basins on the South shore (bottom). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

spatial structure of data, discontinuities in the data, smoothing of maximum events, does not provide correction for sites where the closest observation is too far, and generation of inconsistent reaction times, since the method does not guarantee mass balance preservation. Essentially, OI can lead to discontinuity in the hydrograph because the error field is interpolated each day based on the differences between the observations and the model simulations at the gauged sites. The error field is then spatially applied to all simulated discharges in the modelled region. However, this can cause problems when the hydrological model output is dependent on exogeneous variables such as basin size. Indeed, a large basin could have a delayed response (multiple days) to a precipitation event, whereas a small neighboring (or nested) basin would react on the same day. Therefore, the error field would correct the large peak flood at a later date which could then impose a large (yet unnecessary) correction on the smaller basin. Also, significant discontinuities can be introduced since the closest observed hydrometric stations can vary over time when some enter service and others are shut down. While a technique like OI provides significant gains in terms of performance over the raw simulation from a hydrological model, its various shortcomings are undesirable when conducting a flood frequency analysis, especially on ungauged basins.

The multi-model averaging method has the advantage of not having the shortcomings from a typical data assimilation method such as OI, since the time-series are continuous (except during transitions when weights change) and depend on blending model outputs, which does not produce temporal inconsistencies. Contrary to the OI, sites far away from observations will also be corrected since the same weights will be applied to all simulations, regardless of their location. Also, the multi-model method does not perform any smoothing of time series since the same weights are applied each day, which means the timing of hydrographs' peaks is preserved. Furthermore, additional degrees of freedom can be relatively easily added to improve its performance by adding more simulations and/or by computing the weights over a shorter period. However, temporal breaks can end up introduced when using weighting based on shorter periods (e.g., year-by-year or month-by-month). In an operational context, the challenge lies in the ability to generate a limited number of simulations that are sufficiently different from one another to combine their strengths, while typically using the same hydrological model used by the agency or organization to limit overhead related to maintenance and training on multiple models. There are other shortcomings what should not be overlooked as well: this method does not adjust to the information available at measurement sites, it has no guarantee of mass balance preservation, it can propose negative weights or weights not adding up to 1, which can have a dubious physical meaning and it scores highly on average values (KGE), but can be less efficient on peak values.

It was hypothesized that accessing a wider variety of simulations would give more flexibility to the multi-model averaging scheme, which would then allow for more accurate simulations. By design, this method would not create discontinuities except at the beginning of each period due to the weights being recomputed if a shorter period than the full period is used. Currently, the multi-model approaches using the 144 HYDROTEL simulations achieve similar performance levels as OI (and better results when the period is reduced). While a relatively large number of simulations were used, they were mostly based on various climatological datasets and different calibrations. Ideally, changes in model structure (e.g., snowmelt, routing, and infiltration modules) could have provided more variety and potentially helped improve the results even further, as suggested by the results of the previous section. However, HYDROTEL's internal mechanics could not be modified in this study.

A potential advantage of the multi-model average method over the OI that was not investigated in this study would be the capacity to apply the weighting in a context where observations would not be available. For instance, in a climate change impact study using future simulations, pre-computed weights for the multi-model averaging could still be used.

4.7. Limitations

This study was performed on a regional distributed model and the results were evaluated on pseudo-ungauged sites. However, the gauged basins are typically relatively large (the smallest was 44 km^2, and the median was almost 1000 km^2). It is therefore possible that the results do not scale well with much smaller size basins since this has not been tested. However, this is also true for the OI method application, therefore no strong conclusion on the generalizability of either method can be made. The results seem to point to the methods being applicable without any correlation to basin size, but this cannot be proven at this time. This should be tested in future studies, where very small gauged basins would also be included.

Another limitation is that of observed flow uncertainty. Observed flow measurements are known to be highly uncertain and noisy, which can lead to errors that propagate throughout the process. For example, the multi-model weights are computed by minimizing differences between the weighted streamflow and the observed flows. Any errors or biases in the observed flows would then propagate into the multi-model weights and would affect the quality of the simulation on the gauged and ungauged basins. However, by taking weights over longer periods, the epistemic uncertainties should be abated somewhat, leading to a more robust signal. On the other hand, OI generates error fields that are computed each day, so it could be prone to more uncertainty emanating from the observed flow random errors. Therefore, obtaining a good estimate of observed flow uncertainty would help finding more robust methods that could take this uncertainty into consideration during the calculation of weights.

Finally, this study investigates the performance of the simulation methods using two metrics, KGE and NRMSE Qx1day. It is important to note that the hydrological model was always calibrated using KGE, and therefore is expected to reproduce overall hydrographs well, with a smaller emphasis on the peak flows. It could have been possible to calibrate on an objective function that weights the peak flows more heavily in order to maximize performance of the Qx1day metrics, but this would have been at the expense of the other flow regimes. Therefore, the calibration on KGE is a compromise solution that can still provide estimates of Qx1day with a good level of confidence.

5. Conclusion

Estimating streamflow and peak flows at ungauged sites is still a challenge today due to the various sources of uncertainty and highly spatially heterogeneous nature of hydrological processes. In this study, we attempted to provide the most precise hydrographs possible on a historical period and using a distributed hydrological model, in such a way that it would be possible to process the hydrological model simulations to obtain accurate hydrographs on the entire region. Using a leave-one-out cross-validation methodology applied to 95 basins, we were able to show that the HYDROTEL model simulations could be substantially improved using optimal interpolation (OI). However, some drawbacks associated with this method led us to attempt a second method. The alternative method was the application of multi-model averaging techniques to a large variety of hydrological model simulations. Some multi-model methods performed better than others, but the GRA method showed similar performance to that of OI without any of the drawbacks. This means that the historical period hydrographs could be generated with more confidence, which can then lead to better estimates of peak flows. This is an important step towards modeling peak flows for flood frequency analysis. Given the fact that these results were all obtained considering the basins as ungauged, there is a high confidence that the method can be applied over the entire domain covered by the HYDROTEL model. This also means that estimating flood risk at these ungauged basins can be made with higher confidence.

Interestingly, the multi-model averaging approach does not suffer from the scale issues faced by the OI method. Indeed, OI generates correction factors for each time step independently from one another, meaning that errors computed from small gauges with reactive basins will propagate to larger basins with slower reaction time, therefore impacting the peak flow timing and amplitude. The multi-model averaging method does not have this inconsistency as all weights are computed over longer time horizons and over many basins, but then applied on the simulated flow for the specific basins. Therefore, peak flow amplitude and timing are preserved.

This study also highlighted the fact that multi-model averaging methods require a certain level of variability within the simulations pool to maximize their effectiveness. The HYDROTEL model, although driven with multiple weather forcings, parameters and time horizons, was not able to provide as much structural variability as required and as demonstrated by using the lumped hydrological models. Future research should make use of additional distributed hydrological models to implement regional multi-model simulations with a higher degree of simulation variability.

CRediT authorship contribution statement

Jean-Luc Martel: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Richard Arsenault:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Simon Lachance-Cloutier:** Resources, Supervision, Writing – review & editing. **Mariana Castaneda-Gonzalez:** Resources, Data curation, Writing – review & editing. **Richard Turcotte:** Supervision, Writing – review & editing. **Annie Poulin:** Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: 'Richard Arsenault reports financial support was provided by Ouranos.'

Data availability

Data will be made available on request.

Acknowledgements

The authors would like to thank the teams at the Direction de l'Expertise Hydrique du Québec (DEH) and Ouranos that made this project possible in the context of the INFO-Crue research program. This project was funded by the INFO-Crue program, project #702300.

The authors would also like to thank the ECMWF for providing the ERA5 reanalysis data through the Copernicus Data Store at https://climate.copernicus.eu/climate-reanalysis. They would also like to acknowledge the contribution of Dr. Tang and co-authors for making the SCDNA database available at https://zenodo.org/record/3735534.

References

Arsenault, R., Brissette, F.P., 2014. Continuous streamflow prediction in ungauged basins: the effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res. 50 (7), 6135–6153. https://doi.org/10.1002/2013WR014898.

Arsenault, R., Brissette, F., 2016. Multi-model averaging for continuous streamflow prediction in ungauged basins. Hydrol. Sci. J. 61 (13), 2443–2454. https://doi.org/10.1080/02626667.2015.1117088.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J.-L., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001.

Arsenault, R., Martel, J.L., Brunet, F., Brissette, F., Mai, J., 2023. Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. Hydrol. Earth Syst. Sci. 27 (1), 139–157. https://doi.org/10.5194/hess-27-139-2023.

Bates, J.M., Granger, C.W.J., 1969. The Combination of Forecasts. J. Oper. Res. Soc. 20 (4), 451–468. https://doi.org/10.1057/jors.1969.103.

Bergeron, O., 2016. Guide d'utilisation 2016 - Grilles climatiques quotidiennes du Programme de surveillance du climat du Québec, version 1.2, Québec, ministère du Développement durable, de l'Environnement et de la Lutte contre les changements climatiques, Direction du suivi de l'état de l'environnement.

Blöschl, G., Sivapalan, M., Wagener, T., Savenije, H., Viglione, A., 2013. Runoff Prediction in Ungauged Basins: Synthesis Across Processes, Places and Scales. Cambridge University Press.

C3S, 2019. C3S ERA5-Land reanalysis. Copernicus Climate Change Service (C3S).

Diks, C.G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Env. Res. Risk A. 24 (6), 809–820. https://doi.org/10.1007/s00477-010-0378-z.

Exbrayat, J.-F., Viney, N.R., Seibert, J., Frede, H.-G., Breuer, L., 2011. Multi-model data fusion as a tool for PUB: example in a Swedish mesoscale catchment. Adv. Geosci. 29, 43–50. https://doi.org/10.5194/adgeo-29-43-2011.

Fortin, V., Turcotte, R., 2007. Le modèle hydrologique MOHYSE (in French). Université du Québec à Montréal Note SCA7420, 14 pp.

Fortin, V., Roy, G., Donaldson, N., Mahidjiba, A., 2015. Assimilation of radar quantitative precipitation estimations in the Canadian Precipitation Analysis (CaPA). J. Hydrol. 531, 296–307. https://doi.org/10.1016/j.jhydrol.2015.08.003.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., Villeneuve, J.-P., 2001a. Distributed watershed model compatible with remote sensing and GIS data. I: Description of model. J. Hydrol. Eng. 6 (2), 91–99.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., Villeneuve, J.-P., 2001b. Distributed watershed model compatible with remote sensing and GIS data. II: Application to Chaudière watershed. J. Hydrol. Eng. 6 (2), 100–108.

Fortin, V., 2000. Le modèle météo-apport HSAMI: historique, théorie et application., Institut de Recherche d'Hydro-Québec.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecast. 3 (2), 197–204. https://doi.org/10.1002/for.3980030207.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Heo, J.-H., Ryu, G.-H., Jang, J.-D., 2018. Optimal interpolation of precipitable water using low Earth orbit and numerical weather prediction data. Remote Sens. (Basel) 10 (3), 436.

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

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., Cudennec, C., 2013. A decade of Predictions in Ungauged Basins (PUB)—a review. Hydrol. Sci. J. 58 (6), 1198–1255.

Huot, P.-L., 2014. Évaluation de méthodes d'optimisation pour le calage efficace de modèles hydrologiques coûteux en temps de calcul, École de technologie supérieure.

Kim, N.W., Shin, M.J., 2018. Estimation of peak flow in ungauged catchments using the relationship between runoff coefficient and curve number. Water 10 (11), 1669. https://doi.org/10.3390/w10111669.

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424–425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55 (12), 11344–11354. https://doi.org/10.1029/2019WR026065.

Kumar, R., Samaniego, L., Attinger, S., 2013. Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resour. Res. 49 (1), 360–379. https://doi.org/10.1029/2012WR012195.

Lachance-Cloutier, S., Turcotte, R., Cyr, J.-F., 2017. Combining streamflow observations and hydrologic simulations for the retrospective estimation of daily streamflow for ungauged rivers in southern Quebec (Canada). J. Hydrol. 550, 294–306. https://doi.org/10.1016/j.jhydrol.2017.05.011.

Li, J., Heap, D., 2014. Spatial interpolation methods applied in the environmental sciences: a review. Environ. Model. Softw. 53, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008.

Linacre, E.T., 1977. A simple formula for estimating evaporation rates in various climates, using temperature data alone. Agric. Meteorol. 18 (6), 409–424. https://doi.org/10.1016/0002-1571(77)90007-3.

Ly, S., Charles, C., Degré, A., 2013. Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review. Biotechnol. Agron. Soc. Environ. 17 (2), 392.

Martel, J.-L., Demeester, K., Brissette, F., Poulin, A., Arsenault, R., 2017. HMETS—a simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. Int. J. Eng. Educ. 33 (4), 1307–1316.

Martel, J.-L., Brissette, F., Poulin, A., 2020. Impact of the spatial density of weather stations on the performance of distributed and lumped hydrological models. Can.

Water Resources J. / Revue canadienne des ressources hydriques 45 (2), 158–171. https://doi.org/10.1080/07011784.2020.1729241.

McGuinness, J.L., Bordne, E.F., 1972. A Comparison of Lysimeter-derived Potential Evapotranspiration with Computed Values. US Department of Agriculture.

Oke, P.R., Brassington, G.B., Griffin, D.A., Schiller, A., 2010. Ocean data assimilation: a case for ensemble optimal interpolation. Aust. Meteorol. Oceanogr. J. 59 (1SP), 67–76.

Ouarda, T., St-Hilaire, A., Bobée, B., 2008. Synthèse des développements récents en analyse régionale des extrêmes hydrologiques. Revue des sciences de l'eau / J. Water Sci. 21 (2), 219–232, doi: 10.7202/018467ar.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C., 2005. Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J. Hydrol. 303 (1-4), 290–306.

Papacharalampous, G., Tyralis, H., 2022. Time series features for supporting hydrometeorological explorations and predictions in ungauged locations using large datasets. Water 14 (10), 1657. https://doi.org/10.3390/w14101657.

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

Perez, G., Mantilla, R., Krajewski, W.F., Quintero, F., 2019. Examining observed rainfall, soil moisture, and river network variabilities on peak flow scaling of rainfall-runoff events with implications on regionalization of peak flow quantiles. Water Resour. Res. 55 (12), 10707–10726. https://doi.org/10.1029/2019WR026028.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279 (1), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7.

N.A. Phillips, N.A., 1982. A Very Simple Application of Kalman Filtering to Meteorological Data Assimilation. US Department of Commerce, National Oceanic and Atmospheric Administration.

Razavi, T., Coulibaly, P., 2013. Streamflow prediction in ungauged basins: review of regionalization methods. J. Hydrol. Eng. 18 (8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690.

Razavi, T., Coulibaly, P., 2016. Improving streamflow estimation in ungauged basins using a multi-modelling approach. Hydrol. Sci. J. 61 (15), 2668–2679. https://doi.org/10.5194/adgeo-29-43-2011.

Shu, C., Ouarda, T.B.M.J., 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. Water Resour. Res. 43 (7) https://doi.org/10.1029/2006WR005142.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshmi, V., Liang, X., McDONNELL, J.J., Mendiondo, E.M., O'connell, P.E., Oki, T., Pomeroy, J. W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48 (6), 857–880.

Tabios, G.Q., Salas, J.D., 1985. A comparative analysis of techniques for spatial interpolation of precipitation. JAWRA J. Am. Water Resources Association 21 (3), 365–380. https://doi.org/10.1111/j.1752-1688.1985.tb00147.x.

Tang, G., Clark, M.P., Newman, A.J., Wood, A.W., Papalexiou, S.M., Vionnet, V., Whitfield, P.H., 2020. SCDNA: a serially complete precipitation and temperature dataset for North America from 1979 to 2018. Earth Syst. Sci. Data 12 (4), 2381–2409.

Tarek, M., Brissette, F.P., Arsenault, R., 2020. Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America. Hydrol. Earth Syst. Sci. 24 (5), 2527–2544. https://doi.org/10.5194/hess-24-2527-2020.

Tolson, B.A., Shoemaker, C.A., 2007. Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour. Res. 43 (1) https://doi.org/10.1029/2005WR004723.

Troin, M., Martel, J.-L., Arsenault, R., Brissette, F., 2022. Large-sample study of uncertainty of hydrological model components over North America. J. Hydrol. 609, 127766 https://doi.org/10.1016/j.jhydrol.2022.127766.

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. Hydrol. Res. 38 (3), 211–234. https://doi.org/10.2166/nh.2007.009.

Valéry, A., 2010. Modélisation précipitations–débit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants. Ph.D. thesis, Agro Paris Tech, 417 pp. Available from: <https://webgr.irstea.fr/wp-content/uploads/2012/07/2010-VALERY-THESE.pdf>.

Wackernagel, H., 2003. Multivariate Geostatistics: An Introduction with Applications. Springer Science & Business Media.

Wan, Y., Chen, J., Xu, C.Y., Xie, P., Qi, W., Li, D., Zhang, S., 2021. Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. J. Hydrol. 603, 127065 https://doi.org/10.1016/j.jhydrol.2021.127065.

Wang, W., Zhao, Y., Tu, Y., Dong, R., Ma, Q., Liu, C., 2023. Research on parameter regionalization of distributed hydrological model based on machine learning. Water 15 (3), 518. doi: 10.3390/w15030518.