https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Multimodel implementation of the Raven HBV-EC model for streamflow simulation and uncertainty quantification

Aida Jabbari[1,2], Biljana Music[1,3], David Huard[3], James Craig[4], Richard Turcotte[5], Mohammad Bizhanimanzar[6], Simon Lachance-Cloutier[5], Charles Malenfant[5] and François Anctil[1,2]

[1]Department of Civil and Water Engineering, Université Laval, Québec, G1V 0A6, Canada
[2]CentrEau – Water Research Centre, Université Laval, Québec, G1V 0A6, Canada
[3]Ouranos, 550 Sherbrooke W., West Tower, 19th floor, Montreal, H3A 1B9, Canada
[4]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, N2L 3G1, Canada
[5]Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs (MELCCFP), Direction principale de l'expertise hydrique (DPEH), Québec, G1R 5V7, Canada
[6]Climate risk and resilience consultant, CIMA+, Montréal, Québec, H3C 1J2, Canada

Correspondence to: Aida Jabbari (aida.jabbari.1@ ulaval.ca)

Abstract. Despite advancements in hydrological modelling, quantifying inherent uncertainties in simulation and forecasting remains essential. These uncertainties arise from sources such as initial conditions, input data, parameter estimation, and model structure. While the hydrological community has increasingly focused on uncertainty assessment, most studies concentrate on input data and parameter uncertainty within specific models, leaving model structure uncertainty unexplored. This study introduces an ensemble-based approach to assess hydrological distributed model uncertainty, emphasizing model structure and input data uncertainties concurrently. The study leverages the Raven hydrological modelling framework to create an ensemble of hydrological model structures. This structural ensemble fed noise-perturbed forcing inputs to represent input data uncertainty. The forward greedy method aids in selecting a pool of models from the ensemble, enhancing reliability and reducing the model count. This method is employed to refine the model pool by ensuring that each criterion meets the predefined performance standards. The approach is demonstrated over the southwest portion of the Saint-Laurent watershed in Canada, evaluating model ensembles against observed streamflow. This study advances the understanding of hydrological model uncertainty assessment and emphasizes the significance of a comprehensive, multimodel approach that accounts for structural, input data, and calibration uncertainties for robust streamflow simulations and forecasts. The findings highlight how employing multimodel ensembles is crucial for minimizing different sources of uncertainty, as opposed to relying on a solitary model.

1 Introduction

In recent years, the hydrological modeling community has placed a significant emphasis on addressing uncertainty within hydrological models (Knoben et al., 2020; Moges et al., 2021; Refsgaard et al., 2012; Troin et al., 2022). To tackle these uncertainties, a range of specific analysis methods have been developed. Nonetheless, most of these investigations

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

predominantly concentrate on assessing uncertainty related to input data and model parameterization using a fixed hydrological model, with the model's fundamental structure remaining unchanged. Some studies identified the model structure uncertainty as significant source of uncertainty,(Beven, 2018; Breuer et al., 2009; Clark et al., 2011; Coxon et al., 2019; Di Baldassarre and Montanari, 2009; Seiller et al., 2017; Spieler et al., 2020). Several studies underscore the significance of adopting a multimodel approach to characterize structural uncertainty(Renard et al., 2010; Seiller et al., 2012; Troin et al., 2021; Valdez et al., 2022; Zappa et al., 2011). Moreover, research findings have indicated that, when considering uncertainty partitioning, the influence of model structure uncertainty tends to outweigh model parameter uncertainty(Poulin et al., 2011; Spieler et al., 2020).

One of the challenges posed by most existing hydrological models is their inflexible numerical structure, which hinders the incorporation of modifications. However, this constraint can be overcome through the utilization of flexible hydrological frameworks like Raven developed by(Craig et al., 2020) as a versatile hydrologic modeling framework, offering the ability to construct models using a wide array of pre-existing process algorithms. Raven allows for the comparison of various combinations of model components using identical model forcing and discretization, enabling a comprehensive evaluation of structural uncertainty.

This study introduces an ensemble-based approach for assessing uncertainty in distributed hydrological models. Multiple distributed model ensembles specifically designed for hydrological uncertainty assessment are constructed using Raven. These models are calibrated and verified over the Saint-Laurent Sud-Ouest (SLSO, see Figure 1) domain, which flows into the St. Laurent River, Quebec, Canada. Sub-sets of the many distributed models are then explored to improve the description of the uncertainty.

The study focused on assessing uncertainty arising from model structure in distributed hydrological modeling. Different variants of the Raven model are created to explore structural uncertainty, each with a distinct set of model parameters. However, as the reliability is assessed on the available streamflow observations, the analysis confirms that the (meteorological) input uncertainty needs to be considered at the same time. Through perturbations and comparisons with meteorological reanalysis data that incorporate uncertainties, the research demonstrates a robust method for assessing uncertainty in the context of the study. To the best of our knowledge, there is a scarcity of studies exploring the hydrological uncertainty of distributed models using a multimodel ensemble strategy that enables the simultaneous assessment of uncertainties related to model structure, input data and parameter uncertainty related to the calibration.

This paper is structured as follows. Section 2 introduces the study area and provides information on the meteorological and hydrometric data sets, including details of how the distributed hydrological models were created. Section 3 describe the findings and analyzes how the uncertainty is propagated, while section 4 concludes with final remarks.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

2 Material and methods

In this section the study domain, available climate data, hydrological model and uncertainty assessment analysis method are described.

2.1 Study area

The domain under investigation, known as the SLSO, encompasses a portion of the southern Quebec province in Canada and extends into the state of Vermont (USA), as illustrated in Figure 1.

The land cover within the SLSO is notably diverse, with 31%, classified as deciduous forest, 31% as agriculture, 25% as coniferous forest, 9% as water bodies, 2% as wetland, 1% as impermeable surface, 0.9% as peatland, and 0.1% as bare soil. These land cover data are derived from a new 100-m resolution dataset compiled by the Direction de l'Expertise Biodiversité (DEB) through a collaborative effort that involved multiple Quebec agencies, such the Système d'information écoforestière (SIEF), Ministère du Développement durable, de l'Environnement de la Lutte contre les changements climatiques (MDDELCC), Ministère des Forêts, de la Faune et des Parcs (MFFP), and Ministère de l'Énergie et des Ressources naturelles (MERN). Additionally, they incorporate data from the GlobCover 2009 matrices by the European Space Agency and Circa 2000 matrices by Agriculture and Agri-Food Canada.

The soil texture within the SLSO domain is dominated by loamy sand (55%), sand (30%), loam (10%) and sandy loam (5%). These soil texture data are obtained from two sources. The first source combines data from the Canadian Soil Information Service (CANSIS) and United States Geological Survey (USGS) data, which provide estimates of clay and sand percentages in the top three soil layers. These data cover all North America at a resolution of 1 km. To assign a textural class for the soil column, vertically weighted values of the clay and sand percentages were calculated for each pixel. By employing a pedo-transfer function following the classification proposed by the United States Department of Agriculture (USDA), these values were converted into different textural classes. Nearest neighbor interpolation was applied to infill missing data. The second source corresponds to data produced by (Shangguan et al., 2014).

In the Quebec portion of the SLSO, topographic data at a 10-meter resolution from the Base de Données Topographiques du Québec (BDTQ) were used. For the United States portions, a 30-meter resolution data from the National Elevation Dataset (NED) provided by the U.S. Geological Survey (USGS) were employed. These data sets were aggregated and subsequently converted to a 100-meter resolution. The surface elevation in the SLSO domain ranges from slightly above sea level near the Saint Laurent River to over 900 meters.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

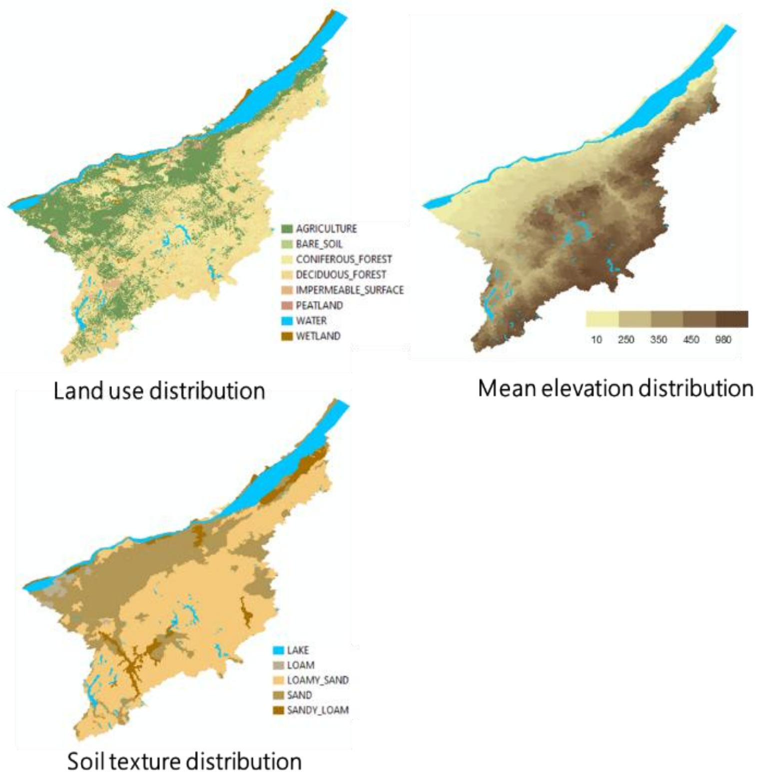


Figure 1: Overview of the SLSO land use, mean elevation (m), and soil texture distribution.

2.2 Meteorological data and observed streamflow

Gridded meteorological data, including daily precipitation, minimum and maximum temperature, covering the period from 1961-2022 and with a spatial resolution of 0.1°, were acquired from the MDDELCC. These data are constructed using ground meteorological stations from the Réseau météorologique coopératif du Québec (RMCQ) and the Direction de la qualité de l'air et du climat (DQAC) using the ordinary Kriging method. According to this dataset, the SLSO domain experiences an average annual precipitation of around 1092 mm, with about 148 days of rainfall per year. Solid precipitation contributes to one-third of the total annual precipitation. Additionally, the mean yearly temperature in the SLSO area is 4.4°C, and there are 235 days each year when the maximum temperature surpasses 0°C.

The streamflow data for various hydrometric stations within the SLSO domain are provided by the MDDELCC for most of the stations from 1961 onwards. As presented in Figure 2, the domain contains 28 hydrometric stations. The characteristics of the stations, including their numbers (provided by MDDELCC), subbasin numbers, areas, and the start and end dates of the available data, are presented in Table 1. In this study, the stations are numbered according to the subbasin numbers associated with each hydrometric station (see Section 2.3.1 and Table 3). Out of the 28 stations, 18 are used for calibration,

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

while all 28 stations are utilized for validation. In this study for calibration, we only used stations that have: a natural flow, which means it does not have large dams that could influence it, a significant number of available data and a reliable rating curve, meaning that we need at least a few events for which we have a gauged flow coupled with a water level.

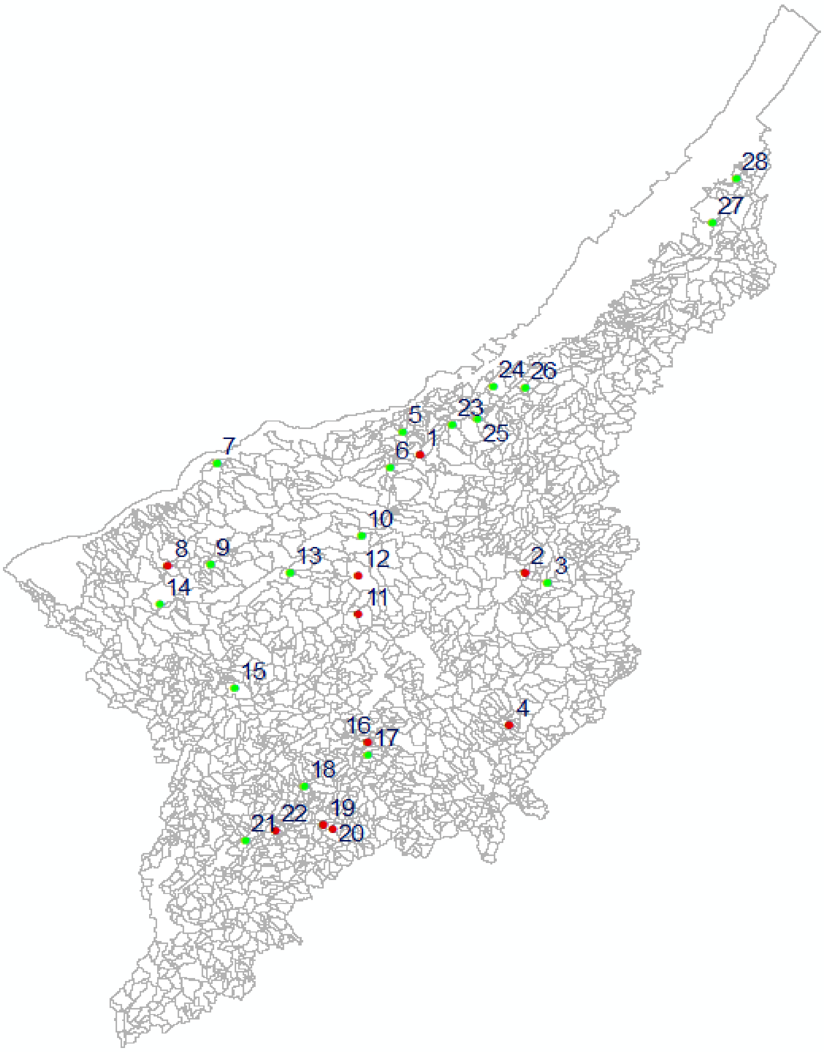


Figure 2: the location of the 28 hydrometric stations in SLSO domain used for the validation (red and green circles) and the stations used for the calibration (green circles)

The streamflow measurements are conducted at 15 minutes intervals. These measurements are then transmitted to an integrated collection system every hour for subsequent validation and processing. It's worth mentioning that the presence of river ice during the winter introduces additional uncertainty to streamflow measurements. Consequently, streamflow data with significant effects of ice are omitted during the calibration.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Table 1: Characteristics of the 28 hydrometric stations in SLSO domain, detailing station ID, number of subbasins, associated drainage area, and the temporal coverage of available data.

Name	Station number	Subbasin	Area (km^2)	Start date	End date
Chaudière	23402	25	5820	1915.02.27	Current time
Chaudière	23426	141	4090	1965.01.20	1981.05.27
Famine	23422	160	696	1964.10.22	Current time
Chaudière	23403	306	1170	1980.10.01	Current time
Beaurivage	23401	585	708	1915.04.10	1983.09.13
Bras dHenri	23432	633	154	1925.09.15	Current time
Petite rivière du Chêne	23702	708	385	1972.06.28	Current time
Bécancour	24007	786	2330	2007.01.11	Current time
Bécancour	24014	796	2163	1970.01.01	2001.10.01
Bécancour	24003	829	914	1999.11.12	Current time
Bécancour	24013	850	227	1966.10.13	Current time
Bullard	24010	880	25.7	1997.05.12	2010.11.19
Bourbon	24015	889	48.1	1977.11.01	1995.10.25
Nicolet	30103	958	1550	2006.07.24	Current time
Nicolet Sud-Ouest	30101	1106	562	1966.10.14	Current time
Au Saumon	30219	1636	839	1929.04.16	Current time
Au Saumon	30282	1659	769	1938.01.01	1977.09.30
Eaton	30234	1945	646	1974.11.27	Current time
Eaton	30242	1972	192.67	1953.04.01	Current time
Eaton	30238	1974	87.2	1966.09.21	1982.09.30
Coaticook	30215	2154	514	1965.08.19	1978.11.15
Aux Saumons	30246	2247	217	1959.07.24	Current time
Etchemin	23303	2388	1152	1968.03.23	1987.08.12
Boyer	23004	2542	191	1996.05.29	Current time
Boyer Sud	23002	2552	61	1993.06.29	Current time
Du Sud	23106	2599	821	1966.10.04	Current time
Ouelle	22704	2792	796	1982.10.01	Current time
Kamouraska	22601	2878	213	2008.09.22	Current time

2.3 Hydrological model

The initial phase of this research involved selecting a suitable distributed hydrological model for Quebec's watersheds, aiming for a complexity level comparable to the HYDROTEL model (Fortin et al., 2001b, a) used in the production of multiple editions of the Hydroclimatic Atlas of Southern Québec (2015, 2023). The chosen base model is the Hydrologiska Byråns Vattenbalansavdelning – Environment Canada (HBV-EC) model (Hamilton et al., 2000), which represents a

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Canadian adaptation of the HBV-96 model originally developed at the Swedish Meteorological and Hydrological Institute (Lindström et al., 1997).

The HBV-EC is a distributed conceptual model that employs daily precipitation, temperature, and long-term monthly potential evaporation as input to simulate streamflow. The decision was made to use a version of the HBV-EC model, implemented within Raven in distributed mode, which operates on a daily time step (Craig et al., 2020). A visual representation of the Raven HBV-EC model is shown in Figure 3. The base HBV-EC is perturbed in structure to generate model ensemble members by modifying the algorithms used to represent individual processes. Note that the potential evapotranspiration in this study was calculated using an empirical formulation (H. Hargreaves and A. Samani, 1985), which provides daily potential evapotranspiration values, replacing the original 'monthly' approach used. Furthermore, the glacier routine was removed from the model since the study area does not include glaciated areas.

2.3.1 Spatial discretization of the SLSO domain

The SLSO river network map is produced utilizing the GIS interface of the HYDROTEL model (Fortin et al., 2001b, a), referred to as PHYSITEL. This process involves utilizing the Digital Elevation Model (DEM) map, along with the altitude raster map and vector map depicting the river network.

In the Raven framework, a subbasin interacts with a particular river reach, and the flow sequencing among these reaches is determined by the downstream ID of the subbasin. The subbasin is subsequently subdivided into Hydrological Response Units (HRUs), each characterized by a unique combination of factors like soil type, land use, vegetation cover, and aquifer type. To assign each HRU to a specific subbasin, the HRU map is intersected with the subbasin map, extracting the relevant subbasin ID. Notably, the process also involves appropriately parameterizing lake HRUs, encompassing aspects like average lake depth, surface area, and total volume. For lake parameterization, we utilized the HYDROLAKES database (Messager et al., 2016). In total, the SLSO domain incorporates 2889 subbasins and 7661 HRUs with 57 lake HRUs. The soil configuration within the model consists of three layers, each associated with specific soil parameters. The spatial distribution of soil varies horizontally across different soil textures, including sand, loam, loamy-sand, and sandy-loam.

For the purpose of setting up, executing, and calibrating the Raven HBV-EC model, RavenPy was applied as a Python wrapper to the Raven hydrological framework (Arsenault et al., 2023). Notably, RavenPy also facilitates the launch of parallel Raven simulations.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

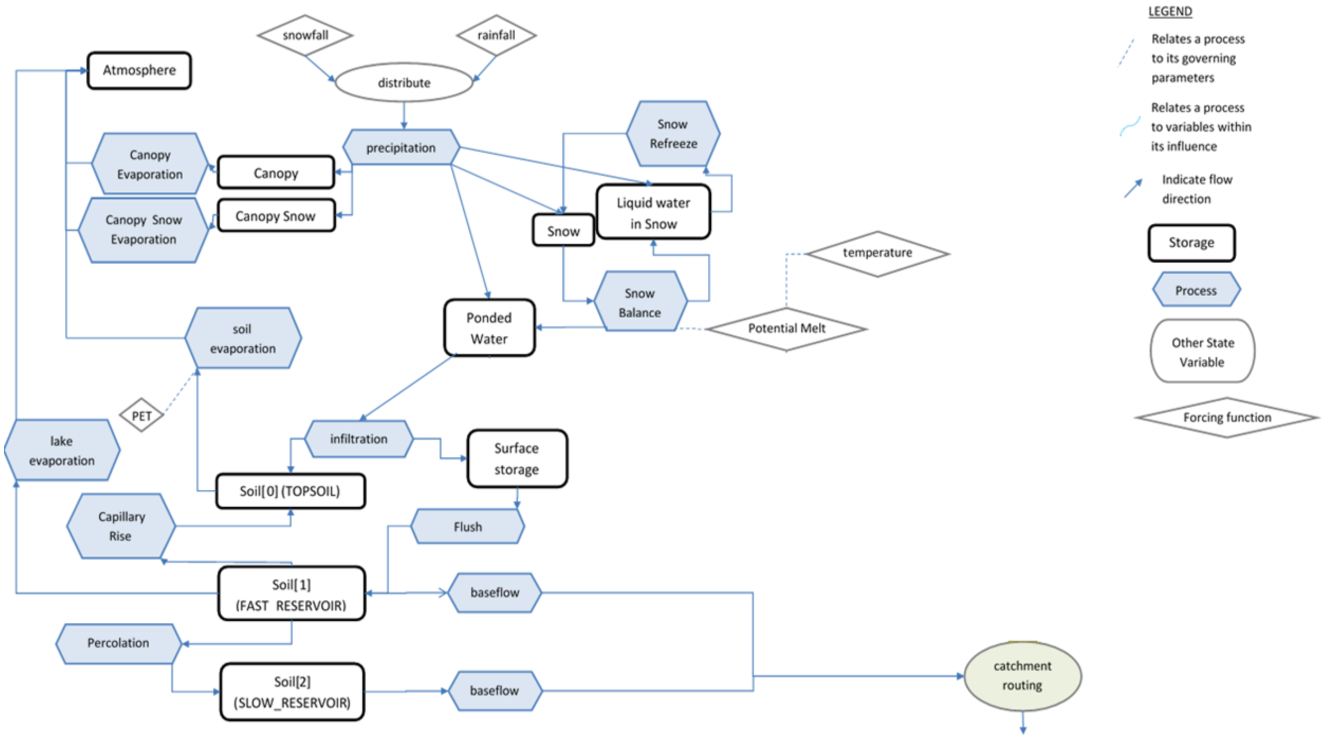


Figure 3: The schematic of Raven HBV-EC (Craig, 2023).

2.4 Sensitivity analysis and regional parameter estimation

Prior to conducting the calibration, a sensitivity analysis was carried out consisting of two steps to find the most sensitive parameters. In the first step, the traditional Elementary Effects (EE) method was employed, utilizing 28 model parameters, 100 trajectories and one iteration (resulting in 2800 model evaluations) to analyze candidate parameters for the base HBV-EC model configuration. The most sensitive parameters were determined by applying an automatically derived threshold within the Efficient Elementary Effects (EEE) method (Cuntz et al., 2015), which led to nine most sensitive parameters (Table 2). As mentioned before, the soil model used in this study comprises three layers and various textures. There are different model parameters for each layer and soil texture, resulting in a total of 31 sensitive model parameters.

Model calibration and validation constitute integral aspects of hydrological modeling. With the initial model parameters in place, Ostrich is employed. Ostrich is an optimization and calibration tool that is independent of the model, employing multiple algorithms. This aids in refining the sensitive parameters to maximize the Kling-Gupta efficiency metric (KGE) (Gupta et al., 2009). The calibration is executed using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007), General-purpose Constrained Optimization Platform (GCOP) is used as the objective function and run

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

for 500 iterations and repeated 5 times. Additionally, calibration iterations of 1000, 2000, and 4000 were performed to compare the KGE variation among different stations.

Table 2: The nine most sensitive HBV-EC parameters as determined through the EEE method.

Parameter	Name	Definition
1	TOPSOIL_THICKNESS	Thickness of 1st soil layer [m]
2	RAINSNOW_TEMP	Rain/snow halfway transition temperature [°C]
3	POROSITY	Effective porosity of the soil
4	HBV_BETA	Soil parameter in HBV infiltration algorithm
5	FIELD_CAPACITY	Field capacity saturation of the soil
6	MELT_FACTOR	Maximum snow melt factor [mm/day/°C]
7	MIN_MELT_FACTOR	Minimum snow melt factor [mm/day/°C]
8	HBV_MELT_FOR_CORR	HBV snowmelt forest correction
9	HBV_MELT_ASP_CORR	HBV snowmelt aspect correction

We employed 28 hydrometric stations for validation and calibrated the model using 18 hydrometric stations. The locations of the hydrometric stations are illustrated in Figure 2, while Table 3 provides the station numbers and the KGE values for both validation and calibration. For validation, the stations are numbered from 1 to 28 according to the subbasin associated with each station, ordered from smallest to largest. For calibration, the stations are numbered from 1 to 18 based on the associated subbasin, listed in descending order.

It's crucial to emphasize that our primary objective is to establish a regional parameter set applicable to the entire SLSO domain, including many ungauged reaches. This strategy aligns with the methodology used in the Hydroclimatic Atlas of Southern Québec, where parameters calibrated in one area are subsequently applied to locations without gauge data.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Table 3: Comparison of KGE between calibration and validation time series across multiple hydrometric stations in the SLSO domain.

Name	Station number	Numbering for validation	Validation KGE	Numbering for calibration	Calibration KGE
Chaudière	23402	1	0.78	-	-
Chaudière	23426	2	0.71	-	-
Famine	23422	3	0.79	1	0.78
Chaudière	23403	4	0.69	-	-
Beaurivage	23401	5	0.74	2	0.71
Bras dHenri	23432	6	0.71	3	0.75
Petite rivière du Chêne	23702	7	0.75	4	0.72
Bécancour	24007	8	0.79	-	-
Bécancour	24014	9	0.73	5	0.78
Bécancour	24003	10	0.66	6	0.70
Bécancour	24013	11	0.75	-	-
Bullard	24010	12	0.65	-	-
Bourbon	24015	13		7	0.74
Nicolet	30103	14	0.70	8	0.73
Nicolet Sud-Ouest	30101	15	0.70	9	0.63
Au Saumon	30219	16	0.77	-	-
Au Saumon	30282	17	0.78	10	0.74
Eaton	30234	18	0.66	11	0.71
Eaton	30242	19	0.65	-	-
Eaton	30238	20	0.70	-	-
Coaticook	30215	21	0.63	12	0.77
Aux Saumons	30246	22	0.68	-	-
Etchemin	23303	23	0.64	13	0.78
Boyer	23004	24	0.77	14	0.71
Boyer Sud	23002	25	0.67	15	0.74
Du Sud	23106	26	0.78	16	0.85
Ouelle	22704	27	0.70	17	0.79
Kamouraska	22601	28	0.78	18	0.81

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

2.5 Uncertainty assessment

2.5.1 Generating Raven HBV-EC model variants

To assess model structural uncertainty, a range of models are generated by employing various combinations of snow melt, snow balance, and potential evapotranspiration algorithms within the Raven model framework. In the initial model configuration (Figure 3), the HBV snow melt, the simple balance, and the Hargreaves potential evapotranspiration methods were used, respectively.

Relying on potential mixture of snow melt, snow balance, and potential evapotranspiration algorithms, a total of 48 models, including the original HBV-EC model, are created from a combination of 4 snow melt methods, 3 snow balance approaches and 4 potential evapotranspiration formulations (Figure 4). The evaluation of all possible model combinations is beyond the scope of this study. More information regarding the above-mentioned methods can be found in Raven manual. In the next step, all variants are calibrated using the approach, resulting in various model parameter sets for each model. The simulations from these calibrated models are then used to explore the uncertainty.

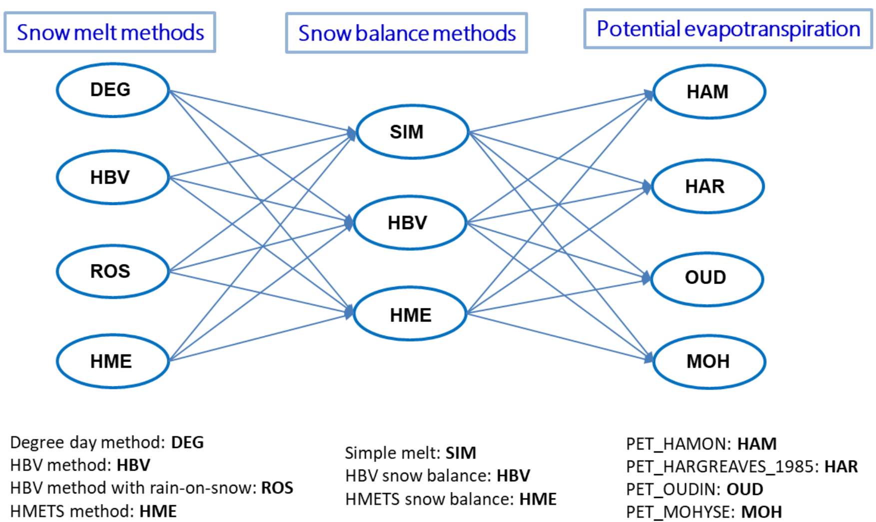


Figure 4: Configurations of snow melt, snow balance, and potential evapotranspiration components leading to 48 unique models.

2.5.2 Sub-model selection

Initially, an assessment of uncertainty is conducted based on the ensemble of 48 models. To streamline the subsequent evaluation process, and to remove ensemble members which were overly similar in character, a sub-selection of models is devised. The sub-model selection is intended to yield a smaller number of the initially created models to reduce the time of calculations and simultaneously receive the benefit of multi model ensembles to explore the uncertainty. There are a great number of possible model combinations. For the purposes of this study, we are looking for a smaller set of models that will still yield comparable or better results than using the entire group of models.

Various techniques can be employed for sub-selection within a dataset, including greedy selection methods (Brochero et al., 2011; Seiller et al., 2017). The forward greedy strategy is adopted in this study to refine the model pool by ensuring that each criterion meets the predefined performance standards.

To enable a comparative analysis of model selections, four different sets are established: set-0 through set-3. Set-0 comprises all 48 models. Set-1 encompasses models chosen through the forward greedy approach, prioritizing performance enhancement by reducing the root mean square error (RMSE) between simulated and observed streamflow. Set-2 consists of models selected through forward greedy optimization, aimed at minimizing the discrepancy between the RMSE and the square root of the average ensemble variance in streamflow, referred to here as spread. Note that this discrepancy is taken here as an estimate of reliability (Fortin et al., 2014). Set-3 incorporates models chosen based strictly on their performance: the models are ranked according to the average KGE across the hydrometric stations, with the top 10 models being selected.

The selection process for models in the forward greedy optimization for set-1 and -2 proceeds as follows:

1. The selection of the top 5 models is determined by identifying those with the lowest RMSE.
2. Subsequently, the remaining models are progressively included into the group, with each addition subjected to an assessment of potential performance enhancement.
3. In cases where the newly added model's contribution to performance, indicated by RMSE in set-1 and the difference between RMSE and spread in set-2, is not substantial, the model is removed from consideration.
4. This iterative process continues until all models have been evaluated, resulting in the determination of the final count of acceptable models.

2.5.3 Accounting for uncertain input data

In this study, to address input data sources of uncertainty, random white noise perturbation is added to the precipitation and temperature data. We followed the approach (Thiboult and Anctil, 2015), which involved creating three precipitation noise levels (gamma distributions with a standard deviation equivalent to 25%, 50%, and 75% of the mean value) and three temperature noise levels (normal distributions with a standard deviation of 1, 2, and 3°C). For each magnitude of the noise added to the meteorological data, an ensemble of 48 members is created. The random noise data are added to each grid inside the SLSO domain independently. We did not account for the spatial covariance since the study area is small. Applying the same perturbation to each grid point generally preserves the existing spatial covariance between the grid points but can obscure fine-scale variability. This impact on spatial patterns becomes more significant in larger study areas.

To ascertain the suitable percentage of noise to be incorporated into precipitation and temperature data, an independent dataset is employed to gauge the dispersion of precipitation and temperature within our noisy data. This independent dataset is derived from the ensemble meteorological dataset for North America (EMDNA), a compilation of 100 members (Tang et al., 2020). This comprehensive dataset includes daily temperature range as well as daily precipitation and temperature data,

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

spanning from 1979 to 2018, at a spatial resolution of 0.1°. To draw a comparison between the precipitation spread within EMDNA and the perturbed precipitation (derived from 48 ensemble members featuring varied percentages of noise), 10 distinct grid points within the SLSO domain were selected. Extraction of data was carried out using the nearest neighbor method, as the grid locations in EMDNA differ from those within the SLSO domain.

2.5.4 Model calibration uncertainty

Hydrological model uncertainties can stem from a variety of sources, including the calibration process, where the DDS method does not always guarantee convergence to the global optimum. These uncertainties manifest through dissimilar parameterizations that involve key parameters which significantly influence the behavior and performance of the model. Given this, it is essential to explore and understand different types of uncertainties that can affect hydrological modeling.
To facilitate a comprehensive comparison of these diverse uncertainty types, this study specifically investigates the uncertainties associated with the model calibration process. This investigation is carried out by selecting two distinct models from a pool of available options and subjecting them to a rigorous calibration procedure. This calibration process is repeated 48 times for each model, with each repetition consisting of 500 iterations using the Ostrich optimization tool. By performing multiple calibration runs, the study aims to explore the variability in model parameters and assess how this variability affects the model's predictions and overall performance.

2.5.5 Evaluation metrics

This study involves a comparison between the RMSE and the spread of ensembles, which provides insight into potential under-dispersion or over-dispersion in the ensembles. The purpose of this comparison is to gauge the dispersion of ensembles relative to their simulation capabilities. This assessment is carried out through the normalized root mean square error ratio (NRR) as in Eq.(1).

$$NRR = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\left[\frac{1}{N}\sum_{n=1}^{N}\hat{y}_t^n\right]-y_t^n\right)^2}}{\frac{1}{N}\left\{\sum_{n=1}^{N}\sqrt{\frac{1}{T}\left[\sum_{t=1}^{T}(\hat{y}_t^n-y_t^n)^2\right]}\right\}\sqrt{\frac{N+1}{2N}}} \tag{1}$$

where y_t is observation, $\hat{y}_t$ is ensemble simulation average, N is the number of ensemble members and t is the time. An NRR value of 1 signifies that the spread is reliable. NRR>1 implies that the ensemble is too narrow, and NRR<1 implies that it is too wide.
Furthermore, the spread skill diagram is constructed, which provides an additional form of verification. This diagram serves as a visual tool to comprehensively assess the ensemble's bias, spread, and overall reliability (Thiboult and Anctil, 2015) and operates under the fundamental principle that achieving reliability requires the RMSE to be in line with the spread (Fortin et

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

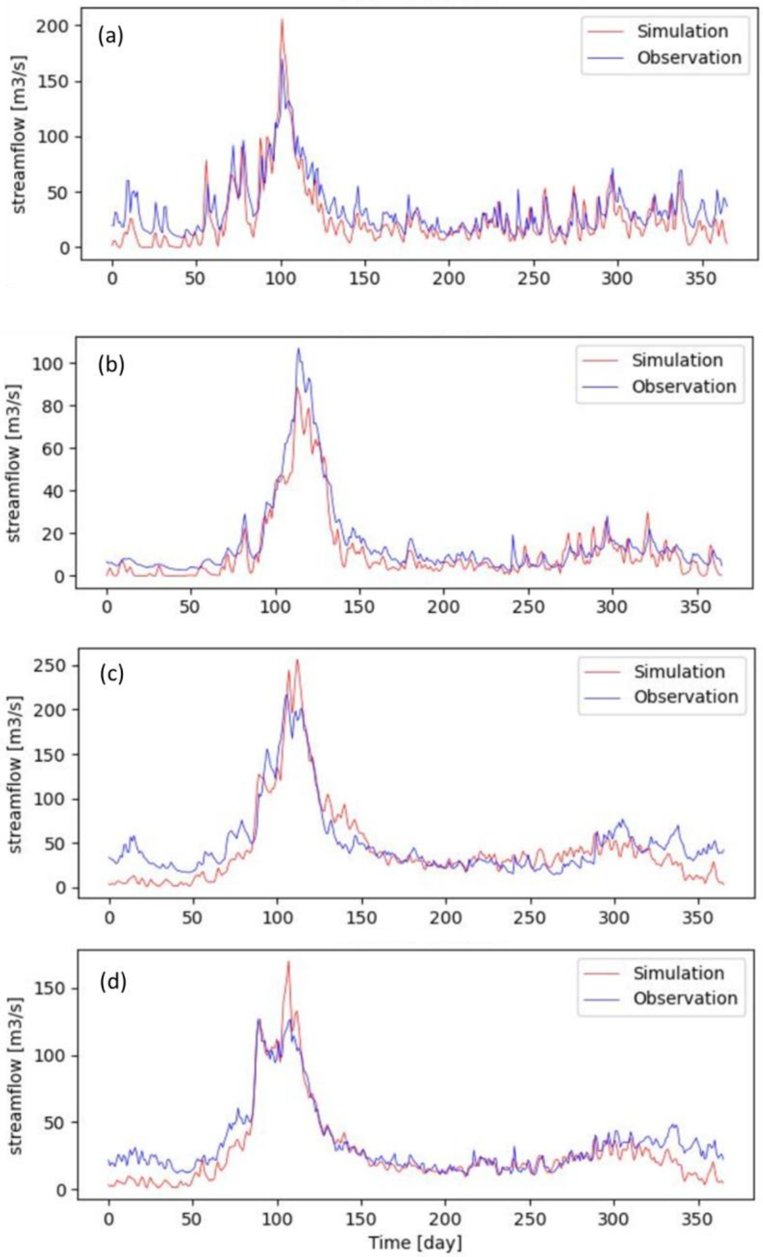
al., 2014). A graphical representation of the ensemble reliability is also provided by the reliability diagram (Thiboult and Anctil, 2015), which plots the simulation probabilities against observed event frequencies. It can provide a diagnostic concerning the bias and the dispersion (Anctil and Ramos, 2019). A perfectly reliable system lies on a diagonal line, which means that the probability of the simulation is equal to the frequency of the event (Valdez et al., 2022).

3 Results

3.1 Implementation of Raven HBV-EC on the SLSO domain

The necessary input data are produced using RavenPy and the initial model parameters are determined to enable the implementation of Raven HBV-EC across the study area. The regional calibration is applied to the period from 2007 – 2017, with one year spin up time (not evaluated within the objective function), to initialize model states considering the pool of hydrometric time series across the SLSO domain. The validation judges the calibrated model over 1961 – 2020, with two years spin up time. During the calibration procedure (2007 – 2017), the KGE is computed for each hydrometric station. The model performance is shown in Figure 5, depicting its ability to replicate streamflow patterns at the hydrometric stations from 1961 to 2020. For example, in the calibration period the model-simulated streamflow closely aligns with the observed flow at the Bécancour 24003 (KGE = 0.70) and the Nicolet (KGE = 0.73) stations (Table 3). Notably, both the Bécancour 24003 station (KGE = 0.66) and the Nicolet station (KGE = 0.70) display favorable levels of agreement for the validation period (Table 3). We consider KGE > 0.6 as acceptable values for our study, however various thresholds for the KGEs can be considered for various studies based on their own objectives. The main modeling issue lies with the winter low flows that are underestimated by the model (Figure 5).

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026



295 **Figure 5: Mean daily streamflows for all years of a) and b) calibration (2007– 2017) in Bécancour and Nicolet stations respectively and c) and d) validation (1961 – 2020) of Raven HBV-EC in Bécancour and Nicolet stations for simulated and observed flows.**

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

3.2 Evaluation of the multimodel ensemble

Following the evaluation of Raven HBV-EC calibration performance with the KGE metric, we created an ensemble of models to assess hydrological uncertainty across the 18 hydrometric stations used for calibration in the SLSO domain. By utilizing a blend of four algorithms for snow melt, three for snow balance calculation, and four formulations for potential evapotranspiration, a total of 48 unique models (including the original HBV-EC) were developed to represent the influence of model structural uncertainty. The average weekly streamflow simulations over the period 2000–2020 for the Au Saumon station are shown in Figure 6, where the pale and dark blue coverages illustrate the distribution of the streamflow ensemble for 48 the models (5 to 95% and 25 to 75%, respectively), while the red line illustrates the median flow and the black line, the observed flow. The observed streamflow falls within the 5 to 95% coverage except from April to May and November to the end of December in which the streamflows are underestimated, as already discussed.

Observations indicate that the model's performance varies across different years and stations, with some years showing overestimation and others showing underestimation. However, on average, the model tends to underestimate peak flows. The largest spread occurs during the mid-April and late May while the smallest spread occurs in November and December. These findings highlight that capturing high flows is more challenging compared to low flows due to their great variability (Seiller and Anctil, 2014). During the transition from winter to spring, there are significant challenges in representing the snowmelt process, and the division of total precipitation between rain and snowfall is highly uncertain. There is a need for accounting for more sources of uncertainty (to increase spread), showed that the hydrological processes are not the only source of uncertainty and climate uncertainty could be an issue that is dealt in a subsequent step below.

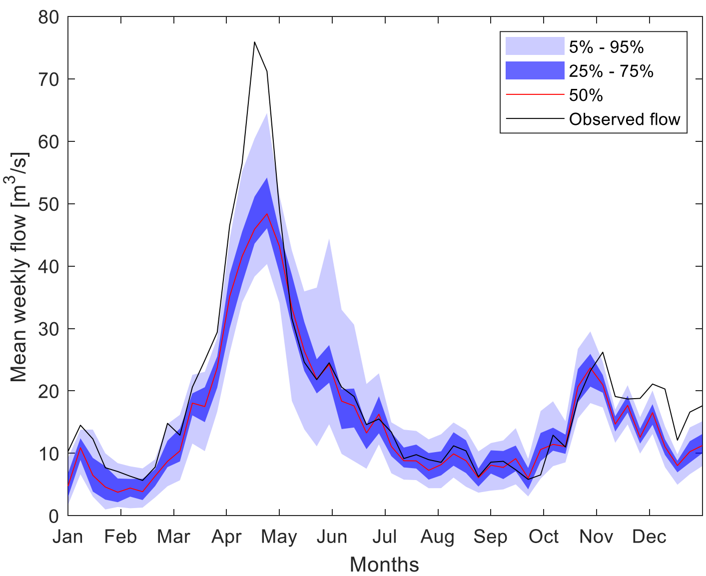


Figure 6: Comparison of the average annual observed streamflow and streamflow ensembles from 48 models during the simulation period (2000 – 2020) depicting median flow simulation, 5th to 95th percentile range, and 25th to 75th percentile range of the streamflow ensemble in Au Saumon station.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

3.3 Uncertainty assessment

3.3.1 Model structure uncertainty

Multimodel ensembles were constructed in the form of the 48 different models (Table 4). First, a simple pooling of all 48 model outputs was considered. The goal of multimodel combination is to extract the maximum amount of information available from a collection of existing models. In the Table 4, the model number and model abbreviations are indicated. The first three characters denote the snow melt method, the next three characters indicate the snow balance method, and the remaining characters indicate the potentiation evapotranspiration method.

Table 4: The 48 models which are used in this study varying snow melt, snow balance and potential evapotranspiration methods from the original HBVEC model

No.	Model	No.	Model	No.	Model
M1	DEG-SIM-HAM	**M17**	HBV-HBV-HAM	**M33**	ROS-HME-HAM
M2	DEG-SIM-HAR	**M18**	HBV-HBV-HAR	**M34**	ROS-HME-HAR
M3	DEG-SIM-OUD	**M19**	HBV-HBV-OUD	**M35**	ROS-HME-OUD
M4	DEG-SIM-MOH	**M20**	HBV-HBV-MOH	**M36**	ROS-HME-MOH
M5	DEG-HBV-HAM	**M21**	HBV-HME-HAM	**M37**	HME-SIM-HAM
M6	DEG-HBV-HAR	**M22**	HBV-HME-HAR	**M38**	HME-SIM-HAR
M7	DEG-HBV-OUD	**M23**	HBV-HME-OUD	**M39**	HME-SIM-OUD
M8	DEG-HBV-MOH	**M24**	HBV-HME-MOH	**M40**	HME-SIM-MOH
M9	DEG-HME-HAM	**M25**	ROS-SIM-HAM	**M41**	HME-HBV-HAM
M10	DEG-HME-HAR	**M26**	ROS-SIM-HAR	**M42**	HME-HBV-HAR
M11	DEG-HME-OUD	**M27**	ROS-SIM-OUD	**M43**	HME-HBV-OUD
M12	DEG-HME-MOH	**M28**	ROS-SIM-MOH	**M44**	HME-HBV-MOH
M13	HBV-SIM-HAM	**M29**	ROS-HBV-HAM	**M45**	HME-HME-HAM
M14	HBV-SIM-HAR	**M30**	ROS-HBV-HAR	**M46**	HME-HME-HAR
M15	HBV-SIM-OUD	**M31**	ROS-HBV-OUD	**M47**	HME-HME-OUD
M16	HBV-SIM-MOH	**M32**	ROS-HBV-MOH	**M48**	HME-HME-MOH

Figure 7 illustrates the assessment of multimodel ensembles using a spread skill diagram. The hydrometric stations which are used in calibration process (the 18 stations, Table 3) are selected to quantify the model structure uncertainty. The plotted results of the spread and the RMSE indicate that the multimodel ensemble simulations are under-dispersed. When the RMSE exceeds the spread, the ensemble demonstrates overconfidence in its simulation abilities, and conversely, when the RMSE is

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

smaller than the spread, it indicates under confidence. As stated before, under-dispersion is expected at this stage since all uncertainty sources are not yet accounted for, the (meteorological) input uncertainty is notably missing.

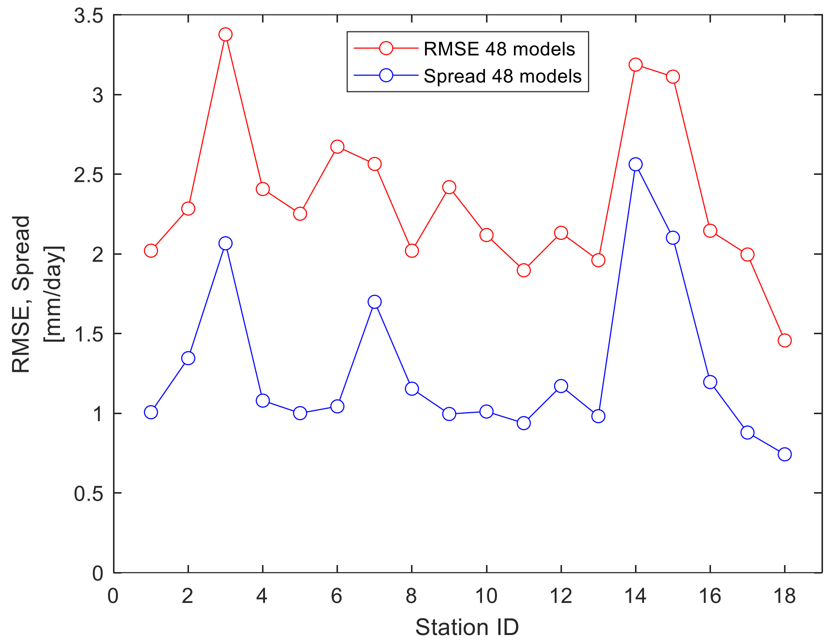


Figure 7: Verification of the 48-member multimodel: RMSE and spread values for each of the 18 hydrometric stations with the ensemble including only the influence of structural uncertainty.

The subsequent phase of uncertainty assessment involves evaluation of various model subsets to seek performance enhancement through model selection and the reduction of the number of models within a multimodel ensemble group. In this study, subsets of the models were identified objectively using the forward greedy method described in Sect. 2.5.2. the forward greedy approach is chosen to select various sub-models by reducing uncertainties through developing a criterion for excluding poor models. Table 5 indicates the characteristics of the different sets. The created sets are then compared with the results obtained from the 48 model ensembles, referred to as set-0. Note that the set-0 uses all the models listed in Table 4. shown in Figure 8 (a).

Table 5: Model names and characteristics in sets 1 to 3 and their method references.

No.	Model	No.	Model	No.	Model
			Set-1: includes 12 model		
M1	DEG-SIM-HAM	M7	DEG-HBV-OUD	M21	HBV-HME-HAM
M2	DEG-SIM-HAR	M10	DEG-HME-HAR	M30	ROS-HBV-HAR

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

M5	DEG-HBV-HAM	M14	HBV-SIM-HAR	M31	ROS-HBV-OUD
M6	DEG-HBV-HAR	M19	HBV-HBV-OUD	M34	ROS-HME-HAR
Set-2: includes 30 model					
M1	DEG-SIM-HAM	M14	HBV-SIM-HAR	M27	ROS-SIM-OUD
M2	DEG-SIM-HAR	M15	HBV-SIM-OUD	M29	ROS-HBV-HAM
M3	DEG-SIM-OUD	M17	HBV-HBV-HAM	M30	ROS-HBV-HAR
M5	DEG-HBV-HAM	M18	HBV-HBV-HAR	M31	ROS-HBV-OUD
M6	DEG-HBV-HAR	M19	HBV-HBV-OUD	M33	ROS-HME-HAM
M7	DEG-HBV-OUD	M21	HBV-HME-HAM	M34	ROS-HME-HAR
M9	DEG-HME-HAM	M22	HBV-HME-HAR	M35	ROS-HME-OUD
M10	DEG-HME-HAR	M23	HBV-HME-OUD	M39	HME-SIM-OUD
M11	DEG-HME-OUD	M25	ROS-SIM-HAM	M45	HME-HME-HAM
M13	HBV-SIM-HAM	M26	ROS-SIM-HAR	M46	HME-HME-HAR
Set-3: includes 10 model					
M1	DEG-SIM-HAM	M7	DEG-HBV-OUD	M11	DEG-HME-OUD
M2	DEG-SIM-HAR	M9	DEG-HME-HAM	M26	ROS-SIM-HAR
M3	DEG-SIM-OUD	M10	DEG-HME-HAR	M27	ROS-SIM-OUD
M5	DEG-HBV-HAM				

The comparison between set-1 and set-0 revealed that the 12 selected models in set-1 exhibited a decrease in RMSE and an improvement in spread relative to set 0 (Figure 8a). In the case of set-2, the difference between RMSE and spread decreased for this group of 30 models, as compared to the initial set of 48 models. All the ensembles exhibit under-dispersion, since the spread is smaller than the RMSE. This evaluation is confirmed by the corresponding reliability diagrams, comparing results of the different sets for Famine watershed (Fig. 8 (b)). Overall, these results confirm that it is possible to work with a smaller number of models, which is operationally and computationally beneficial, without loss in performance or reliability. In fact, small gains are achieved.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

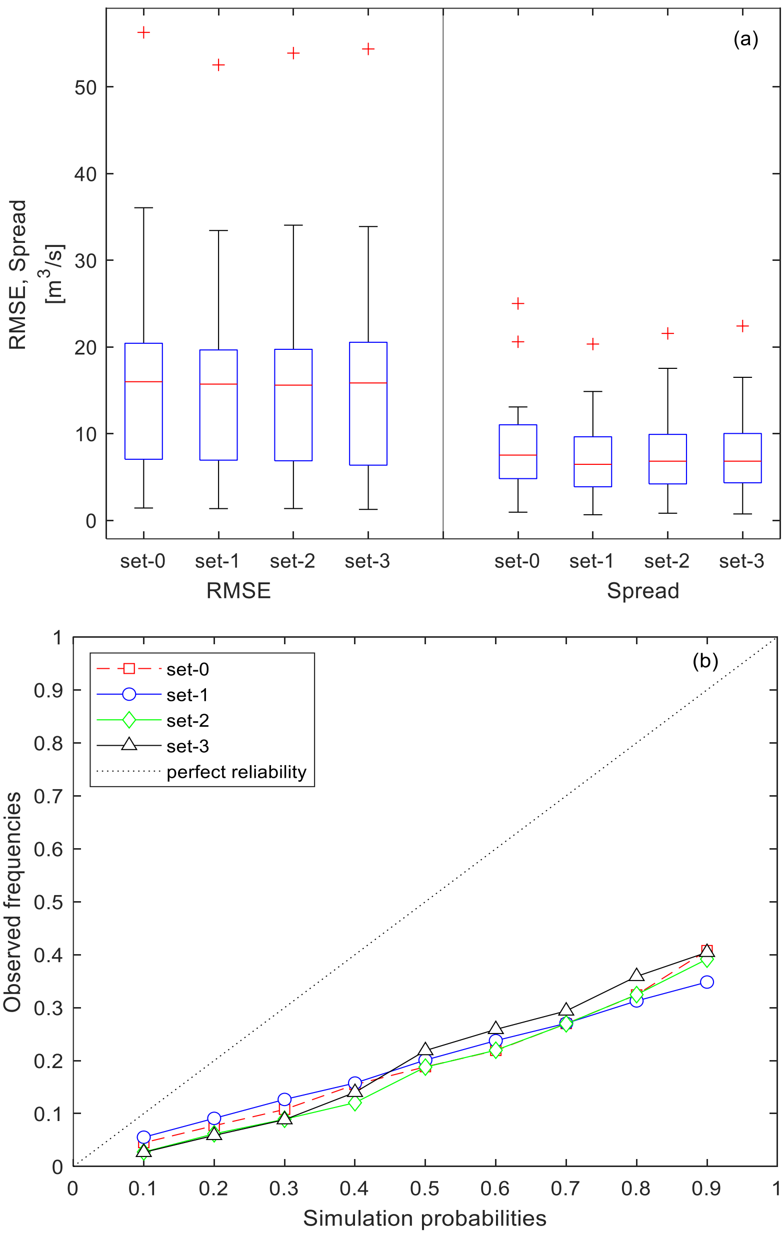


Figure 8: Exploring the uncertainty for different sets. a) Comparison of the RMSE and spread for set-0, set-1, set-2 and set-3. The boxplot captures the distribution formed by the 18 hydrometric stations available within the SLSO domain. b) Reliability diagrams for set-0, set-1, set-2 and set-3, for the Famine River hydrometric station (station ID:1).

One issue that always comes to mind when dealing with multimodel ensembles is: would it not be better (and simpler) to systematically work with the single best model for each watershed, even if that would prevent having any clue about the related uncertainty. This issue is evaluated here, notably to further guide in the selection of the best set out of the four

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

considered here. This evaluation involved comparing the RMSE of the best model for each of the different hydrometric stations (this model thus differs from on site to the other, out of the pool of 48) to the RMSE values of the four multimodel sets.

Figure 9 displays these results as relative RMSE variations. More specifically, the relative RMSE is a statistical measure employed to gauge the precision of a model's simulations through a comparison of the root mean square error with the range of observed values. It serves as a standardized assessment of the root mean square error, computed as the difference between the RMSE of the best model and the RMSE of the different sets, divided by the RMSE of the best model. Relative RMSE is frequently presented as a percentage, with higher positive values indicating reduced RMSE within the sets which is calculated for each hydrometric station.

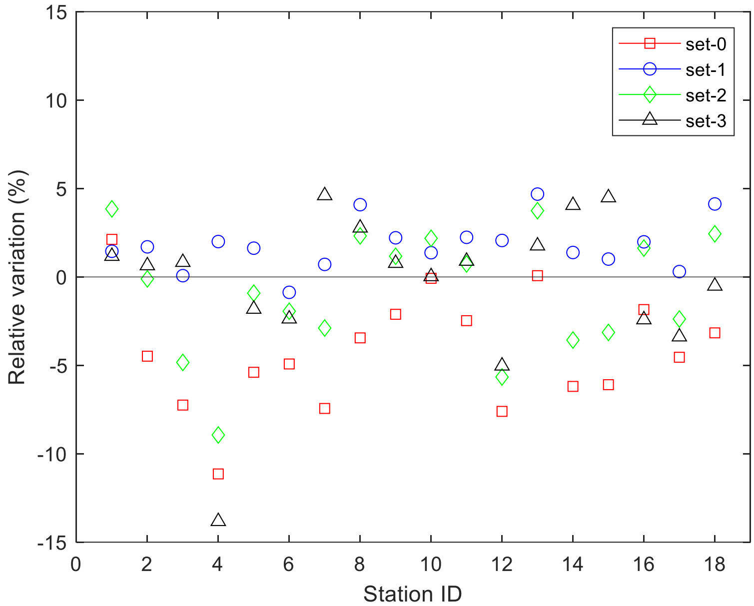


Figure 9: Relative RMSE variation for different sets comparing with the best individual model in each hydrometric station.

By comparing all performance values in Figure 9, it was found that the ensemble from set-1 exhibited superior results in most cases compared to the best model among the 48 available ones. It is noteworthy that the pool of all 48 model (set-0) is systematically the worst option and that the pool of the 10 best individual model (set-3) is generally surpassed by set-1 that originates from a forward greedy optimization of the RMSE that ended up identifying a pool of 12 models (Table 5).

3.3.2 Calibration uncertainty

Following the selection of a subset of models for uncertainty exploration, the subsequent phase of the study involved exploring the uncertainty associated with the identification of the effective parameters of the models. Uncertainty in these

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

parameters can arise due to various factors. Firstly, the inability to estimate or measure these effective parameters, which integrate and conceptualize processes, contributes to their uncertainty. Secondly, parameter uncertainty can be a result of natural process variability and observation errors. Errors in the calibration process can lead to parameter uncertainty, even if the model perfectly represents the hydrologic system. Thus, the combination of inaccurate estimation of effective parameters, difficulties in measuring natural variability, and observation errors collectively contribute to parameter uncertainty. As a consequence, model calibration often lacks a single optimal set of parameters.

The assessment of calibration uncertainty involves the selection of two models, each calibrated 48 times (using the Ostrich - DDS algorithm with 500 iterations). This process generates 48 distinct parameterizations for each chosen model. Out of the 48 available models, we specifically selected Model 1 (M1) and Model 14 (M14) to investigate calibration uncertainty. M1 utilizes the degree day method for snowmelt, the simple snow balance approach for snow balance computations, and the Hamon method for potential evapotranspiration, denoted as DEG-SIM-HAM. furthermore, M14 represents the original Raven HBV-EC model, abbreviated as HBV-SIM-HAR, which employs the HBV method for snowmelt, a simple snow balance method for snow balance calculations, and the Hargreaves 1985 method for potential evapotranspiration estimation.

The outcomes shown in Figure 10 depict the impact of evaluating model parameter uncertainty related to the calibration uncertainty for M1 and the 12 selected models (set-1 listed in Table 4) in the spread-skill (a) and reliability diagram (b). These results reveal a consistent trend: as calibration uncertainty is assessed, there is a reduction in spread and a concurrent increase in RMSE. Comparing the performance of these individual models (M1 and M14) with that of the multimodel ensembles, a decrease in performance becomes evident. Specifically, when evaluating only calbration uncertainty, Models 1 and 14 exhibit a noticeably smaller spread and larger RMSE than that of the structural uncertainty ensembles.

Figure 10(b) depicts a reliability plot that compares calibration uncertainty to a multimodel ensemble composed of 12 selected models. The outcomes reveal that optimizing subsets leads to improved performance, which, in turn, significantly bolsters ensemble reliability. Notably, reliability concerning calibration uncertainty substantially decreased in M1. These findings underscore the effectiveness of the multimodel ensemble approach using set-1 in both exploring uncertainty and enhancing reliability when compared to calibration uncertainty alone. This evaluation underscores the superior reliability and skill of the multimodel ensemble in contrast to the more traditional single-model ensemble approach in all of the hydrometric stations.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

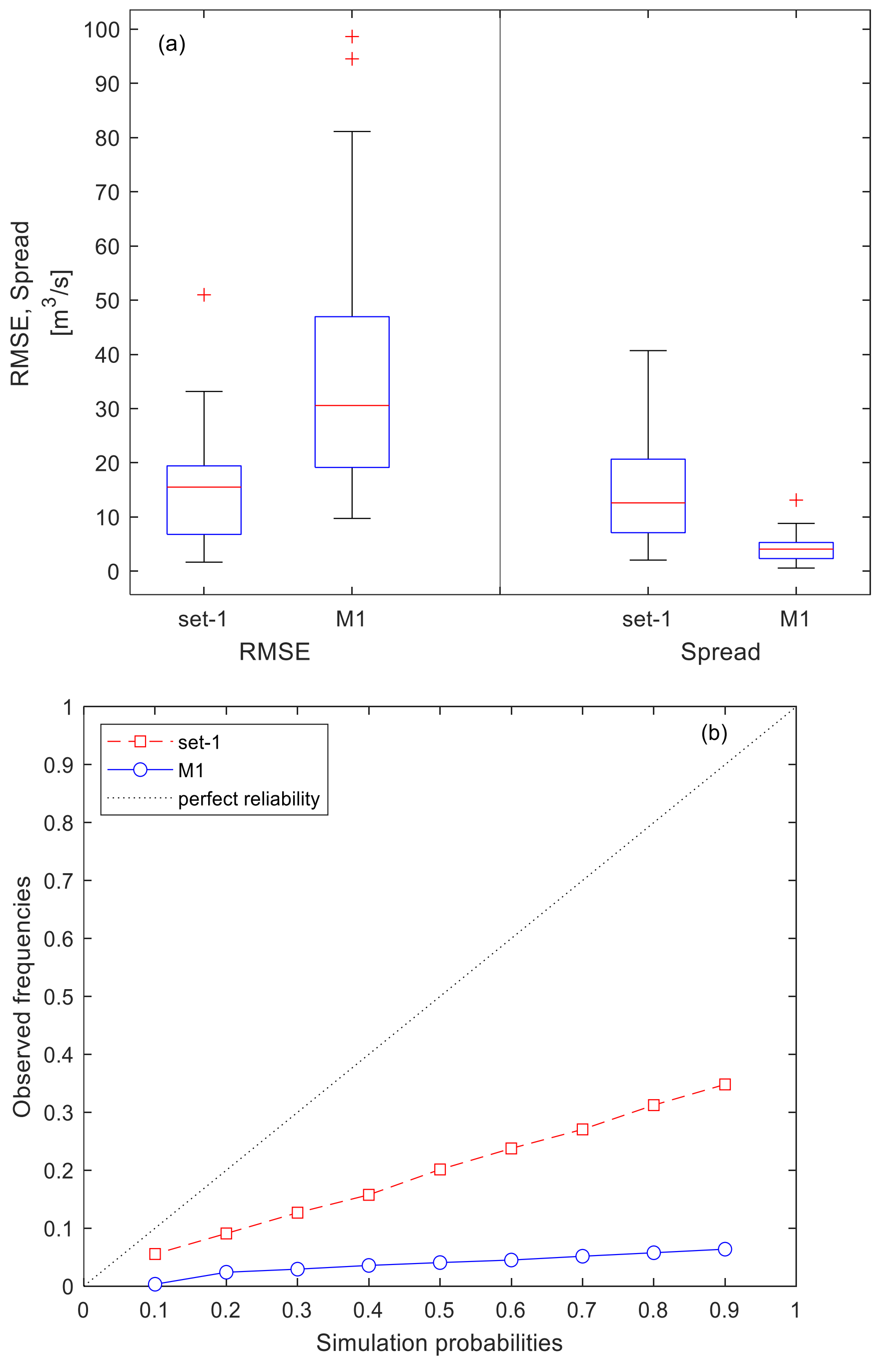


Figure 10: Evaluation of the calibration uncertainty. a) Spread-skill diagram for set-1 models and different model parameters for model M1 (DEG-SIM-HAM) calculated over the evaluation period. The boxplot captures the distribution formed by the 18 hydrometric stations available within the SLSO domain. b) Reliability diagram investigating the calibration uncertainty in Famine River hydrometric station (station ID:1).

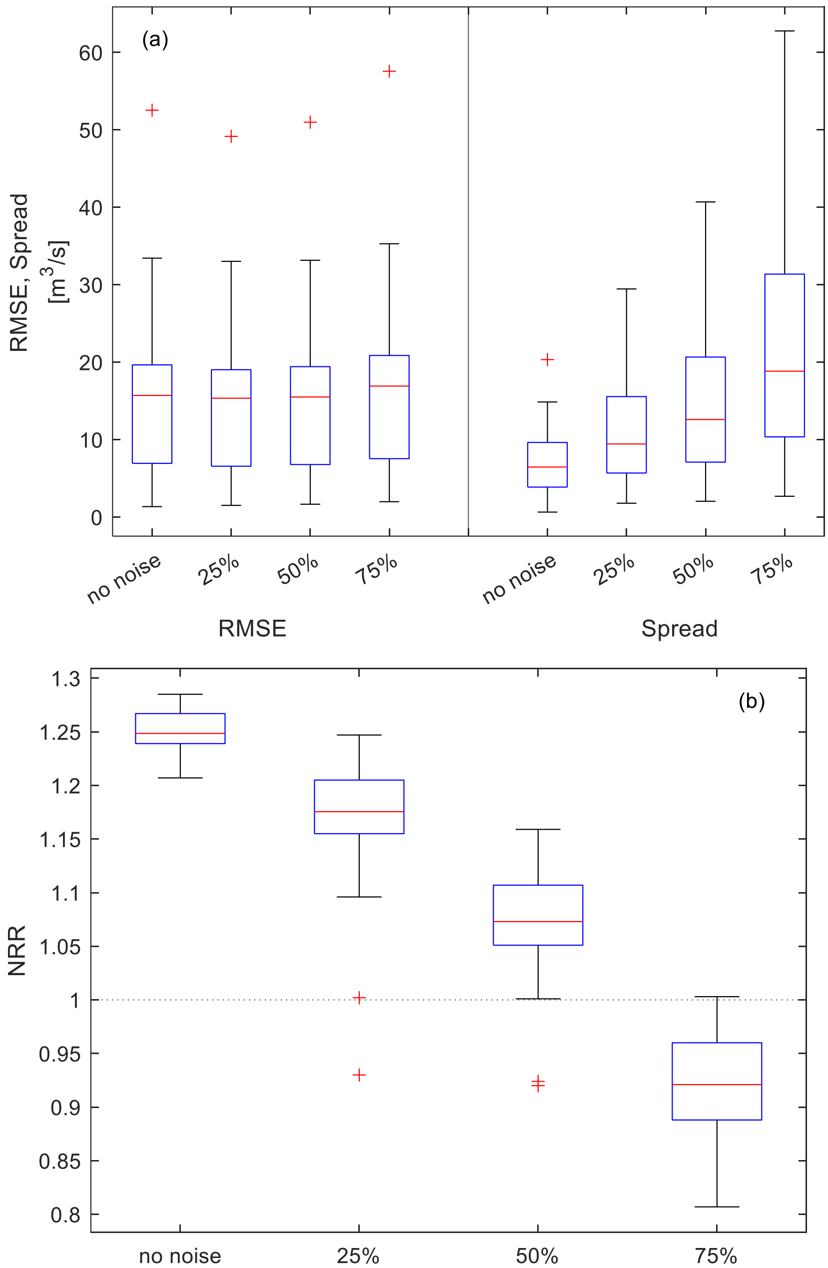
Due to space constraints, we have opted to include only the figure related to the model M1 (DEG-SIM-HAM) in this paper. Furthermore, the outcomes of the NRR evaluation reveal that incorporating noise into the input data of the multimodel ensemble enhances the NRR values, effectively bringing them closer to 1. This improvement is evident when comparing the results with those obtained from introducing no noise to the forcing data. Moreover, when investigating calibration uncertainty for Models 1 and 14, it becomes evident that the NRR values increase, particularly towards the optimal value of one. Notably, better NRR values are achieved with higher levels of noise. These findings underscore that introducing noise to the forcing data within a multimodel ensemble framework offers a more effective means of exploring uncertainty, in contrast to the uncertainty stemming from model parameter variation within individual hydrological models.

3.3.3 Input data uncertainty

The final phase of the study involved introducing noise to the input data to assess the influence of input uncertainty within the multimodel ensembles. Based on the outcomes of the preceding stage, set-1 was identified for further analysis, comprising 12 models. To investigate uncertainty, the precipitation data were perturbed at noise levels of 25%, 50%, and 75%. This perturbation was applied to 48 climate data series to thoroughly explore the effects of noise at each percentage.

The comparison of the multimodel ensemble results with respect to RMSE and spread revealed that as the noise level increased, the spread also escalated substantially while the RMSE is only slightly different (depicted in Figure 11(a)). Figure 11 (a) further indicates that a noise level of 50% for precipitation leads to a spread that is quite similar than the RMSE, the same noise level that was retained by (Thiboult and Anctil, 2015) in the context of the implementation of a multimodel ensemble prediction system in the same geographical region (using a set of lumped models while this study exploits a set of distributed models). The assessment of outcomes from data with and without noise incorporation indicated that better NRRs were achieved with higher perturbation percentages. Specifically, NRR values were consistently above 1 for both 25% and 50% noise levels, while they dipped below 1 for the 75% noise level. It is noteworthy that an ideal NRR value is 1 (Figure 11 (b)).

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026



445 **Figure 11: Exploring the uncertainty by adding noise to the precipitation data. a) Spread skill diagram comparing RMSE and spread for no noise and different percentage of the noises. The boxplot captures the distribution formed by the 18 hydrometric stations available within the SLSO domain. b) Influence of the different precipitation noises on NRR comparing with no noise over the SLSO domain.**

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

To determine the climatic appropriateness of the percentage of additional noise introduced into the system, an assessment is conducted by comparing the precipitation and temperature variability between an ensemble dataset called EMDNA (Tang et al., 2020) and our perturbed data. This comparison is carried out at 10 distinct locations dispersed across the SLSO domain. The information related to these locations is shown in the Figure 12 and Table 6.

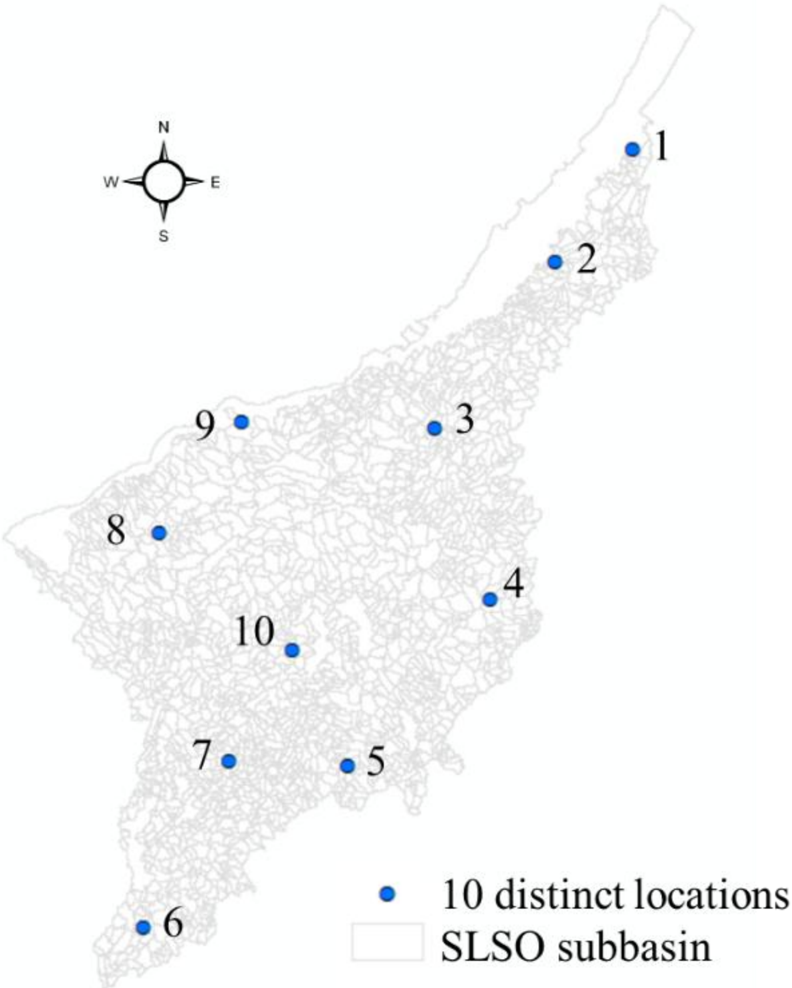


Figure 12: The locations of various sites within the SLSO domain to check the spread of precipitation and temperature using the EMDNA dataset

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Table 6: The information of the 10 locations that the spread of the precipitation and temperature in EMDNA dataset and the noisy (25, 50 and 75%) SLSO data are checked

No.	Longitude	Latitude	Total precipitation (mm)	Average temperature (°C)
1	-69.8	47.6	952.17	4.53
2	-70.2	47.2	1017.22	4.81
3	-70.8	46.6	1184.56	4.41
4	-70.5	46.0	1076.63	4.61
5	-71.2	45.4	1255.59	5.19
6	-72.2	44.8	1221.00	6.38
7	-71.8	45.4	1031.50	6.12
8	-72.2	46.2	1082.62	5.69
9	-71.8	46.6	1148.00	5.20
10	-71.5	45.8	1217.81	4.64

The spread of the precipitation and temperature of the EMDNA are extracted and compared to our noisy dataset in Figure 13. The results of this comparison reveal that the spread of EMDNA closely aligns with the spread observed in the noisy data with a 50% noise addition as illustrated in Figure 13 (a), confirming that the latter is realistic from a climatic perspective. Consequently, the 50% noise level is selected for subsequent uncertainty evaluations. The comparison between the temperature spread showed that, the EMDNA dataset spread is close to 2 °C. Therefore, the 50% noise level of precipitation and 2 °C noise level of temperature are chosen for the further evaluations.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

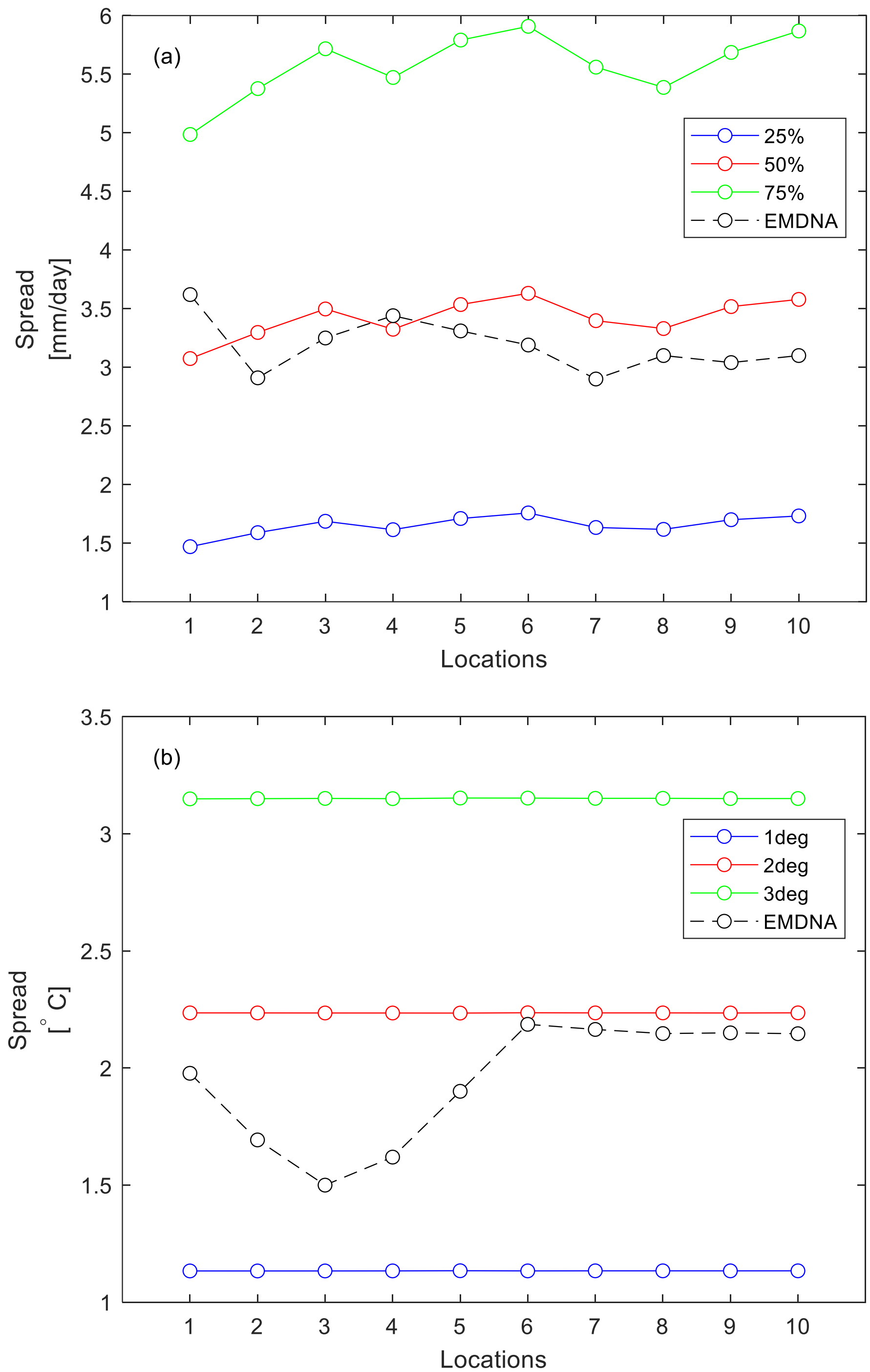
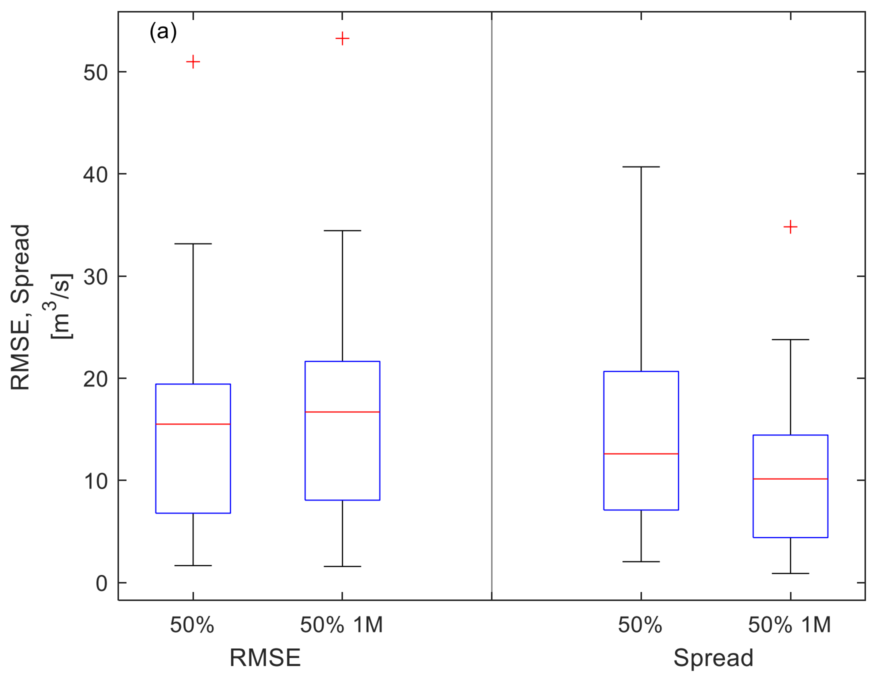


Figure 13: Comparison of the precipitation and temperature spread of the EMDNA data set and of the 48-member ensembles for three different percentage of noises, for 10 locations dispersed across the SLSO domain a) for different precipitation noise levels and b) for different temperature noise levels.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

In Figure 14 (a), the spread and RMSE of 12 models are compared to the initial Raven HBV-EC model, it is shown that using the selected group of the models is doing a reasonable job to explore the uncertainty. In this figure the term 1M refers to using the initial Raven HBV-EC model while others referring to the 12 selected multimodel ensembles. Comparing multimodel ensembles with the use of a single hydrological model revealed that introducing noise to the input meteorological data can help mitigate under-dispersion. This improvement was observed in both scenarios: when employing a single hydrological model and when working with a multimodel ensemble.

The results in terms of the reliability diagram in the Famine watershed illustrated that using a pool of the combination of the 12 models for different percentages of the noises, improved the reliability comparing with using the initial Raven HBV-EC model. In terms of reliability (Figure 14 (b)), using a single hydrological model (green line) did not improve the reliability. However, for the systems when a multimodel approach is used the systems become more robust by improving the reliability. These results suggest that using a combination of multimodel ensembles can explore uncertainty better comparing with using a single hydrological model.



https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

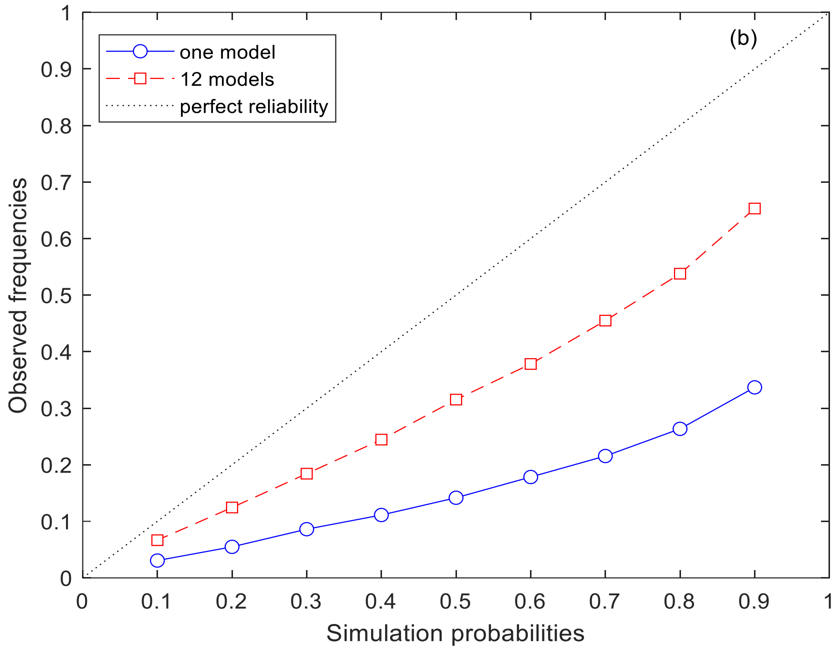


Figure 14: Exploring the uncertainty comparing the multimodel ensemble and one single model over the SLSO domain, the precipitation noise is 50%. a) Spread-skill comparison for one model and 12 selected models to indicate the variation of RMSE and spread. The boxplot captures the distribution formed by the 18 hydrometric stations available within the SLSO domain. b) Reliability diagram comparing one model and 12 models in Famine River hydrometric station (station ID:1).

To sum up, the initial set of 48 models revealed the necessity for improvement. The refinement process involved selecting a subset of 12 models, 25% of the initial models shown in the Figure 6 and introducing noise into the precipitation and temperature data series, resulting in increased reliability. The first conclusions drawn from the initial results are confirmed over the Au Saumon station. Figure 15 illustrates the distribution of the streamflow ensemble, consisting of the 12 selected models, in conjunction with the 48 climate data series with added noise, specifically for the Au Saumon station. A comparison with Figure 6 shows that observed streamflow falls within the 5% to 95% coverage range from April to May. However, for the periods spanning January to April and November to the end of December, the streamflows are consistently underestimated.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

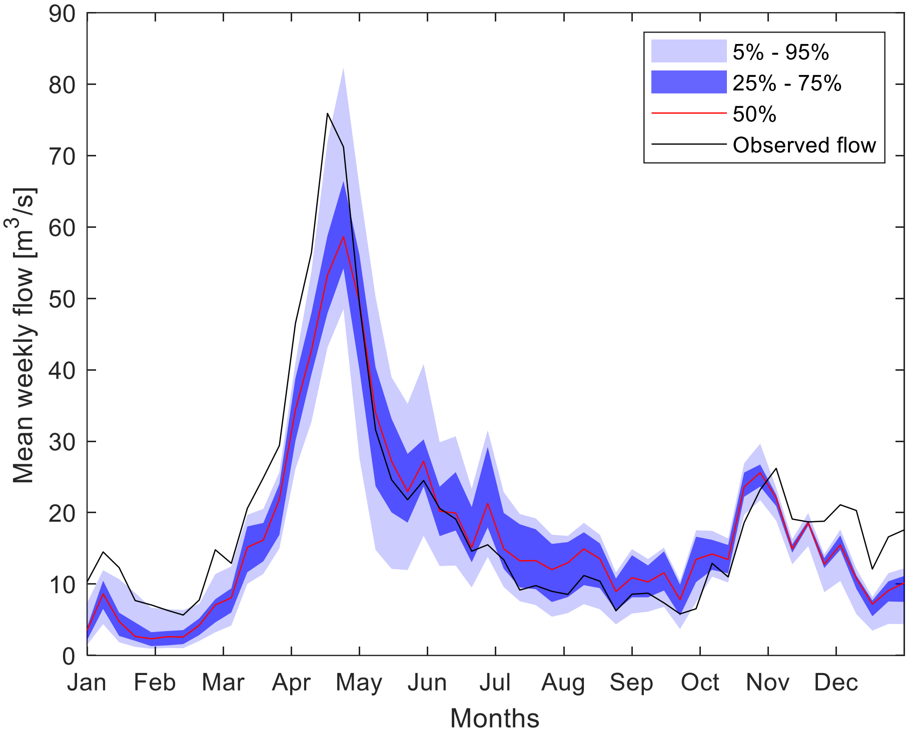


Figure 15: illustration of the comparison between the average annual observed streamflow (the black line) and streamflow ensembles generated by the 12 selected models over the simulation period from 2000 to 2020. It displays the median flow simulation (red line), the 5th to 95th percentile range (the pale blue), and the 25th to 75th percentile range (the dark blue) of the streamflow ensemble at the Au Saumon station.

4 Discussion

The presented paper introduces a framework for assessing and understanding the multifaceted nature of model uncertainty. Hydrological models are essential tools in studying complex hydrological processes, and as the study points out, acknowledging and quantifying uncertainties within these models is crucial for reliable streamflow simulations and forecasts.

This paper focuses on structural uncertainty in distributed hydrological models. While much effort has been invested in improving parameter estimation and refining input data quality, the inherent structural limitations of hydrological models can introduce substantial uncertainties that remain unexplored. By leveraging the Raven hydrologic modeling framework, the study employs a flexible multimodel ensemble approach that provides a more comprehensive exploration of uncertainty stemming from model structure, input data variability and calibration related model parameter uncertainty.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Initially, a pool of 48 different models was created from the original Raven HBV-EC. The flexibility of the Raven model enabled the development of various variants. A sub-model selection using a forward greedy approach demonstrated that a smaller number of models could still provide the benefits of using multi-model ensembles. Investigating the model structure uncertainty stemming from the different Raven HBV-EC variants, each with distinct parameter sets, highlighted another significant source of uncertainty - the meteorological input data.

One of the noteworthy findings is the observation that introducing noise into the input data enhances the reliability and spread of multimodel ensembles, confirming that measurement errors are a substantial source of hydrologic uncertainty. This finding aligns with the understanding that uncertainty is inherent in real-world data, and incorporating this uncertainty within the modeling process results in more robust predictions. The analysis showed that the uncertainty introduced by the model structure and input data is dominating calibration uncertainty.

Furthermore, the study's identification of the optimal noise level offers valuable insights for future studies aiming to replicate similar methodologies. In conclusion, the paper's approach to addressing hydrological model uncertainty through multimodel ensembles, with a specific focus on structural uncertainty, offers insights for advancing the uncertainty hydrological modeling in distributed models. In future research, the ultimate objective is to utilize various climate scenarios to describe the uncertainties within the realm of climate change studies.

5 Conclusion

This study assessed the uncertainty of hydrological model structures by exploring a group of 48 models that are variants of the Raven HBV-EC model. Changes were made to the snow melt, snow process, and potential evapotranspiration methods to create a group of models implemented over the SLSO domain, in the province of Québec, Canada. The forward greedy method was used to make subsets out of the initial 48 models led to selection of 12 models. The results revealed that creating a subset of the initially created models enhances the reliability of multimodel ensembles while reducing the overall number of models used. The evaluation of the model parameter uncertainty in calibration using two models out of 12 models, indicated that the spread and reliability decreased comparing with the multimodel ensemble. The introduction of noise/perturbation into precipitation and temperature data further improved the reliability and spread within the multimodel ensemble system.

Comparison of perturbed precipitation and temperature with an independent dataset (EMDNA) indicated that adding 50% noise to precipitation data and 2˚C was a reasonable choice for this study. The addition of noise to temperature data demonstrated that increasing the noise level led to improved reliability and spread.

The uncertainty evaluation highlighted the enhanced reliability and skill of the multimodel ensemble compared to more traditional single-model approaches. The most effective strategy for reducing uncertainty related to different sources of uncertainty was found to be the development of a multimodel ensemble, achieved by combining various distributed hydrological model structures and introducing noise into the forcing data. Other source of uncertainty is evaluated by exploring the calibration uncertainty for two models. This uncertainty assessment is related to calibration procedure using Ostrich, DDS algorithm. This calibration process is repeated 48 times for each model, with each repetition consisting of 500 iterations. In contrast, uncertainty related to model parameter values, which comes from the calibration of the various models, did not significantly reduce overall uncertainty or improve the model's performance.

Code and data availability

The hydrological modelling framework used in this study is Raven, model setup and execution were conducted using RavenPy, a Python interface for Raven archived at https://doi.org/10.5281/zenodo.7972347 (Huard et al., 2026). The Raven model files, RavenPy workflow scripts, model configuration files, input data, tutorials, and supporting documentation used in this study are permanently archived on Zenodo at https://doi.org/10.5281/zenodo.21859618 (Jabbari et al., 2026). Observed streamflow data were obtained from the Hydroclimatic Atlas of Québec, provided by the Ministère du Développement durable, de l'Environnement et de la Lutte contre les changements climatiques (MDDELCC) (https://www.cehq.gouv.qc.ca/atlas-hydroclimatique/stations-hydrometriques/index-en.htm). Meteorological forcing data were obtained from two sources: the Ensemble Meteorological Dataset for North America (EMDNA), a compilation of 100 members of daily precipitation, mean daily temperature, and temperature range, publicly available at https://doi.org/10.20383/101.0275; and the RSCQ daily climate grids for Québec (GCQ-V3), provided by the Ministère de l'Environnement, de la Lutte contre les Changements Climatiques, de la Faune et des Parcs (MELCCFP) through the Données Québec open-data portal. The RSCQ data were interpolated/extracted for the SLSO region for use in this study. The original RSCQ dataset is distributed under a Creative Commons Attribution (CC BY) license, and the interpolated SLSO data used in this study are included in the Zenodo archive.

Author contributions

AJ, BM, FA, JC, DH, RT, MBM, SLC, and CM contributed to the conceptualization and methodology; AJ, DH, JC, RT, MBM, SLC, and CM contributed to software, resources and investigation; AJ performed the formal analysis, data curation, validation, and visualization; BM and FA contributed to supervision, project administration, and funding acquisition; AJ wrote the manuscript draft; and BM, DH, JC, RT, MBM, SLC, CM, and FA reviewed and edited the manuscript.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Competing interests

The authors declare that they have no competing financial or non-financial interests that could have influenced the research, analysis, or conclusions presented in this study.

Acknowledgements

This work was made possible with funding from the Gouvernement du Québec through the INFO-Crue research program. The authors express their gratitude to the Centre d'Expertise Hydrique du Québec for supplying hydrometeorological data. Special thanks to Juliane Mai and Christian Jauvin for assisting with the sensitive parameter estimation and preparation of RavenPy. Additionally, the authors acknowledge Ouranos for their computing support and resources.

Financial support

The authors acknowledge financial support from the Ministère de l'Environnement, de la Lutte contre les changements climatiques, de la Faune et des Parcs (MELCCFP).

References

Anctil, F. and Ramos, M.-H.: Verification Metrics for Hydrological Ensemble Forecasts, in: Handbook of Hydrometeorological Ensemble Forecasting, edited by: Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L., and Schaake, J. C., Springer Berlin Heidelberg, Berlin, Heidelberg, 893-922, 10.1007/978-3-642-39925-1_3, 2019.

Arsenault, R., Huard, D., Martel, J.-L., Troin, M., Mai, J., Brissette, F., Jauvin, C., Vu, L., Craig, J. R., Smith, T. J., Logan, T., Tolson, B. A., Han, M., Gravel, F., and Langlois, S.: The PAVICS-Hydro platform: A virtual laboratory for hydroclimatic modelling and forecasting over North America, Environmental Modelling & Software, 168, 105808, https://doi.org/10.1016/j.envsoft.2023.105808, 2023.

Beven, K. J.: On hypothesis testing in hydrology: Why falsification of models is still a really good idea, WIREs Water, 5, e1278, https://doi.org/10.1002/wat2.1278, 2018.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H. G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Advances in Water Resources, 32, 129-146, https://doi.org/10.1016/j.advwatres.2008.10.003, 2009.

Brochero, D., Anctil, F., and Gagné, C.: Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria, Hydrol. Earth Syst. Sci., 15, 3307-3325, 10.5194/hess-15-3307-2011, 2011.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, https://doi.org/10.1029/2010WR009827, 2011.

Coxon, G., Freer, J., Lane, R., Dunne, T., Knoben, W. J. M., Howden, N. J. K., Quinn, N., Wagener, T., and Woods, R.: DECIPHeR v1: Dynamic fluxEs and ConnectIvity for Predictions of HydRology, Geosci. Model Dev., 12, 2285-2306, 10.5194/gmd-12-2285-2019, 2019.

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., Mai, J., Serrer, M., Sgro, N., Shafii, M., Snowdon, A. P., and Tolson, B. A.: Flexible watershed simulation with the Raven hydrological modelling framework, Environmental Modelling & Software, 129, 104728, https://doi.org/10.1016/j.envsoft.2020.104728, 2020.

Cuntz, M., Mai, J., Zink, M., Thober, S., Kumar, R., Schäfer, D., Schrön, M., Craven, J., Rakovec, O., Spieler, D., Prykhodko, V., Dalmasso, G., Musuuza, J., Langenberg, B., Attinger, S., and Samaniego, L.: Computationally inexpensive identification of noninformative model parameters by sequential screening, Water Resources Research, 51, 6417-6441, https://doi.org/10.1002/2015WR016907, 2015.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol. Earth Syst. Sci., 13, 913-921, 10.5194/hess-13-913-2009, 2009.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., and Villeneuve, J.-P.: Distributed Watershed Model Compatible with Remote Sensing and GIS Data. I: Description of Model, Journal of Hydrologic Engineering, 6, 91-99, 10.1061/(ASCE)1084-0699(2001)6:2(91), 2001a.

Fortin, J.-P., Turcotte, R., Massicotte, S., Moussa, R., Fitzback, J., and Villeneuve, J.-P.: Distributed Watershed Model Compatible with Remote Sensing and GIS Data. II: Application to Chaudière Watershed, Journal of Hydrologic Engineering, 6, 100-108, 10.1061/(ASCE)1084-0699(2001)6:2(100), 2001b.

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R.: Why Should Ensemble Spread Match the RMSE of the Ensemble Mean?, Journal of Hydrometeorology, 15, 1708-1713, https://doi.org/10.1175/JHM-D-14-0008.1, 2014.

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

H. Hargreaves, G. and A. Samani, Z.: Reference Crop Evapotranspiration from Temperature, Applied Engineering in Agriculture, 1, 96-99, https://doi.org/10.13031/2013.26773, 1985.

Hamilton, A. S., Hutchinson, D. G., and Moore, R. D.: Estimating Winter Streamflow Using Conceptual Streamflow Model, Journal of Cold Regions Engineering, 14, 158-175, 10.1061/(ASCE)0887-381X(2000)14:4(158), 2000.

Huard, D., Smith, T. J., Jauvin, C., Mai, J., Arsenault, R., Arnal, L., and Rondeau-Genesse, G.: CSHS-CWRA/RavenPy: v0.21.0 (v0.21.0), Zenodo [code], 10.5281/zenodo.19056710, 2026.

Jabbari, A., Music, B., Huard, D., Craig, J., Turcotte, R., Bizhanimanzar, M., Lachance-Cloutier, S., Malenfant, C., and Anctil, F.: Raven and RavenPy files, tutorials, and data repository for "Multimodel implementation of the Raven HBV-EC model for streamflow simulation and uncertainty quantification" [dataset], 10.5281/zenodo.21859618, 2026.

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., and Woods, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments, Water Resources Research, 56, e2019WR025975, https://doi.org/10.1029/2019WR025975, 2020.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, Journal of Hydrology, 201, 272-288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., and Schmitt, O.: Estimating the volume and age of water stored in global lakes using a geo-statistical approach, Nature Communications, 7, 13603, 10.1038/ncomms13603, 2016.

Moges, E., Demissie, Y., Larsen, L., and Yassin, F.: Review: Sources of Hydrological Model Uncertainties and Advances in Their Analysis, 10.3390/w13010028, 2021.

Poulin, A., Brissette, F., Leconte, R., Arsenault, R., and Malo, J.-S.: Uncertainty of hydrological modelling in climate change impact studies in a Canadian, snow-dominated river basin, Journal of Hydrology, 409, 626-636, https://doi.org/10.1016/j.jhydrol.2011.08.057, 2011.

Refsgaard, J. C., Christensen, S., Sonnenborg, T. O., Seifert, D., Højberg, A. L., and Troldborg, L.: Review of strategies for handling geological uncertainty in groundwater flow and transport modeling, Advances in Water Resources, 36, 36-50, https://doi.org/10.1016/j.advwatres.2011.04.006, 2012.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., and Franks, S. W.: Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, Water Resources Research, 46, https://doi.org/10.1029/2009WR008328, 2010.

Seiller, G. and Anctil, F.: Climate change impacts on the hydrologic regime of a Canadian river: comparing uncertainties arising from climate natural variability and lumped hydrological model structures, Hydrol. Earth Syst. Sci., 18, 2033-2047, 10.5194/hess-18-2033-2014, 2014.

https://doi.org/10.5194/egusphere-2026-4806
Preprint. Discussion started: 28 September 2026

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth Syst. Sci., 16, 1171-1189, 10.5194/hess-16-1171-2012, 2012.

Seiller, G., Anctil, F., and Roy, R.: Design and experimentation of an empirical multistructure framework for accurate, sharp and reliable hydrological ensembles, Journal of Hydrology, 552, 313-340, https://doi.org/10.1016/j.jhydrol.2017.07.002, 2017.

Shangguan, W., Dai, Y., Duan, Q., Liu, B., and Yuan, H.: A global soil data set for earth system modeling, Journal of Advances in Modeling Earth Systems, 6, 249-263, https://doi.org/10.1002/2013MS000293, 2014.

Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., and Schütze, N.: Automatic Model Structure Identification for Conceptual Hydrologic Models, Water Resources Research, 56, e2019WR027009, https://doi.org/10.1029/2019WR027009, 2020.

Tang, G., Clark, M. P., Papalexiou, S. M., Newman, A. J., Wood, A. W., Brunet, D., and Whitfield, P. H.: EMDNA: Ensemble Meteorological Dataset for North America [dataset], 2020.

Thiboult, A. and Anctil, F.: On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments, Journal of Hydrology, 529, 1147-1160, https://doi.org/10.1016/j.jhydrol.2015.09.036, 2015.

Tolson, B. A. and Shoemaker, C. A.: Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, Water Resources Research, 43, https://doi.org/10.1029/2005WR004723, 2007.

Troin, M., Arsenault, R., Fournier, E., and Brissette, F.: Catchment Scale Evaluation of Multiple Global Hydrological Models from ISIMIP2a over North America, 10.3390/w13213112, 2021.

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

Valdez, E. S., Anctil, F., and Ramos, M. H.: Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems, Hydrol. Earth Syst. Sci., 26, 197-220, 10.5194/hess-26-197-2022, 2022.

Zappa, M., Jaun, S., Germann, U., Walser, A., and Fundel, F.: Superposition of three sources of uncertainties in operational flood forecasting chains, Atmospheric Research, 100, 246-262, https://doi.org/10.1016/j.atmosres.2010.12.005, 2011.