Integrated implementation of the Raven HBV-EC model and describing uncertainty in a multimodel ensemble evaluation

Aida Jabbari, Biljana Music, David Huard, James Craig, Richard Turcotte, Mohammad Bizhani Manzar, Juliane Mai, Simon Lachance-Cloutier, Charles Malenfant and François Anctil

2024

Title:

Integrated implementation of the Raven HBV-EC model and describing uncertainty in a multimodel ensemble evaluation

Aida Jabbari, Biljana Music, David Huard, James Craig, Richard Turcotte, Mohammad Bizhani Manzar, Juliane Mai, Simon Lachance-Cloutier, Charles Malenfant and François Anctil

Abstract

Despite advancements in hydrological modeling, 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 a novel ensemble-based approach to assess hydrological model uncertainty, emphasizing model structure and input data uncertainties concurrently. The study leverages the Raven hydrological modeling framework to create an ensemble of hydrological models. This ensemble is further perturbed with noise to represent input data uncertainty. The approach is demonstrated over the southwest portion of the Saint-Laurent watershed in Canada, evaluating model ensembles against observed streamflow. The forward greedy method aids in selecting sub-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. Additionally, calibration uncertainty and input data uncertainty are evaluated. The results underscore the importance of multimodel ensembles in reducing various sources of uncertainty, with noise-perturbed data improving reliability. 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.

Key words:

Model uncertainty, forward greedy, Noise, Raven HBV-EC, Reliability, Spread

Introduction

In recent years, the hydrological modeling community has placed a significant emphasis on addressing uncertainty within hydrological models (Troin et al., 2022; Knoben et al., 2020; Fu et al., 2015; Butts et al., 2004). To tackle these uncertainties, a range of specific analysis methods have been developed. Nonetheless, most of these investigations predominantly concentrate on assessing uncertainty related to input data and model parameterization within a particular hydrological model, with the model's fundamental structure staying unchanged. Several studies underscoring the significance of adopting a multimodel approach to characterize structural uncertainty (Zappa et al., 2011; Seiler et al., 2012; Troin et al., 2021; Valdez et al., 2022). Moreover, research findings have indicated that, when considering uncertainty partitioning, the influence of model structure uncertainty tends to outweigh uncertainty in parameter estimation (Poulin et al., 2011).

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. Raven was 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 (Craig, 2023). Leveraging Raven allows for the comparison of various combinations of model components, enabling a comprehensive evaluation of how ensemble dynamics impact streamflow simulations.

This study introduces an ensemble-based approach for assessing uncertainty in distributed hydrological models. Multiple 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 analysis focused on assessing uncertainty arising from model structure within the framework of reconstructing historical data, employing a widely recognized approach. The study concludes that the current approach inadequately represents this uncertainty. The paper suggests that the limited spread in results may be attributed to uncertainties in meteorological observations. Through perturbations and comparisons with meteorological reanalysis data that incorporate uncertainties, the research demonstrates a more 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, calibration and input data.

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

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 notable diverse, with 31%, classified as deciduous forest, 31% as agriculture, 25% as coniferous forest, 9% as water bodies, 1.2% as wetland, 1% as impermeable surface, 0.4% 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, it incorporates data from the GlobCover 2009 matrices by the European Space Agency and Circa 2000 matrices by Agriculture and Agri-Food Canada.

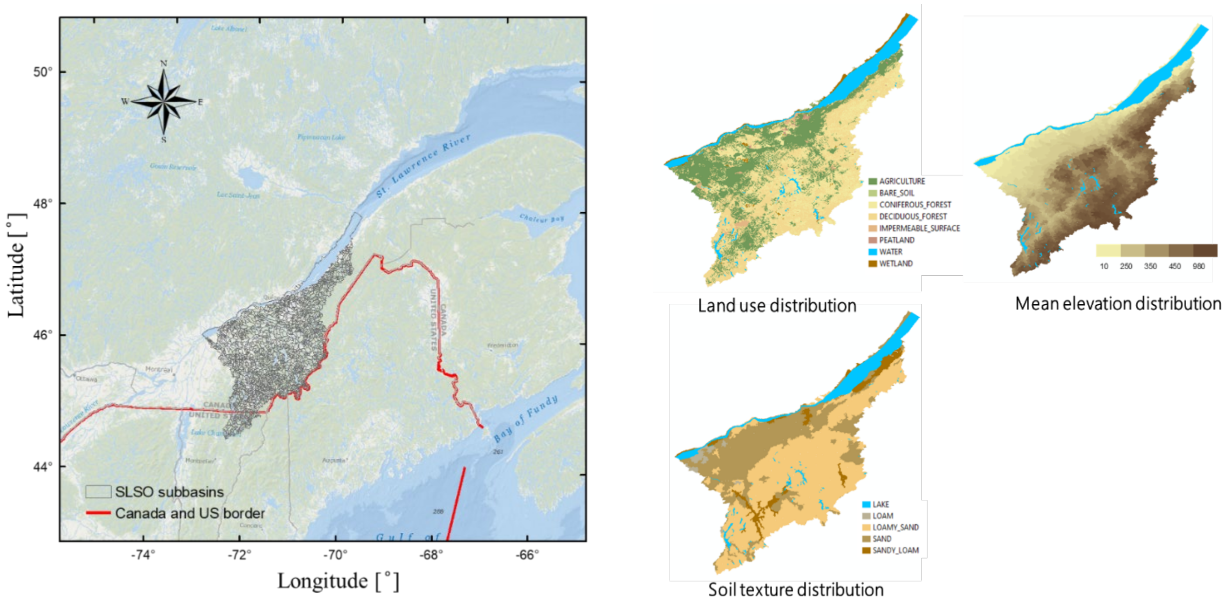


Figure 1: Overview of the SLSO domain location, land use, mean elevation (m), and soil texture distribution.

The soil texture within the SLSO domain is dominated by loamy sand (55%), sand (30%), loam (9.5%) and sandy loam (5.5%), with water body is excluded. 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 of North America with a resolution of 1 km. To assign a textural class for the soil column, vertically weighted values of the clay and sand percentages was 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 assign a class to most of the missing data. The second source corresponds to the data produced by Shangguan et al., 2014.

In the Quebec portion of the SLSO, topographic data with 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. According to this data, the elevation of the ground in the SLSO domain ranges from slightly above sea level near the Saint Laurent River to over 900 meters above it.

2.2. Meteorological data and observed stream flows

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 Direction de l'Expertise Hydrique (DEH) of Quebec. These grids are constructed using ground meteorological stations from both the Réseau météorologique coopératif du Québec (RMCQ) and the Direction de la qualité de l'air et du climat (DQAC). According to this dataset, the SLSO domain experiences an average annual precipitation of around 1092 mm, with about 148 days of rainfall. Solid precipitation contributes to approximately one-third of the total annual precipitation. Additionally, the mean yearly temperature in the SLSO area is roughly 4.4°C, and there are approximately 235 days each year when the maximum temperature surpasses 0°C.

The streamflow data for various hydrometric stations within the SLSO domain are accessible online from 1960 to the present through the following link: https://www.cehq.gouv.qc.ca/atlas-hydroclimatique/stations-hydrometriques/index-en.htm. It's important to note that the domains comprise data from 28 hydrometric stations, and 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 adjusted to account for the effects of ice are omitted during the calibration process.

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 used in the production of multiple editions of the Hydroclimatic Atlas of Southern Québec (2015, 2023). The chosen model is the Hydrologiska Byråns Vattenbalansavdelning – Environment Canada (HBV-EC) model, which represent a 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 (Bergström 1995). The choice was made to employ a version of the HBV-EC model emulated within Raven in distributed mode which operates on a daily time step (Craig et al., 2020). A visual depiction of the Raven HBV-EC model is presented in Figure 2. As mentioned earlier, Raven offers the flexibility to manipulate numerical schemes, interpolation methods, all of which align perfectly with the objectives of this study. Note that the potential evapotranspiration in this study was calculated using an empirical formulation (Hargreaves and Samani, 1985; Craig, 2023), which provides daily potential evapotranspiration values, replacing the original 'monthly' approach used (Bergstrom, 1995). Furthermore, the glacier routine was removed from the model since the study area does not include glacier 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, 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. These combined resources aid in establishing the runoff direction on a cell-by-cell basis. Once the river network delineation is accomplished, the watershed's outlet is identified, subsequently allowing for the determination of the surface area that contributes to the outlet's drainage (equivalent to the total surface area of the watershed).

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. In total, the SLSO domain incorporates 2889 subbasins and 7661 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.

To initiate the simulations, the Raven model relies on configuration files denoted as rv* files (Craig, 2023). For the purpose of setting up, executing, and calibrating the Raven HBV-EC model, RavenPy (available on https://zenodo.org/doi/10.5281/zenodo.7972347) comes into play as a Python wrapper to the Raven hydrological framework, designed to simplify the setup of hydrological models, the initiation of simulations, and the extraction of results (Arsenault et al., 2023).

RavenPy, an open-source software solution, serves as a versatile tool for generating rv* files, running the model, and subsequently accessing the model's output. Notably, RavenPy facilitates the launch of parallel Raven simulations.

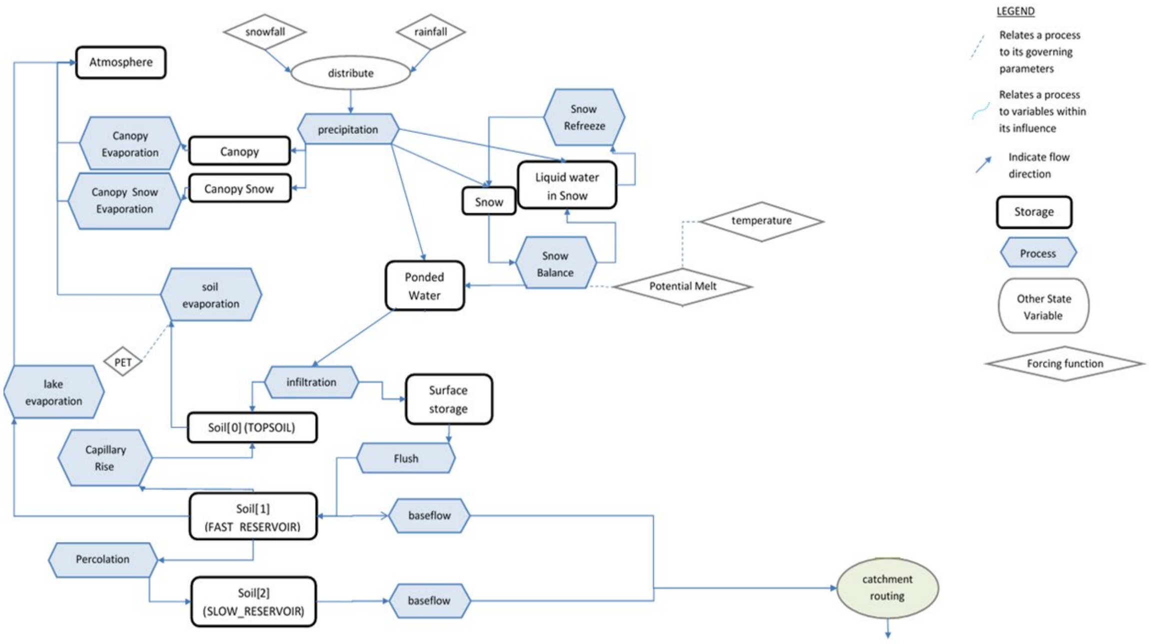


Figure 2: The schematic of Raven HBV-EC (Craig, 2023).

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 (Morris, 1991) was employed, utilizing 28 model parameters, 100 trajectories and one iteration (resulting in 2800 model evaluations) to analyze candidate parameters. 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 1). As mentioned before, the soil model used in this study comprises three layers and various textures. As a result, there are different model parameters for each layer and texture, resulting in a total of 31 sensitive model parameters.

Table 1: 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

Model calibration and validation constitute integral aspects of hydrological modeling. Calibration is a two-fold procedure. In the initial stage, model parameters are manually assessed through trial and error to establish initial values. Subsequently, an automatic parameter assessment strategy is implemented, employing a numerical optimization technique. With the initial model parameters in place, OSTRICH (Matott, 2013) is employed. OSTRICH is a versatile optimization and calibration tool that is independent of the model, employing multiple algorithms. This aids in refining the sensitive parameters based on the Kling-Gupta efficiency (KGE, Gupta et al., 2009). The calibration is executed using the Dynamically Dimensioned Search (DDS) algorithm (Tolson and Shoemaker, 2007), run for 500 iterations and repeated 5 times. Additionally, calibration iterations of 1000, 2000, and 4000 were performed.

We employed 28 hydrometric stations for validation and calibrated the model using 18 hydrometric stations due to data constraints. 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.

2.5. Uncertainty assessment

2.5.1. Diverse algorithm combinations for 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 3). 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 (Craig, 2023). In the next step, all variants are calibrated using the approach described above, and calibrated model's simulations are used to explore the uncertainty.

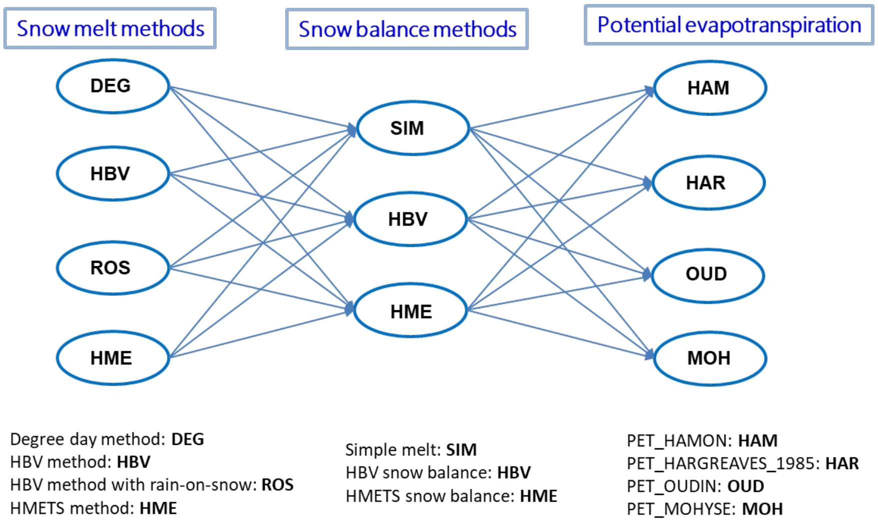


Figure 3: Configurations of snow melt, snow balance, and potential evapotranspiration components leading to 48 unique models.

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, a sub-selection of models is devised. The sub-model selection is done to have 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 distinct sets are established. The set-0 comprises all 48 models, while the set-1 encompasses models chosen through the forward greedy approach, prioritizing performance enhancement by reducing the root mean square error (RMSE). The set-2 consisting of models selected through forward greedy optimization, aimed at minimizing the discrepancy between the RMSE and the square root of the average ensemble variance, referred to here as spread (Fortin et al. 2014). Note that this discrepancy is taken here as an estimate of reliability. The set-3 incorporates models chosen based 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 Ensemble Kalman Filter (EnKF), for state-parameter estimation, the forcing data is perturbed by adding noise to the variables (observations), to ultimately generate reliable ensembles in which the spread correctly reflects the uncertainty. Here the variance of this noise is set proportional to the magnitude of each variable

In this study, to address input data sources of uncertainty, random white noise perturbation is added to the precipitation and temperature data. We followed Thiboult and Anctil's research in 2015, which involved creating three precipitation noise levels (a standard deviations equivalent to 25%, 50%, and 75% of the mean value with a gamma distribution) and three temperature noise levels (a standard deviation of 1, 2, and 3 °C with a normal distribution). For each magnitude of the noise added to the meteorological data, an ensemble of 48 members is created.

To add noise to the precipitation data, the random data with a gamma distribution are generated and multiplied to the precipitation data. To generate random data by gamma law, shape (k) and scale (θ) parameter should be calculated using the following equations. In which the σ is the standard deviation and P_{mean} is the mean of the precipitation.

$$k = \frac{p_{mean}^2}{(\sigma(\%) * p_{mean})^2}$$

$$\theta = \frac{(\sigma(\%) * p_{mean})^2}{p_{mean}}$$

For example when the noise is equivalent to 50% of the standard deviation of the mean value, the k =4 and θ = 0.25, such that mean = 1 and $\sigma = 0.5$.

To add noise to the temperature data, random data with a normal distribution centered on zero and standard deviation set to 1, 2 and 3 °C. It should be noted that to generate the noisy data ensembles, for each of the noise values 48 climate date series are created for precipitation and temperature respectively.

In order 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 as presented by Tang et al. (2021). This comprehensive dataset spans daily temperature range as well as daily precipitation and temperature data, 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 arise from many sources, including the calibration process as DDS does not systematically converge. These dissimilar parameterizations involve the key parameters that govern the behavior of the model. To facilitate a comprehensive comparison of diverse uncertainty types, this study also delves into the uncertainty related to model calibration. This is achieved by selecting two models from the pool of available ones and subjecting them to a calibration process repeated 48 times, each involving 500 iterations through the utilization of the Ostrich optimization tool.

2.5.5. Evaluation metrics

This study involves a comparison between the root mean square error (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) (Abaza et al., 2014):

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

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 desirable (reliable). NRR>1 implies that the ensemble is too narrow, and NRR<1 shows that it is too wide.

Furthermore, the Spread Skill plot is constructed, which provides an additional angle of verification to the NRR. The spread skill diagram serves as a visual tool to comprehensively assess the ensemble's bias, spread, and overall reliability (Thiboult & Anctil, 2015). This diagram operates under the fundamental principle that achieving reliability requires the RMSE to be in line with the spread (Fortin et al., 2014).

A graphical representation of the ensemble reliability is provided by the reliability diagram (Wilks, 2011), 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 done over 2007 – 2017, with one year spin up time to initiate hydrological processes (not evaluated within the objective function), considering the pool of hydrometric time series across the SLSO domain. The validation allows to judge 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. For example, the model-simulated streamflow closely aligns with the observed flow at the Bécancour (KGE = 0.70) and the Nicolet (KGE = 0.73) stations. The model performance is shown in Figure 4, depicting its ability to replicate streamflow patterns at the hydrometric stations from 1961 to 2020. Notably, both the Bécancour station (KGE = 0.66) and the Nicolet station (KGE = 0.70) display favorable levels of agreement. The main modeling issue lies with the winter low flows that are underestimated by the model (Figure 4).

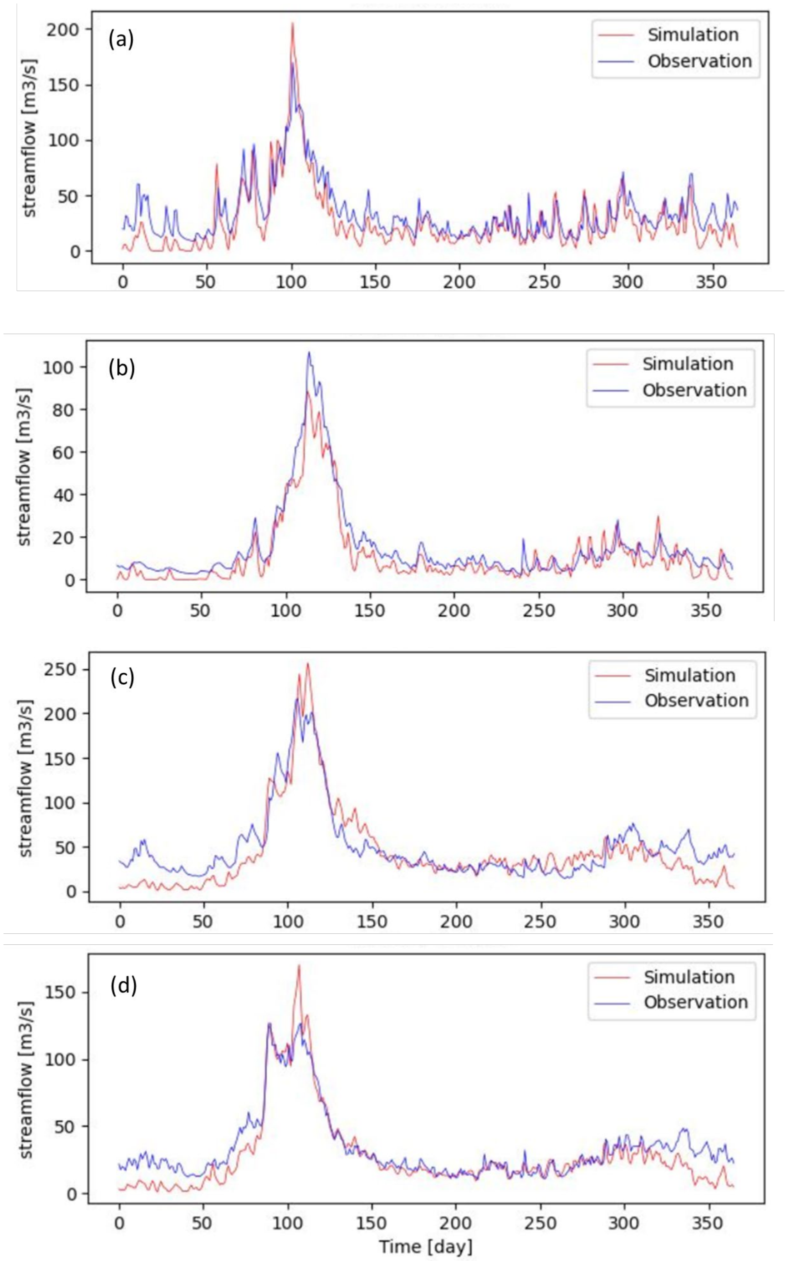


Figure 4. Mean daily streamflows for all years of a) and b) calibration (2007– 2017) in Bécancour and Nicolet stations respectively and c) and d) validation (1961 – 2020) of Raven HBV-EC in Bécancour and Nicolet stations for simulated and observed flows.

3.2. Evaluation of the multimodel ensemble

Following the evaluation of Raven HBV-EC calibration performance using the KGE metric, an ensemble of models is created to assess hydrological uncertainty. 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 assess uncertainty.

The average weekly streamflow simulations over the period 2000–2020 for the Au Saumon station are shown in Figure 5, 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. The comparison of the distribution of streamflow ensembles indicated that the largest spread occurs during the mid-April spring floods while the smallest spread occurs in November and December. These findings validate that capturing high flows is more challenging compared to low flows, likely due to their erratic nature (Seiller and Anctil, 2014). The figure also illustrates the need for accounting for more sources of uncertainty (to increase spread), an issue that is dealt in a subsequent step below.

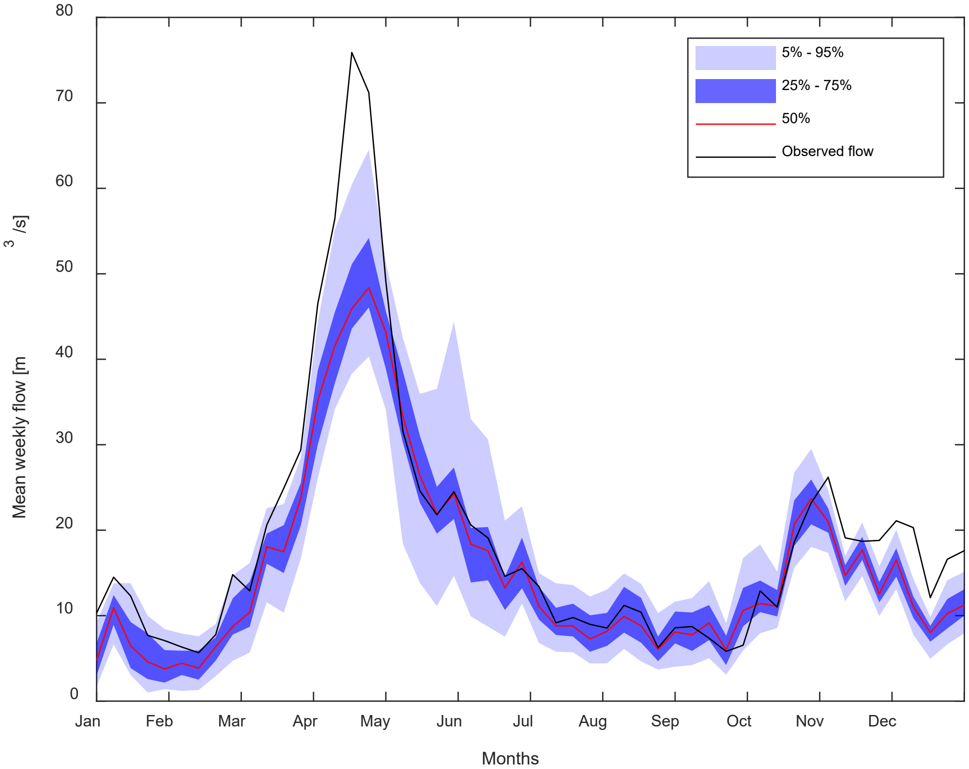


Figure 5: Comparison of the average annual observed streamflow and streamflow ensembles from 48 models during the simulation period (2000 – 2020) depicting median flow simulation, 5th to 95th percentile range, and 25th to 75th percentile range of the streamflow ensemble in Au Saumon station.

3.3. Uncertainty assessment

Multimodel ensembles were constructed in the form of the 48 different models. 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. Figure 6 illustrates the assessment of multimodel ensembles using a spread skill diagram. The plotted results of the spread and the RMSE in Figure 6 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 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 input uncertainty is notably missing. The differences in values in Figure 6 reflects some local performance issues from one hydrometric site to the other but more importantly the fact that each site corresponds to a different drainage area (mean streamflow).

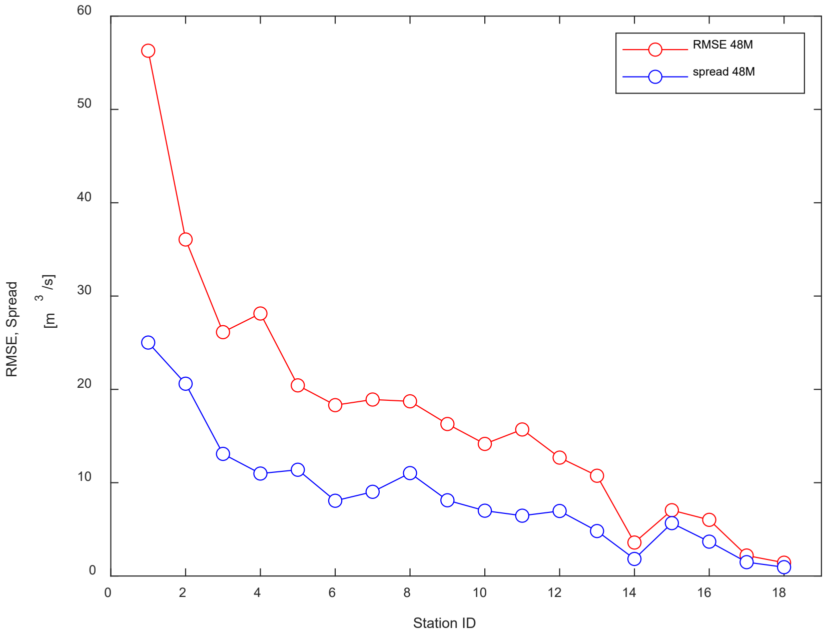
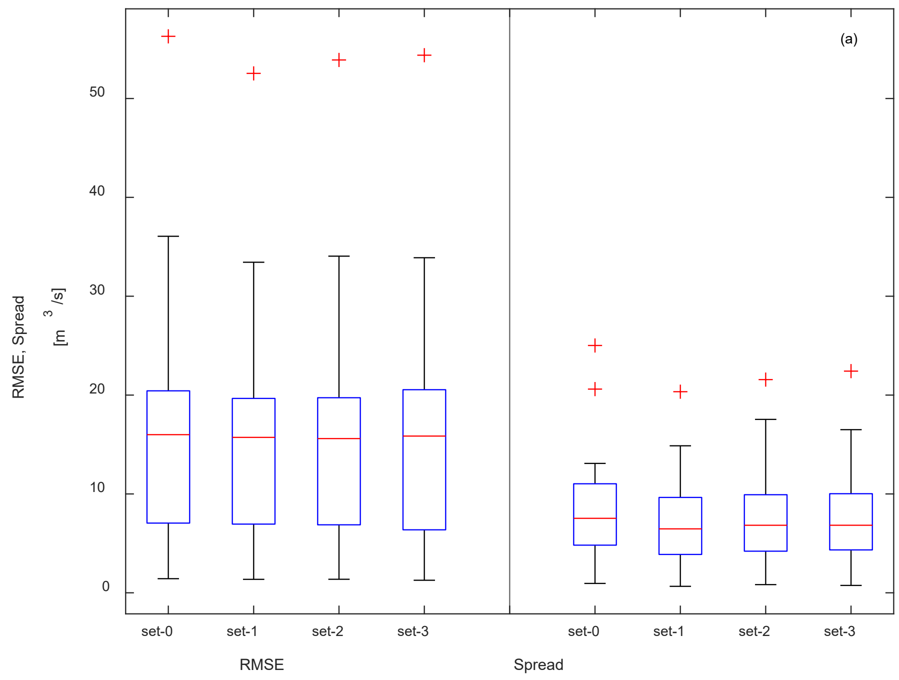


Figure 6: Verification of the 48-member multimodel: RMSE and spread values for each of the 18 hydrometric stations. The hydrometric stations are arranged in descending order of size, with the largest ones coming first.

The subsequent phase of uncertainty assessment for the multimodel ensembles involves proposing various model subsets to seek performance enhancement through model selection and the reduction of the number of models within a multimodel ensemble group. The selection of each hydrological model member of the ensemble can be done in different ways. 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. The created sets are then compared with the results obtained by the 48 model ensembles, named set-0, shown in Figure 7 (a). The comparison between set-1 and set-0 revealed that the selected 12 models exhibited a decrease in RMSE and an improvement in spread. 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.

The different sets show 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. 7 (b)). Overall, these results confirm that it is possible to work with a smaller number of models, which is operationally and computationally interesting, without loss in performance or reliability. In fact, small gains are achieved. The distribution depicted by the boxplots in Figure 6 reflects performance and drainage area divergences from one hydrometric site to the other, as already mentioned.



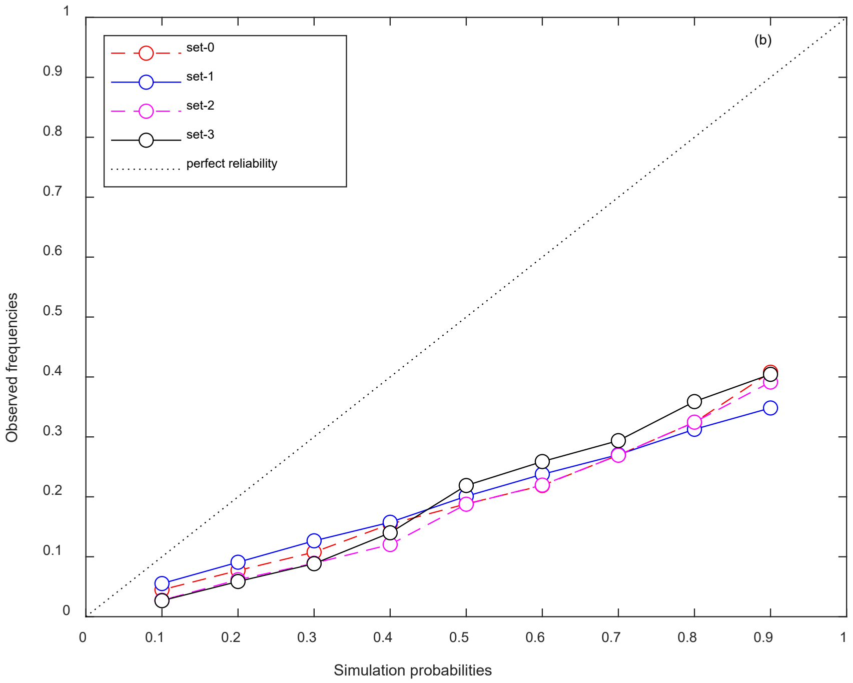


Figure 7: Exploring the uncertainty for different sets. a) Comparison of the RMSE and spread for set-0, set-1, set-2 and set-3. The boxplot captures the distribution formed by the 18 hydrometric stations available within the SLSO domain. b) Reliability diagrams for set-0, set-1, set-2 and set-3, for 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 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 8 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.

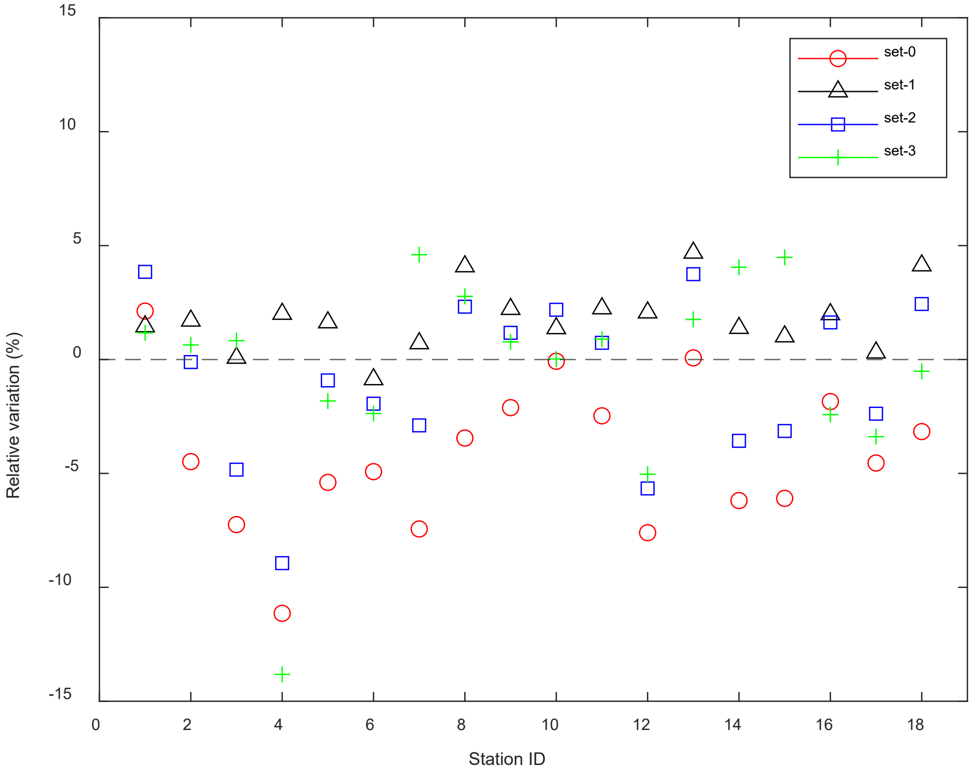


Figure 8: Relative RMSE variation for different sets comparing with the best individual model in each hydrometric station.

By comparing all performance values in Figure 8, 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 2).

Table 2: Model names in Set-1 and their method references. The first three characters denote the snow melt method, followed by the next three characters signifying the snow balance method, with the remaining characters indicating the potentiation evapotranspiration method.

No.	Model name abbreviation
M1	DEG-SIM-HAM
M2	DEG-SIM-HAR

M5	DEG-HBV-HAM
M6	DEG-HBV-HAR
M7	DEG-HBV-OUD
M10	DEG-HME-HAR
M14	HBV-SIM-HAR
M19	HBV-HBV-OUD
M21	HBV-HME-HAM
M30	ROS-HBV-HAR
M31	ROS-HBV-OUD
M34	ROS-HME-HAR

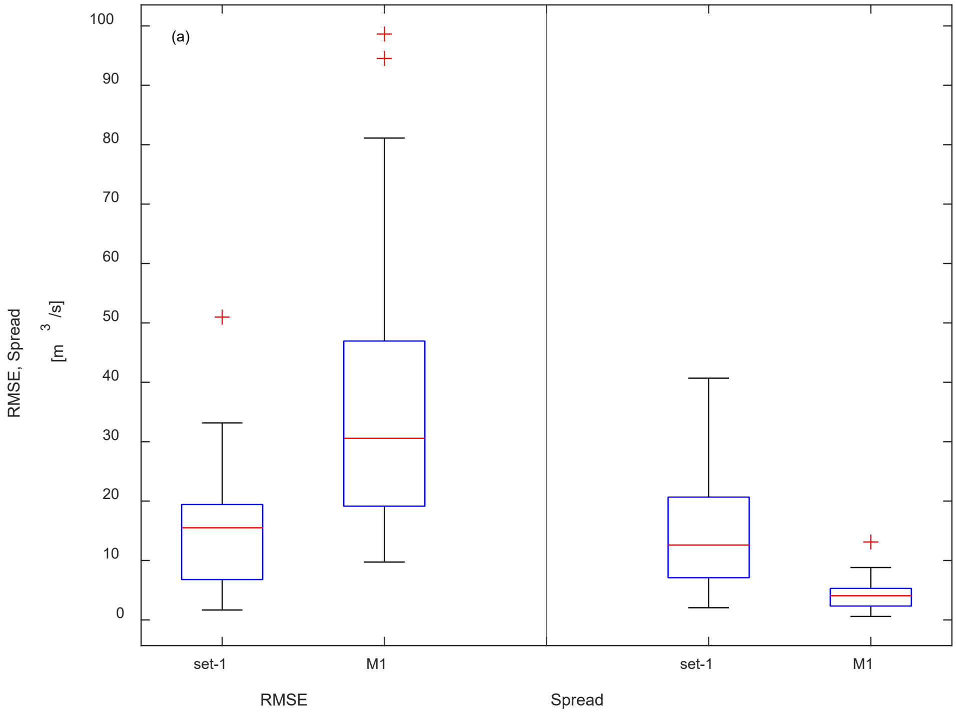
Following the selection of a subset of models for uncertainty exploration, the subsequent phase of the study involved exploring the uncertainty associated to the identification of the effective parameters of the models. Uncertainty in these parameters can arise due to various factors. Firstly, the inability 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 due to this uncertainty.

The assessment of parameter-related 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 parameter 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 9 depict the impact of evaluating model parameter uncertainty for Model M1 and the 12 selected models (set-1 listed in Table 2) in the spread-skill (a) and reliability diagram (b). These results reveal a consistent trend: as model parameter-related 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 exploring model parameter uncertainty, Models 1 and 14 experience a noticeable decrease in spread accompanied by a simultaneous rise in RMSE.

Figure 9(b) depicts a reliability plot that compares model parameter 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 model parameter 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 model parameter uncertainty alone. This evaluation underscores the superior reliability and skill of the multimodel ensemble in contrast to the more traditional single-model ensemble approach.



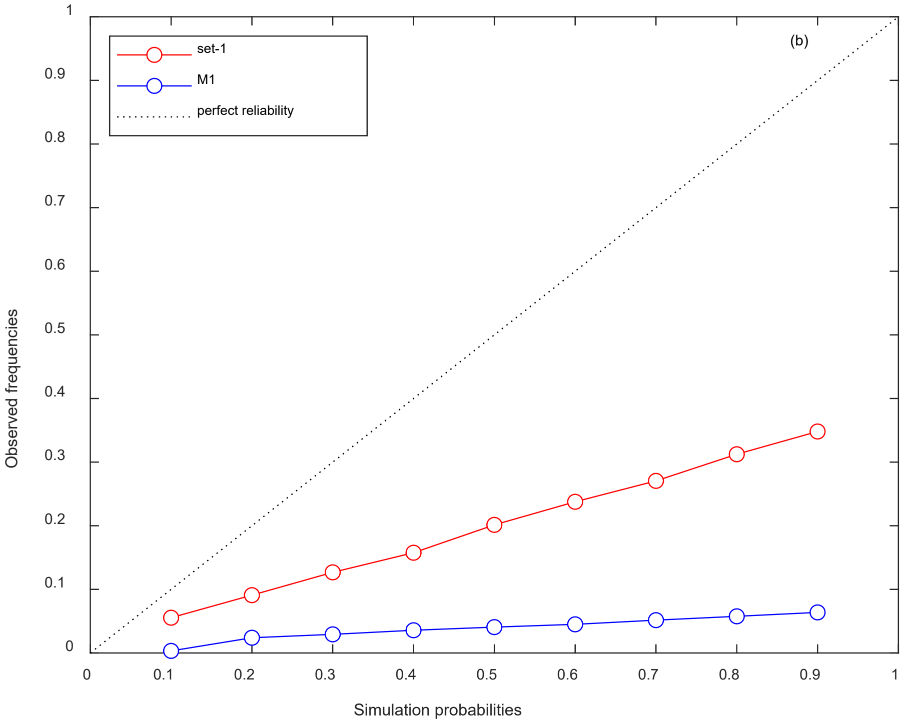


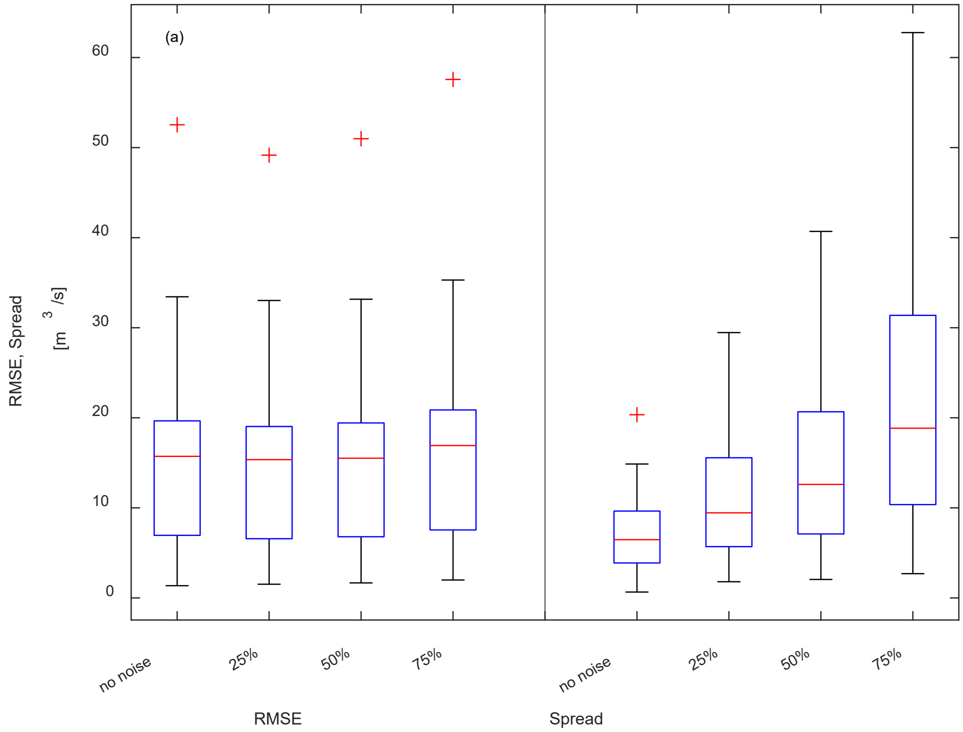
Figure 9: Evaluation of the model parameter uncertainty. a) Spread-skill diagram for set-1 models and different model parameters for model M1 (DEG-SIM-HAM) calculated over the evaluation period. b) Reliability diagram investigating the model parameter uncertainty in Famine River hydrometric station (station 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 model parameter 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.

Following the selection of a subset of models, and exploring the model parameter uncertainty, the subsequent phase of the study involved introducing noise to the input data to assess 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 10(a)), which was expected from the EnKF literature already stated. Figure 10 (a) further indicates that a noise level of 50% for precipitation lead to a spread that is quite similar than the RMSE, the same noise level that was retained by Thiboult and Anctil (2015) in their EnKF analysis 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 10 (b)).



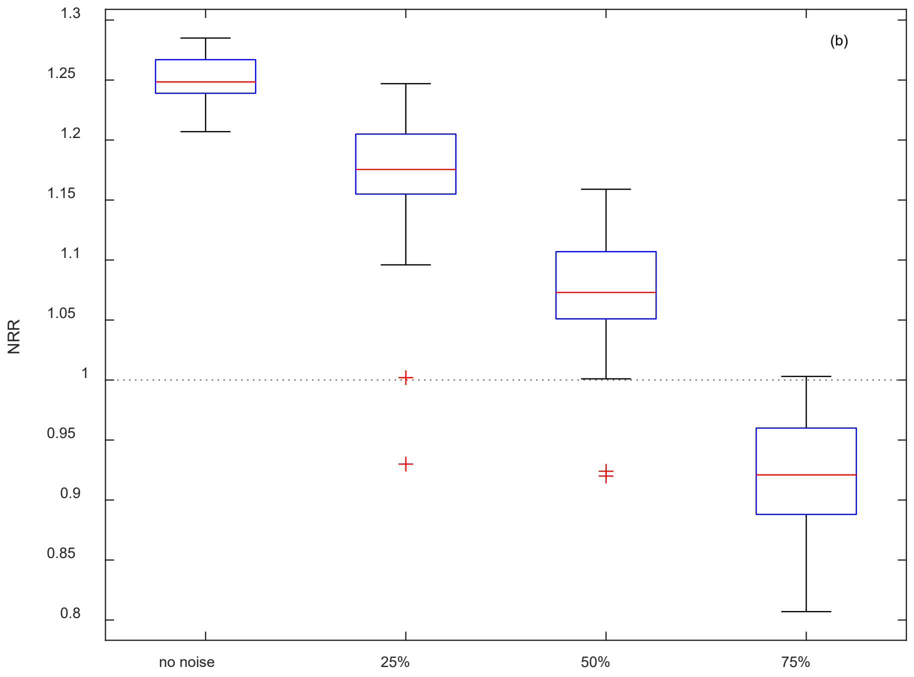


Figure 10: Exploring the uncertainty by adding noise to the precipitation data. a) Spread skill diagram comparing RMSE and spread for no noise and different percentage of the noises. b) Influence of the different precipitation noises on NRR comparing with no noise over the SLSO domain.

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 and our perturbed data. This comparison is carried out at 10 distinct locations dispersed across the SLSO domain. The spread of the precipitation and temperature of the EMDNA are extracted and compared to our noisy dataset in figure 11. 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 11 (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.

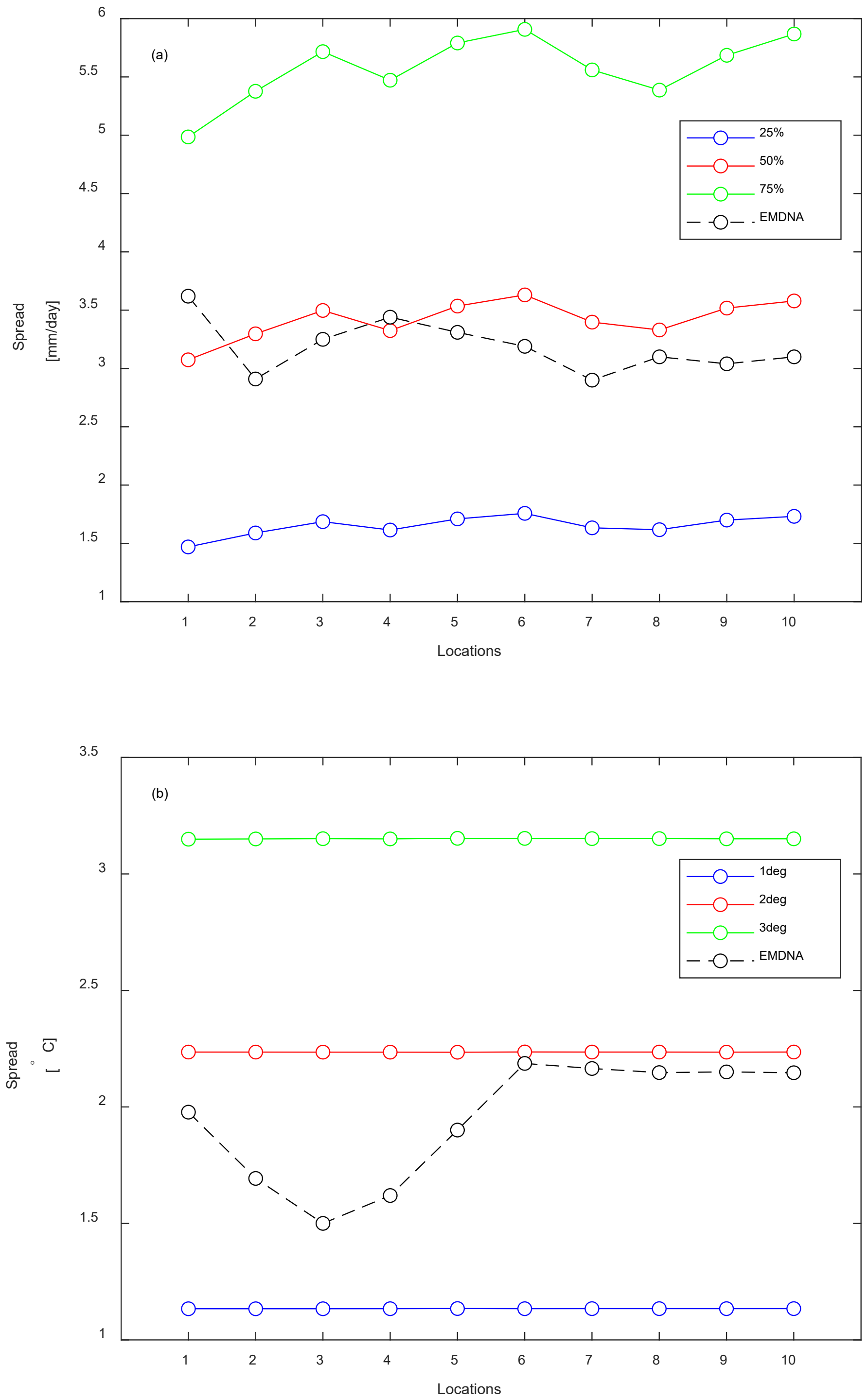


Figure 11: Comparison of the precipitation and temperature spread of the EMDNA data set and of the 48-member ensembles for three different percentage of noises, for 10 locations dispersed across the SLSO domain a) for different precipitation noise levels and b) for different temperature noise levels.

To examine the fluctuations in the RMSE and spread of the multimodel ensemble, a spread-skill diagram is generated, allowing for a comparison between results obtained with the inclusion of different noises (1, 2 and 3°C) in the temperature data. Illustrated in Figure 12 (a), the comparison focuses on the influence of different noise levels introduced into the temperature data. By incorporating a noise level that is 50% of the precipitation data's noise, the figure vividly demonstrates an enhancement in the data spread. Notably, as the noise added to the temperature data increases, the spread becomes more pronounced. This finding underscores the sensitivity of the spread to the magnitude of noise introduced. In light of these findings, noise levels of 50% and 2°C were selected for further uncertainty evaluation in subsequent steps.

In Figure 12 (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 utilization of a single hydrological model revealed that introducing noise to the input meteorological data could enhance the mitigation of under-dispersion. This improvement was observed in both scenarios: when employing a single hydrological model and when working with a multimodel ensemble. However, it's noteworthy that in the case of the multimodel ensemble, the addition of noise specifically contributed to the enhancement of under-dispersion.

The results in terms of the reliability diagram in 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 (Fig. 12(b)), using a single hydrological model (green line) did not improve the reliability however for the systems when a multimodel approach is used (the 12 selected ones) the systems become more robust by improving the reliability (the green and blue lines). These results suggest that using a combination of multimodel ensembles can explore uncertainty better comparing with using a single hydrological model.

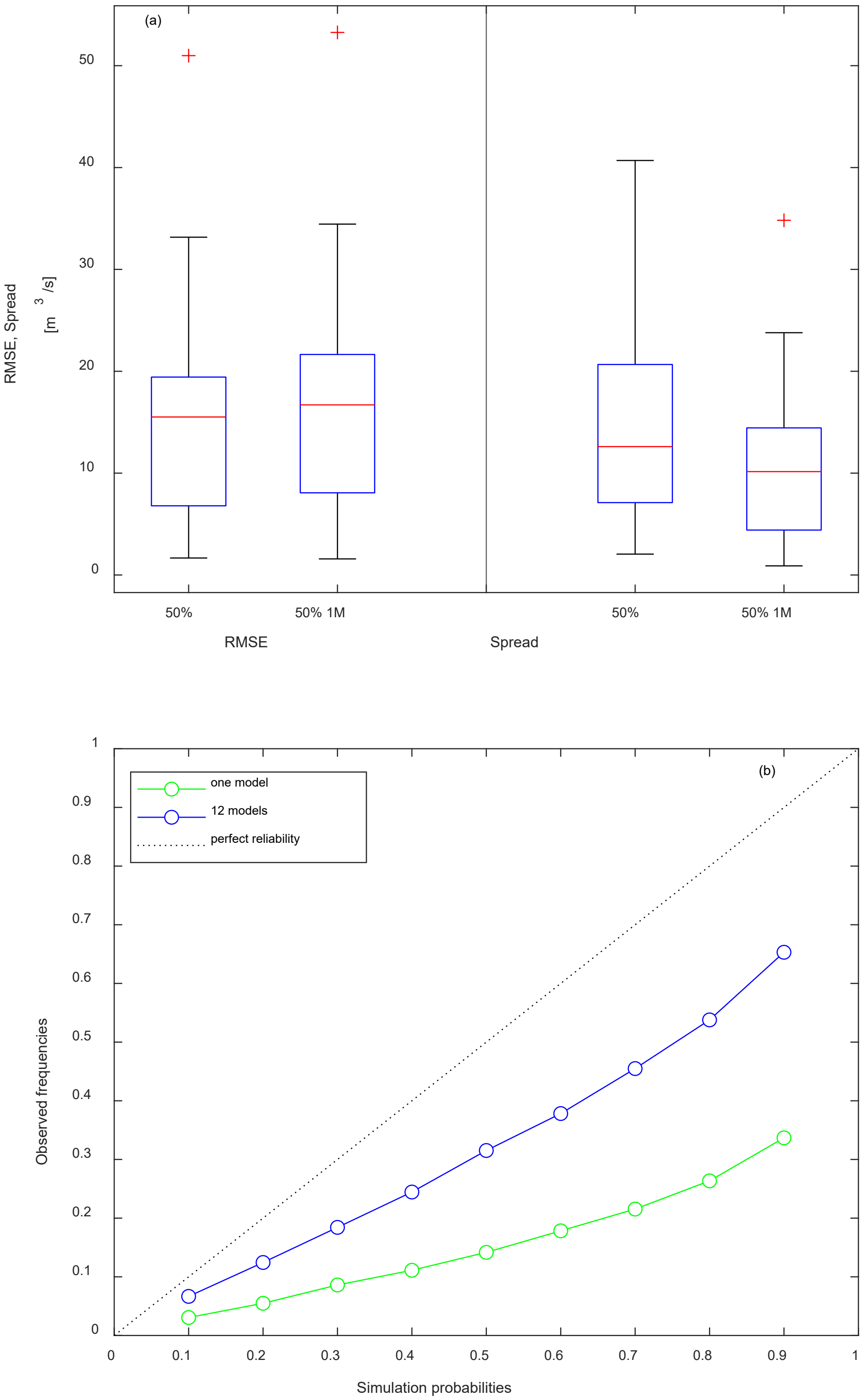


Figure 12: Exploring the uncertainty comparing the multimodel ensemble and one single model over the SLSO domain, the precipitation noise is 50%. a) Spread-skill comparison for one model and 12 selected models to indicate the variation of RMSE and spread. b)

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 5 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 13 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 5 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.

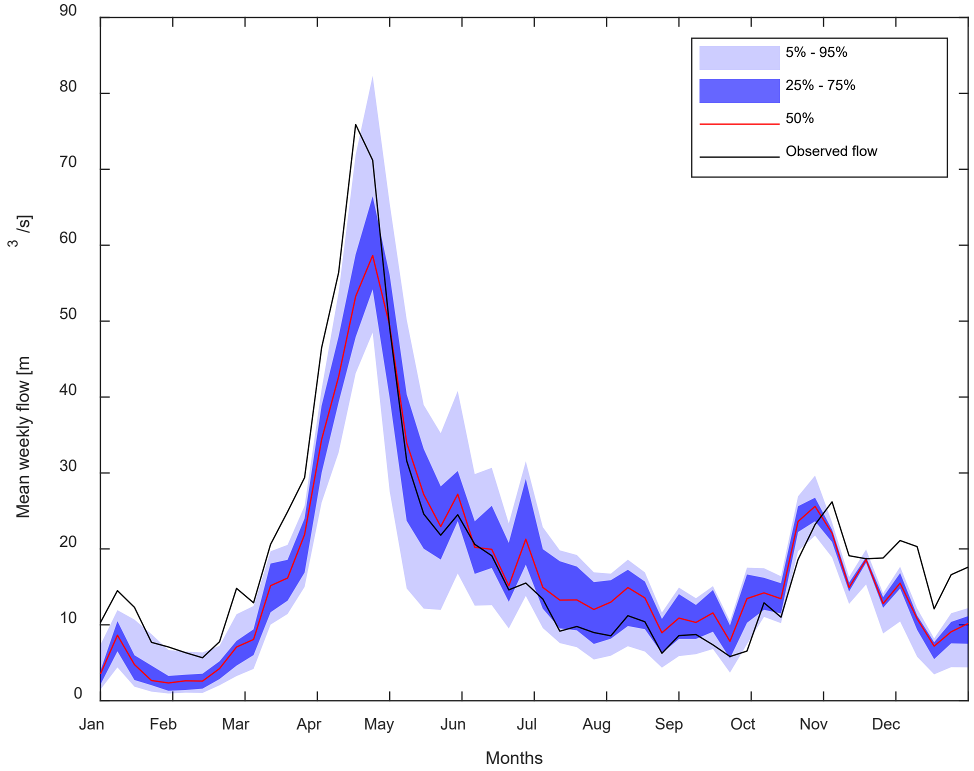


Figure 13: illustration of the comparison between the average annual observed streamflow (the black line) and streamflow ensembles generated by the 12 selected models over the simulation period from 2000 to 2020. It displays the median flow simulation (red line), the 5th to 95th percentile range (the pale blue), and the 25th to 75th percentile range (the dark blue) of the streamflow ensemble at the Au Saumon station.

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 the group of the 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. 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 using two models out of 48 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. However, when considering the spread, the variation among different noise values for temperature was small. As a result, a noise value of 2°C was considered reasonable for subsequent evaluations.

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 hydrological model structures and introducing noise into the forcing data. In contrast, uncertainty related to model parameter values did not significantly reduce overall uncertainty or improve the model's performance.

6. 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 which is an important step forward in addressing exploring the uncertainty. 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 comprehensive exploration of uncertainty stemming from model structure, input data variability and parameter estimation.

The paper's detailed implementation process, from model calibration to the creation of multimodel ensembles, demonstrates a robust methodology that enhances the study's credibility. The results, presented in various diagrams and figures, effectively communicate the effects of different uncertainties on model performance. Particularly notable is the comparison between single-model simulations and multimodel ensemble simulations, which underscores the value of employing diverse model structures to capture a broader range of uncertainties.

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. 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 future research, the ultimate objective is to utilize various climate scenarios to describe the uncertainties within the realm of climate change studies.

References:

Abaza M, Anctil F, Fortin V, Turcotte R. 2014. Sequential streamflow assimilation for short-term hydrological ensemble forecasting. Journal of Hydrology 519, 2692-2706. DOI: 10.1016/j.jhydrol.2014.08.038.

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., and Cloke, H. L., and Schaake, J. C., Springer Berlin Heidelberg, 1–30, https://doi.org/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. and Logan, T., 2023. The PAVICS-Hydro platform: A virtual laboratory for hydroclimatic modelling and forecasting over North America. Environmental Modelling & Software, 168, p.105808.

Bergström, S., 1995. The HBV model. In: V.P. Singh, ed. Computer models of watershed hydrology. Highland Ranch, CO: Water Resources Publications, 443–476.

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

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., et al. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. Environmental Modelling & Software, 129, 104728. https://doi.org/10.1016/j.envsoft.2020.104728.

Craig, J. R. (2023). Raven: User's and developer's manual v3.7. The Raven Development Team. http://raven.uwaterloo.ca/Downloads.html.

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., & Samaniego, L. (2015). Computationally inexpensive identification ofnoninformative model parameters by sequential screening. Water Resources Research, 51(8), 6417–6441. https://doi.org/10.1002/2015WR016907.

Fortin, V., Abaza, M., Anctil, F., Turcotte, R., 2014. Why should ensemble spread match the RMSE of the ensemble mean? J. Hydrometeorol. 15, 1708–1713.

Fu, C., James, A.L., Yao, H., 2015. Investigations of uncertainty in SWAT hydrologic simulations: a case study of a Canadian Shield catchment. Hydrol. Process. 29, 4000–4017. https://doi.org/10.1002/hyp.10477.

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, J. Hydrol.,377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hargraves, G.H. and Z.A. Samani, (1985) Reference Crop Evapotranspiration from Temperature, Appl. Eng. Agric., vol 1, no. 2, pp. 96-99.

Hydroclimatic Atlas of Southern Québec; Government of Québec: Quebec City, QC, Canada, 2015 and 2023. Centre d'Expertise Hydrique du Québec; Ministère du Développement Durable, de l'Environnement et de la Lutte Contre les Changements Climatiques.

Knoben, W.J.M., Freer, J.E., Peel, M.C., Fowler, K.J.A., Woods, R.A., 2020. A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. Water Resour. Research 56, 1–23. https://doi.org/10.1029/2019WR025975.

Lindström, G., et al., 1997. Development and test of the distributed HBV-96 hydrological model. Journal of Hydrology, 201 (1–4), 272–288. doi:10.1016/S0022-1694 (97)00041-3.

Matott, L., 2013. an Optimization Software Tool, Documentation and User's Guide, Version 17.12.19. Univ. Buffalo Cent. Comput. Res.

Morris, M. D. (1991), Factorial sampling plans for preliminary computational experiments, Technometrics, 33(2), 161–174.

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, J. Hydrol., 409, 626–636, https://doi.org/10.1016/j.jhydrol.2011.08.057, 2011.

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, https://doi.org/10.5194/hess-16-1171-2012, 2012.

Seiller, G.; 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. 2014, 18, 2033–2047.

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

Shangguan, W., Dai, Y., Duan, Q., Liu, B., & Yuan, H. (2014). A global soil data set for earth system modeling. Journal of Advances in Modeling Earth Systems, 6(1), 249-263.

Tang, G., Clark, M. P., Papalexiou, S. M., Newman, A. J., Wood, A. W., Brunet, D., & Whitfield, P. H. (2021). EMDNA: Ensemble meteoro-logical dataset for North America. Earth System Science Data, 13, 3337–3362. https://doi.org/10.5194/essd-13-3337-2021.

Thiboult, A., & Anctil, F. (2015). On the difficulty to optimally implement the ensemble Kalman filter: An experiment based on many hydrological models and catchments. Journal of Hydrology,529(Part 3), 1147–1160.

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

Troin, M., Arsenault, R., Fournier, E., Brissette, F., 2021. Catchment Scale Evaluation of Multiple Global Hydrological Models from ISIMIP2a over North America. Water 2021 (13), 3112. https://doi.org/10.3390/w13213112.

Troin, M., Martel, J.-L., Arsenault, R., & Brissette, F. (2022). 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.

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

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences (3rd edn.), Academic Press, Oxford, UK, 2011