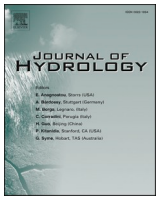
Journal of Hydrology 658 (2025) 133186

Contents lists available at ScienceDirect

Journal of Hydrology

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



Research papers

Value of various elements of the hydrological forecasting chain: Is there a successful pathway for improving the overall performance?

Jonathan Davidson-Chaput [a,*], Richard Arsenault [a], Jean-Luc Martel [a], Magali Troin [a,b]

[a] Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Université du Québec, 1100 Notre-Dame Street West, Montreal, Quebec H3C 1K3, Canada

[b] HydroClimat, TVT, Maison du Numérique et de l'Innovation, place Georges Pompidou, 83 000 Toulon, France

A B S T R A C T

Water resources management relies heavily on hydrological forecasting, and continuous improvements are made to better manage hydropower reservoirs and improve their profitability. The significance of hydrological forecasting has been extensively studied in the literature, but the relative value of the various elements composing a forecasting system has been less investigated to date, making it hard to pinpoint which element to focus research on to improve the overall profitability of hydropower systems. This paper investigates if one or more of the following four elements of the hydrological forecasting chain has more impact on the variance in profit generation in an operational context, namely 1) the hydrological model, 2) the hydrometeorological dataset, 3) the objective function used for calibration, and 4) the bias/dispersion in the ensemble streamflow prediction (ESP) system. The value of these elements is assessed by making a full factorial design experiment. The elements are changed in various combinations to generate various ESPs, feeding a test bench which simulates a single hydropower generating reservoir. A linear programming algorithm is then used to optimize water management decisions. The value of the analyzed elements in the forecasting chain is evaluated by comparing the variance in the average profit generated by each of the ESPs, grouped by combination for each element. The impacts of other constraints such as energy purchase price and minimum load constraints are also evaluated. For the studied system, results show that the elements taken independently have little impact on the average profit variance, while higher-order interactions between the elements lead to a larger impact on profitability. However, bias/dispersion and its interactions with other elements show no significant impact on the profit variance under the operational conditions in this study. Results show that multiple elements need to be simultaneously improved to achieve this goal.

1. Introduction

Hydropower is a major source of energy worldwide and its contribution to the global energy portfolio could grow in the coming years as energy production is transitioning towards renewable sources. Hydrological forecasting is relied upon to manage water resources although it is inherently uncertain (Georgakakos et al., 1998). To address forecast uncertainty, hydrological ensemble forecasting systems and approaches have been developed with the aim of increasing confidence in the hydrological forecasts, specifically for their operational use in water management (Troin et al., 2021). One of the commonly used forecasting techniques is the Ensemble Streamflow Prediction (ESP) system, which is a stochastic approach in which each year of the historical meteorological dataset is used as input into a hydrological model over the watershed of interest, providing the user with a range of probable streamflows.

In the operational context of hydropower reservoir management, the ESPs are then used as inputs to an optimization algorithm which optimizes the profit from energy production based on the probable inflows. There are various types of optimization algorithms, and Cassagnole et al. (2021) identify three main classes: non-linear and linear programming (LP), dynamic programming and stochastic variants, and heuristic programming. The best optimization algorithm for a given task depends on many factors, such as the computing power and available time to converge to a solution, the nature and size of the problem, and the type and quality of available information. In a hydropower reservoir management context, the optimization is based on various constraints such as reservoir levels, expected long-term value of the water stock, and minimum load constraints (MLC). The MLC is the minimal amount of energy that the system must provide at each time step to meet its obligations, either through generation or purchase of energy.

Assessing forecast quality involves examining three aspects (or attributes) of the forecasts, as per Murphy (1993): consistency (the correspondence between the forecasters' judgments and their forecasts), accuracy (the correspondence between the forecasts and the matching observations), and value (the incremental economic and/or other benefits realized by decision makers through the use of the forecasts). The quality of a hydrological forecast is affected by the different elements

* Corresponding author.
E-mail address: davidson-chaput.1@ens.etstml.ca (J. Davidson-Chaput).

https://doi.org/10.1016/j.jhydrol.2025.133186
Received 9 April 2022; Received in revised form 14 December 2024; Accepted 22 March 2025
Available online 25 March 2025

composing the forecasting chain, such as the hydrological model structure (Butts et al., 2004), the hydrometeorological input dataset (Guo et al., 2018; Schreiner-McGraw and Ajami, 2020), the objective function for model calibration (Meng-Xuan et al., 2016), and the bias/dispersion in the forecasting ensembles (Zalachori et al., 2012). No study has demonstrated that, given an adequate hydrological model and calibration, a better model calibration score leads to more valuable hydrological forecasts, even if this practice is widely accepted in the forecasting community.

Only a few studies have analyzed the issue of the value of each element composing the hydrological forecasting chain in an operational context. For instance, the value of post-processing is assessed by Boucher et al. (2011) who compared post-processed ESPs generated from low-resolution meteorological forecasts to a high-resolution deterministic forecast in a case study of a flood that caused management difficulties of the Baskatong Reservoir in Quebec, Canada. They showed that, even with a basic post-processing method, the ESPs perform better than the high-resolution deterministic forecast. Cassagnole et al. (2021) evaluated the influence of bias, a property of generated forecasts, on the value of hydrological forecasting in an operational context by using a LP algorithm. They generated synthetic 7-day ESPs from the observed outflows of ten watersheds in France by adding various degrees of bias before feeding them in a test bench to simulate their corresponding reservoirs and energy generation. When compared to the perfect forecast (i.e., the observed outflow), they found that biases diminish the generated profits by 1 % to 3 %, whereas positive biases (overestimating the outflows) diminished profits the most. Arsenault and Côté (2019) also evaluated the influence of bias on energy production for the Saguenay-Lac-Saint-Jean hydropower system in Quebec, Canada. They generated 120-day ESPs by feeding historical hydrometeorological data to a hydrological model. Introduction of biases was done by multiplying the outflows by a ratio, thus generating an ensemble of systematically biased ESPs. The forecasts were then passed to a test bench to simulate the system and optimize decisions through three LP algorithms. They showed that a 5 % positive bias (overestimating the outflows) led to a more profitable forecasting system.

However, no study to date has evaluated the operational value of the set of the elements composing the forecasting chain, as well as the interactions between elements, making it hard to pinpoint which element would provide the largest improvement in profitability of the hydropower system.

This study aims at assessing which element of the forecasting chain has the most impact on the optimization of hydropower production and profit. More specifically, hydrological ensemble forecasts are generated by using an ESP system and the test bench is conducted over the Lac-Saint-Jean watershed in Quebec, Canada. This watershed is exploited for hydropower generation, which highlights some implications regarding the elements of the system that could be improved to increase the profitability of the operational forecasting system.

2. Methods

2.1. Overview

This study uses a test bench that simulates the Lac Saint-Jean (LSJ) reservoir, in Quebec, Canada, and its snowmelt-dominated unregulated watershed and inflows. The land use on the watershed is predominantly boreal forest (virgin or logged), with some agriculture and urban areas around the reservoir. Fig. 1 presents the watershed and reservoir, while Tables 1 and 2 depict the properties of the watershed and the reservoir, respectively. The complex, cascading hydrological system of the Lac Saint-Jean is simplified to reduce the amount of computational power required for the study. As such, only a single, unregulated watershed and the most downstream reservoir is modeled. This was done for two reasons. First, the actual operational software is extremely precise and detailed and takes multiple hours to run each day on a high-performance

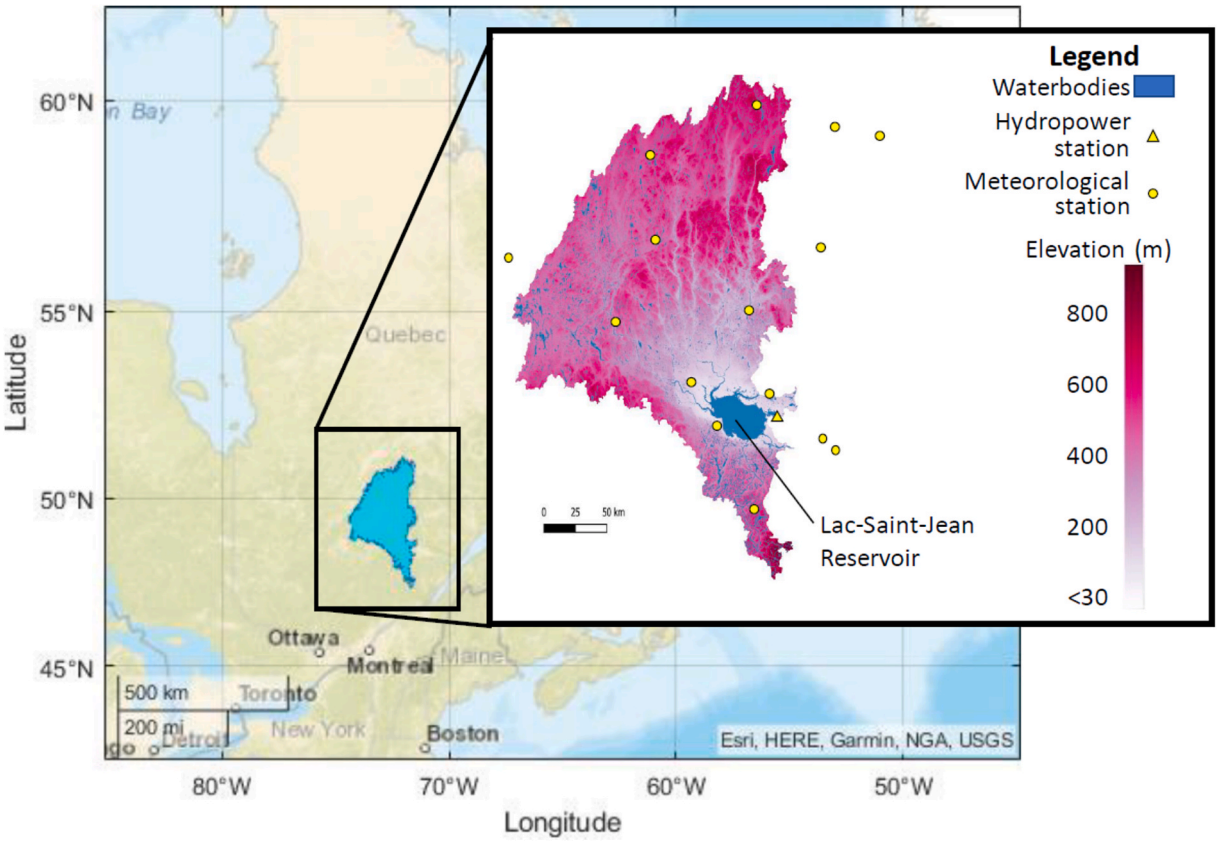


Fig. 1. The Lac-Saint-Jean (LSJ) watershed used in the test bench.

Table 1
The Lac-Saint-Jean watershed properties.

Watershed name	Lac-Saint-Jean (LSJ)
Drainage area [km^2]	45,362
Centroid coordinates (Latitude/Longitude) [°]	49.504°N, 72.696°W
Average streamflow [m^3/s]	865
Average maximum annual streamflow [m^3/s]	4010
Average elevation [m above average sea level]	200
Average total annual precipitation [mm]	930
Average total annual snowfall (water equivalent) [mm]	240
Average annual temperature [°C]	1.6

Table 2
Reservoir properties.

Reservoir name	Lac-Saint-Jean (LSJ)
Surface area [km^2]	1041
Storage capacity [hm^3]	4550
Water level range [m]	7.9
Maximum flow rate before spillage [m^3/s]	1600
Hydropower station name	Isle-Maligne
Installed capacity [MW]	454

cluster. For research purposes, the problem needs to be simplified to be solvable in a reasonable timeframe. However, impacts are limited to linearization of splines and other curves, and as such the underlying mechanics are decently well approximated. Also, limiting the number of reservoirs to one ensures that the results can be more easily interpreted.

The test bench, provided by the project's industrial partner Rio Tinto, manages the virtual reservoir by operating a hydropower station and its spillway. The management rules are as close as possible to the ones used to manage the real-world reservoir, and the LP algorithm optimizes the production and spills according to ESPs. The decisions are based on horizons ranging from short-term (next time steps; three days in the current study to save on computing power) to long-term (15 months ahead in the current study). As in real-world operations, the test bench balances energy production, optimal operating head, and spilled water from floods (either planned or not). To manage the reservoir, the test bench uses ESPs to evaluate the probabilities of inflows into the reservoir and generates power or spills water accordingly. Since Rio Tinto's aluminum smelters are powered by the hydropower station, it is constrained by a MLC which is considered as the default setting of the test bench; the algorithm evaluates if it is more profitable to generate power or buy some on the markets to compensate insufficient production. There are four inputs to the test bench: (1) the ESPs, (2) the weight of each member of the ESP, which are all considered equiprobable in this study, (3) the observed inflows into the reservoir, and (4) the length of each time step. The test bench itself has nine hyperparameters: (1) selling price of excess energy, (2) buyout price of energy (purchase price), (3) cost of not meeting the MLC (penalty), (4) the MLC power to maintain (in MW), (5) penalties for operating above the maximum operating level of the reservoir (MOL), (6) penalties for buying more energy than agreed by contract, (7) maximum reservoir outflow (in m^3/s), 8) maximum quantity of energy available for buyout at each time step (in MW), and (9) initial (and maximum) capacity of the reservoir. In this study, changes will be made to the hyperparameters number (2), (4) and (9) to assess their potential impact on the system's behaviour. To protect strategic business information from the industrial partner, all prices and penalties are a multiplier of the unitary sales prices and are therefore unitless.

2.2. Hydrological forecasting chain variations

Fig. 2 shows the considered elements of the hydrological forecasting chain used to create the total of 225 sets of ESPs. The considered elements are the following: the hydrometeorological input dataset (Fig. 2, box a) the hydrological model (box b), the objective function for model calibration (box c) and the bias/dispersion in the ESP (box d). Changes are made to each of these elements in a discrete and combinatorially complete manner to generate new ESPs, which are then provided as input into the test bench. Each element of the hydrological forecasting chain as well as the process to generate the ESPs are described hereafter.

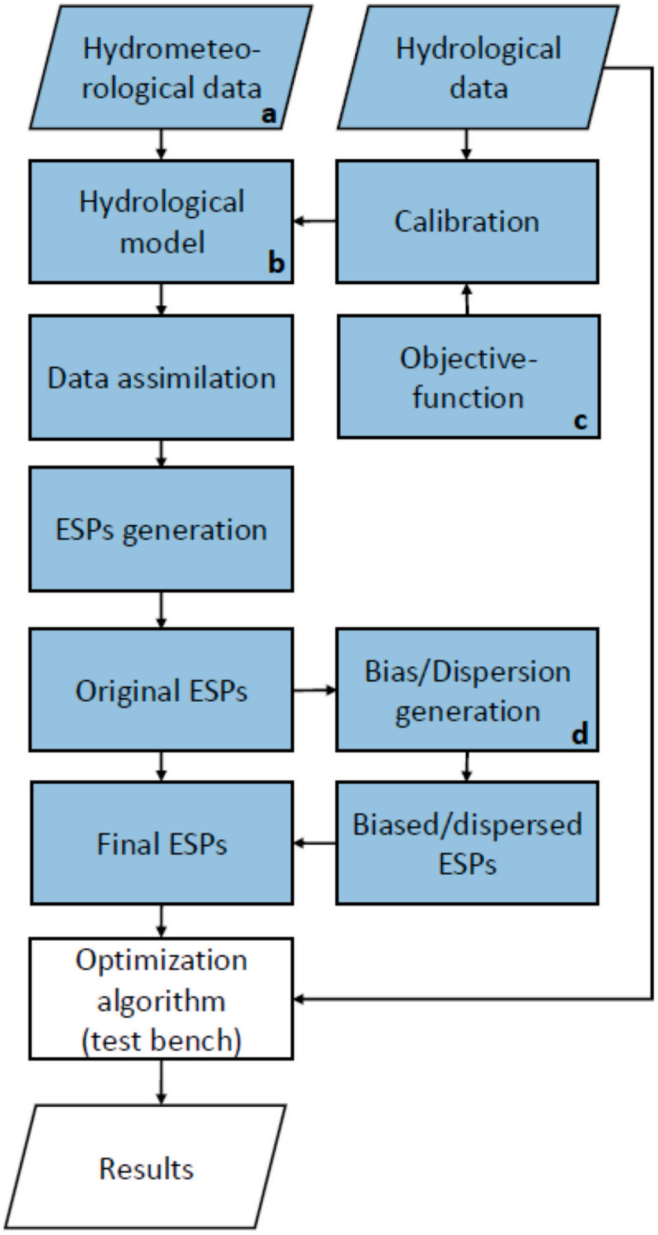


Fig. 2. Hydropower modelling flowchart. The letter in the bottom-right corner of each box refers to each of the four elements that are modified in this study as defined above. Boxes in colored background represent the elements that correspond to the generation of the ESP forecasts (i.e. the hydrological forecasting chain).

2.2.1. Hydrological models

Three lumped hydrological models are used to generate inflows to the reservoir and are part of the Hydrological Prediction Laboratory (HOOPLA) framework (Thiboult et al., 2019). This framework regroups different tools for hydrological modelling and forecasting such as models, calibration algorithms, data assimilation modules and ESP generation capabilities and has been used in previous studies on this catchment (e.g. Dion et al. 2021). Each hydrological model in the

HOOPLA framework was adapted in such a way that they only use minimum, maximum and mean daily temperature, and total precipitation as inputs. They can be run at the daily or 3-hour time step, but the daily time step was used in this study to accommodate the available meteorological input temporal resolution. In this framework, snow accounting (SA) and potential evapotranspiration (PET) routines are processed outside of the hydrological model, allowing the use of the same routine for all hydrological models. This allows to control for confounding factors by standardizing the processes. The SA and PET routines are not changed to reduce computation time (each change doubling the number of ESPs generated). The SA and PET routines used are CemaNeige (Valéry et al., 2014) and Oudin (Oudin et al., 2005), respectively. This selection is supported by the fact that these two routines provide good objective function values for large-sample studies of uncertainty (Troin et al., 2022).

The three hydrological models used in the present study are adapted from (a) CEQUEAU (Girard et al., 1972), (b) GR4H (Mathevet, 2005), and (c) TOPMODEL (Beven et al., 1984) as examples of models with varying degrees of complexity. To ensure all models can run as lumped models at the daily timestep and use the same input data, CEQUEAU and GR4H were slightly modified from their original versions (with GR4H becoming essentially the daily variant GR4J, with some small differences in parameterization; Perrin et al. 2003), while TOPMODEL was substantially modified (Thiboult et al., 2019). Changes that were made in HOOPLA pertain to three aspects: models were changed from distributed to lumped when applicable, some parameters with low sensitivities were set to fixed values, and models share the same potential evapotranspiration and snow module to preserve the same inputs to all models. The versions of the models used herein can be summarized as follows. Note that parameters exclude those in the CEMANEIGE snow module:

- CEQUEAU has 9 calibration parameters and was transformed from its original distributed form to a lumped model such that the routing is performed through a lag function with calibrated parameters for shape and duration. It has two internal states (surface reservoir and ground reservoir) and simulates percolation as well as flow partitioning between the two water storage states.
- GR4H has 4 calibration parameters and uses calibrated triangular unit hydrographs for runoff routing. It has two internal states (production store and routing store). There is a soil moisture accounting routine which controls percolation rates. Water is then partitioned into the water storage reservoirs and flows are convoluted through the unit hydrographs to obtain the final outflow.
- TOPMODEL has 7 calibration parameters and three water storage states (Interception reservoir, ground reservoir and quadratic routing reservoir). It models interception and percolation into the various water stores, before routing runoff from the lower two stores using a similar calibrated lag function to that of CEQUEAU.

2.2.2. Hydrometeorological data

The meteorological input data required for the three hydrological models are the minimum (tasmin), mean (tas) and maximum (tasmax) daily air temperature, and the total daily precipitation (pr). The data come from three different sources: meteorological stations provided by the industrial partner Rio Tinto, which are averaged at the catchment scale using Thiessen polygons, the NRCan gridded and interpolated observational database (Hutchinson et al., 2009; Hopkinson et al., 2011; McKenney et al., 2011) and the ERA5 reanalysis database from the Copernicus Data Store (Hersbach et al. 2018), which was shown to perform well for hydrological modelling in North America (Tarek et al. 2020). The uncertainty associated to the three datasets introduces variability as complexity varies between products: observational stations include reading errors, missing data, and the averaging is done in a specific manner, gridded observational data interpolate data from stations using complex algorithms and terrain data, and reanalysis data are simulations of the best estimate of the atmosphere state but do not rely on ground-based instruments for many variables, including precipitation. One dataset of each type of product is retained to limit the required computing time. Other data sources (such as the Ensemble Meteorological Dataset for North America (Tang et al. 2021) could also be considered, although due to the combinatorial nature of this study, it was preferred to keep only these "raw" data sources rather than combined products.

The timeseries of observed inflows to the reservoir are obtained by mass balance and provided by the industrial partner. The data are used for the calibration of the three hydrological models and provided to the test bench as the observed reservoir inflows to update its internal states. All the data are available over a 31-year period (from 1980 to 01-01 to 2010–12-31) which is used as the study period.

2.2.3. Objective functions used in model calibration

Meng-Xuan et al. (2016) show that no single objective function performs best in all conditions. The objective function must be chosen according to the phenomenon to be modelled, such as spring floods or summer low flows, which can affect the overall performance of the hydrological model and resulting forecasts. To ensure that the entirety of the hydrograph components is considered during calibration for this study, five objective functions were used to calibrate the hydrological models. They are: (1) the Nash-Sutcliffe Efficiency (NSE; Nash and Sutcliffe, 1970), (2) the Kling-Gupta Efficiency (KGE; Gupta et al., 2009), (3) the root mean squared error (RMSE), and both the NSE and the RMSE applied to the natural-logarithmic transformation (ln) of the observed and simulated outflows, (4) lnNSE, and (5) lnRMSE. As explained by Gupta et al. (2009), the widely used NSE (and RMSE to some extent) objective function tends to poorly model the low flow periods in watersheds with a large range of flows. The KGE addresses this issue by balancing the variance, bias, and correlation aspects into the calculations to make it more versatile. By applying a logarithmic transformation to the outflows, the range between low and high flows is reduced. The lnNSE and lnRMSE thus tend to have better results at modelling low flow periods (Santos et al., 2018). NSE, RMSE as well as their log-transformed variants are all monotonic functions of MSE and should theoretically all return the same optimal parameter set (i.e. the parameter set that optimizes NSE should also optimize the others). However, in practice, the transformations change how the calibration algorithms establish gradients and converge, leading to a wider variety of "optimal" parameter sets depending on the local minimum attained (Gupta and Kling, 2011). Calibration was performed using all data from 1980 to 2010 inclusively at the daily time step, as recommended in Arsenault et al. (2018) and Shen et al. (2022). The hydrological models were all calibrated using a budget of 10,000 evaluations using the CMAES optimization algorithm (Hansen and Ostermeier 2001), as recommended in Arsenault et al. (2014).

2.2.4. Data assimilation

Data assimilation consists in updating the state variables of the hydrological model at each time step, providing initial conditions consistent with the last generated output and the current inputs, thus improving hydrological forecasts. In an operational setting, these updates are either done manually or using algorithms (Mai et al., 2020). A data assimilation scheme is implemented to keep the models from going astray and to be more representative of operational conditions. This allows the generation of ESPs with lower initial error between the observed and simulated flows at the time that the forecast is issued. While this masks the imperfections of hydrological models during simulation, it is a standard practice in operational settings to ensure forecasts begin at reasonable initial states. Ensuing forecasts are still therefore affected by model structure and parameterization.

The Ensemble Kalman Filter (EnKF) scheme (Evensen, 2003) is used as a data assimilation method in this study as it was shown to perform well on this catchment and with the hydrological models (Dion et al.

2021, Sabzipour et al. 2023). EnKF generates ensembles of probable state variables considering the uncertainty inherent to the observed streamflow and hydrometeorological inputs of the model. The set of probable initial conditions generated for each time step by the data assimilation algorithm are averaged to generate the most realistic single initial state used to initialize the model's internal state variables. The best estimate of the actual initial state is therefore obtained without adding a layer of complexity linked to the initial state uncertainty. This averaging approach is the same as successfully implemented in Dion et al. (2021).

2.2.5. ESP generation

ESPs are generated for each day of the study period with a horizon of 15 months, covering the period 1980–01-01 to 2010–12-31. The forecasting horizon is chosen to deal with the terminal water value issue (Côté and Leconte, 2016), which forces the optimization algorithm to maximize the profit over a long period rather than emptying the reservoir to generate a larger profit at the first occasion if left unchecked.

To reduce the computational time required to run the test bench, the daily timestep of both the ESPs and the observed outflows are aggregated at coarser temporal resolutions. The ESPs are aggregated in two steps. First, the first three months are aggregated at the 3-day scale to serve as the short- to medium-term forecasts. Then, the remaining twelve months are aggregated at the monthly scale to provide the optimization algorithm with a long-term trend of water availability. Observed inflows are aggregated at the 3-day scale only. The selection of the periods is based on the fact that the hydrometeorological information in the short- and medium-terms have more impact in decision-making compared to the long-term information. For instance, considering that the hydropower reservoir is filled during the spring flood and before winter, all subsequent information will have a low impact on decision-making. This information is added uniquely to deal with the issue of terminal water value and to stop the optimizer from aiming to empty the reservoir at the end of the forecasting horizon.

The hydrological models and data assimilation scheme are first run over the entire study period to generate state variables for each day. Then, to generate the ESPs, 15-month simulations using each of the assimilated days are conducted for each of the three hydrometeorological datasets. Even though the state variables are processed with the data assimilation scheme, there is still a spin-up time to ensure that the data assimilation method converges. Thus, the first year of data is considered as a spin-up period which does not contribute to the ESPs generation or profit evaluation.

When working in hindcast with ESPs generated from historical hydrometeorological data, the hydrometeorological data for the current year cannot be considered as a member, because it would represent a forecast for which the actual weather to come is perfectly known (i.e. not a forecast but a simulation). The last 2 years of available data are also not available as forecast members since the 15-month forecast period would extend beyond the range of available data. Thus, each ESP has a total of 28 possible members (corresponding to the remaining years, i.e. 1981 to 2008 inclusively).

2.2.6. Bias and dispersion

Bias and dispersion are applied to the ESPs by removing members from the original ESP ensemble to create new ensembles as follows:

(1) a negative bias (dry bias with an underestimation of outflows) by removing the four members with the largest total water volume.

(2) a positive bias (wet bias with an overestimation of outflows) by removing the four members with the smallest total volume.

(3) an overdispersion by removing the four members closer to the median total volume (OD bias; two members on both sides).

(4) an underdispersion by removing the four members furthest from the median volume (UD; two members on both sides).

(5) the original ESPs that are not modified in any way from the original set.

The choice of removing four members is made to keep a balance between creating a bias and dispersion and not losing information compared to the original ESPs. Each new ESP is composed of 24 members, while the unbiased ESP is composed of 28 members. This step is performed independently for each forecast issue date, meaning that selected members differ from one forecast to the next.

2.3. Test bench

The test bench used in this study simulates a hydropower station running from a single reservoir. It runs a simplified one-reservoir version of the LP optimization algorithm, which is used by the industrial partner for the Lac-Saint-Jean reservoir management. The test bench simulates a real system by utilizing ensemble forecasts as inputs for each day, then determining the best decision (i.e. water drawdown, spilling and storage) to make to maximize profit (or minimize cost). The decision is made, and then the real observations are used to compute the real power generation and reservoir states at the end of each day. Therefore, the model never has access to the upcoming inflows when making a decision, but energy generation, profits, costs and reservoir states are updated using the observed inflows for each day, simulating a real-world scenario. The test bench uses inflows and their probabilities of occurrence (based on the ensemble distribution) to manage risk using a linear programming approach. The equation that describes the variable to optimize, in this case the profit, is detailed in Eq. (1):

$$P^i = \frac{\sum_{t=1}^{n} \left(G_t^i \bullet SP - \left(B_t^i \bullet BC + O_t^i \bullet OP + BP \right) \right)}{n} \tag{1}$$

where P is the average profit generated; G is the energy generated; B is the energy purchased; SP is the selling price (which has a unitary value of 1); BC is the buyout cost of energy; O is the number of overtopping events; OP is the overtopping penalty; BP is the buying penalty when buying more energy than available by contract; the subscripts t and i are the time step and the combination index, respectively; and n is the total number of time steps. All energy production, sales and buyouts are in MW, the buyout cost (BC) is a unitless multiplier of the sale price per MW, while the penalties are a unitless multiplier of the sale price. Recall that to protect strategic information for the industrial partner, prices and penalties are a multiplier of the unitary sales price and are therefore unitless.

The testbed optimizes profits, which are easy to calculate in terms of energy buyouts and sales, as well as contract violation costs. However, in some cases, there is no direct monetary value to attribute to constraint violations, such as when the reservoir levels exceed the maximum limit (i.e. flooding or dam safety risk). In these cases, penalties are applied using a factor that represents the risk profile that is acceptable to the water resources managers. For example, in the case of maximum reservoir storage violation (overtopping penalty; OP), a very high penalty value (e.g. 10 000 units of the energy sale price) would exert pressure on the optimizer to minimize this occurrence as it would be costly if the risk materializes. Oppositely, if the penalty is low (e.g. 1 unit of the energy sale price), the optimizer might be more aggressive and preserve higher levels to maximize energy generation due to the low cost of violations. In this study, the overtopping penalty was set to 100 units of energy sale price, which was found to be a good compromise in-line with expected overtopping rates.

Furthermore, MLC constraints are imposed to ensure sufficient energy is provided to the smelters. If the generated energy is sufficient for a given day, there is no penalty, and any excess energy can be sold at the energy sale price. Oppositely, if the energy generation is insufficient, the difference must be purchased on the market at a fixed price (buyout cost; BC). These costs are negotiated by contract up to a certain limit. Any excess energy must be purchased at a much higher cost which is penalized using the buying penalty (BP). To ensure that the model does not rely on energy purchases too much, the BC is adjusted such that it is

higher than the energy sale price. In this study, multiple BCs are tested to assess the model's adaptability to progressively more constraining costs. In essence, the higher the buyout cost, the less the model will attempt to generate excess energy to sell as it will prefer preserving any excess water to turbine on days with lower inflows, thus minimizing energy purchases.

All of these constraints were programmed and provided to the optimizer, which was optimized using the CPLEX linear programming solver (CPLEX User Manual, 1987). The testbed then provides values of energy generation, sales and purchases, constraint violations, profits, and other details to assess how the model would have fared on the historical period given the various ensemble inflow forecasts provided as inputs.

As mentioned previously, to further assess the impacts on profits of changes in the hydrological forecasting chain, the following test bench parameters are modified: (1) activating or deactivating the MLC, which imposes a minimal production or buyout of energy for each time step on the optimization algorithm, thus reducing the decision possibilities due to extra constraints, (2) raising the cost of energy buyout to various values (from 1.01 to 1.4 times the energy purchase price), and (3) the volume of the reservoir (+50 % and −50 %), both with and without MLC.

Each ESP generated by the various combinations of the forecasting chain's elements (see Fig. 2) is provided as an input to the test bench. A total of 225 sets of ESPs are generated over a 15-month forecasting horizon (3 hydrometeorological datasets X 3 hydrological models X 5 objective functions X 5 bias/dispersion variants = 225 ESPs).

The observed inflows to the reservoir is considered as an input to the test bench. This allows the test bench to update both the water volume of the reservoir and the hydraulic head at which the hydropower station operates, while computing the electricity generation for any given day.

2.4. Analyzed metrics

Various metrics are used to assess the impacts of the changes in the hydrological forecasting chain on the profit generated by the hydropower station. The main metrics are the energy generation as well as the profit, which can be evaluated and compared directly.

To assess the impacts of the various components of the forecasting chain, the sum of squares (SS) of the average profit of the various combinations P^i (of equation (1), presented as a percentage of the total SS (the relative SS), was used as the main metric. An analysis of variance (ANOVA) was performed to evaluate the variance contribution of each element to the simulated profits in the test bench. An ANOVA allows to analyze the impact of various elements on the variations in results. The following terms are defined here to ensure clarity in the results discussion:

- Factor: elements of the hydrological forecasting chain being changed.
- Source: the factor or combination of factors being analyzed for its contribution to the variance.
- Order: the order of the source represents the number of factors contained in the source.

For instance, the combination of the hydrological model and hydrometeorological dataset is a second order source which combines two factors, and the hydrological model factor is composed of three elements (CEQUEAU, GR4H, and TOPMODEL).

To evaluate the overall contribution of a factor to the variability in the generated profits, the relative SS of the profit from each source containing the factor is summed up. The ranked-based nonparametric Kruskall-Wallis test is performed (at a significance level of $\alpha = 0.05$) to determine if the average profit attributed to a particular element within a factor (i.e., a particular model against others) is statistically different from the others.

3. Results and discussion

3.1. Model calibration results

The hydrological model calibration results are presented in Table 3, for each of the model/dataset combinations and for each objective function.

It can be seen that the models are generally well calibrated, following guidance of Moriasi et al. (2017), whereby the NSE values are found to be good above values of 0.65 and very good above 0.75. Other metrics are also satisfactory when compared to operational results and to results of previous studies. The NRCan dataset displayed the worst calibration score in almost all cases and ERA5 provided the best results for most cases as well. This shows that the importance of selecting high-quality datasets for model calibration and simulation. However, it remains to be seen how this impacts the hydropower generation in an operational forecasting context, as will be seen below.

3.2. Average profit by source of variance

Average profit by factor is presented in Fig. 3. The boxplots for the bias/dispersion (X4) show marginal variability between the five elements when comparing the size of the interquartile range. Slightly more impactful is the hydrological model (X1), however, the choice of dataset and objective function (X2 and X3 respectively) show the most variability. This seems to correlate with the calibration results shown in Table 3. One particular element is that of the NRCan dataset. While it showed generally lower performance in calibration in table 3, it also seems to provide more profits on average than the station dataset. It is possible that the RMSE metric (for which NRCan is the best dataset) allowed generating more profits for this dataset, which would also explain the larger spread than the other models. Indeed, NRCAN is the best dataset for some metrics and the worst for other metrics. Ultimately, the combination of dataset and metric seems to affect the average profits in this case.

A Kruskal-Wallis statistical test was then performed to evaluate the contribution of each element of the forecasting chain to the average profit. Results are shown in Fig. 4. For the elements composing the hydrological model factor (Fig. 4a), average profits are not significantly different from each other, although it can be seen that the top and bottom quartiles of the GR4J average profits are slightly higher than those of the other two models.. The ERA5 reanalysis dataset generates a slightly higher median average profit with a smaller spread than the other two datasets (Fig. 3). The ERA5 dataset is also significantly different than the others according to Fig. 4b, suggesting that more profit can be generated if the optimization algorithm is fed the ESPs generated from the ERA5 dataset. Most results are not significantly

Table 3
Calibration scores for all model-metric-dataset combinations. Bold values indicate the best value in each row.

		CEQUEAU	GR4H	TOPMODEL
NSE	Stn	0.7091	**0.7230**	0.6916
	NRCan	**0.6861**	0.5252	0.5999
	ERA5	**0.7649**	0.7325	0.7392
KGE	Stn	0.7897	**0.8502**	0.8002
	NRCan	**0.8039**	0.7631	0.7747
	ERA5	0.8509	**0.8630**	0.8531
RMSE (mm)	Stn	0.8327	**0.7982**	0.8423
	NRCan	**0.8497**	1.0451	0.9593
	ERA5	**0.7304**	0.7845	0.7746
lnNSE	Stn	**0.8243**	0.8066	0.7678
	NRCan	**0.7368**	0.6998	0.7171
	ERA5	**0.8152**	0.7945	0.7694
lnRMSE	Stn	**0.3308**	0.3519	0.3855
	NRCan	**0.4049**	0.4384	0.4255
	ERA5	**0.3429**	0.3627	0.4030

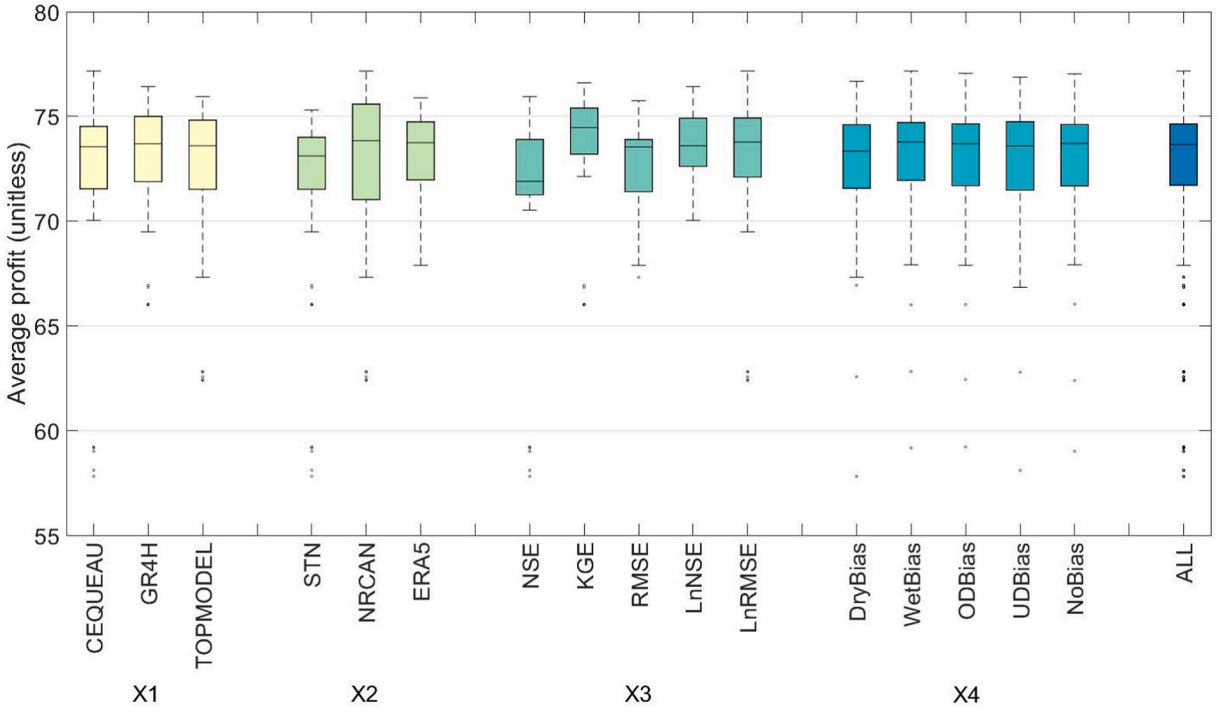


Fig. 3. Comparison of average profit when maintaining one element constant in each factor: the hydrological model (first from left, X1), the hydrometeorological dataset (second from left, X2), the objective function (middle, X3), and the bias/dispersion (second from right, X4). The results of the entire 225-member ensemble are presented in the rightmost boxplot (ALL). For example, the first model boxplot, CEQUEAU, shows the average profit of the 75 combinations (1 hydrological model x 3 hydrometeorological datasets x 5 objective functions x 5 bias/dispersion levels).

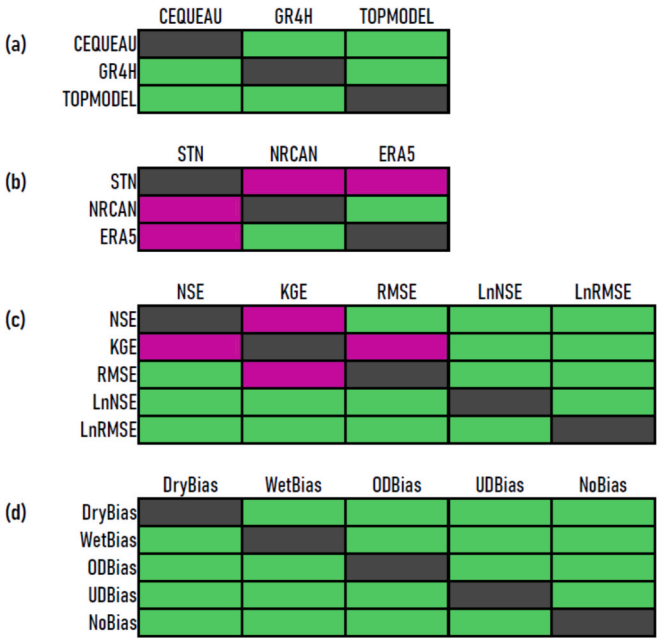


Fig. 4. Results of the Kruskal-Wallis tests comparing the average profit generated by the elements composing the factors: (a) the hydrological model, (b) the hydrometeorological dataset, (c) the objective function, and (d) the bias/dispersion factor. Failure to reject the the null hypothesis indicating that there is no significant difference (at a 5% level) is shown in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

different from the others regarding the elements composing the objective function factor (Fig. 4c), except for the KGE, which is also the one displaying the higher profits. The KGE was devised to solve inherent problems in the interactions between elements composing the NSE (Gupta et al., 2009), which might explain why it stands out from the other objective functions. It is also the only one that is not a transformation of the Mean Square Error (MSE). Then, the elements composing the bias/dispersion factor are not significantly different between themselves (Fig. 4e), which supports the results of Figs. 3 and 5 (below), and will be discussed in more detail in section 3.6. With respect to the model calibration score in Table 3, it is expected that the higher the score, the more generated profits. However, it was previously demonstrated that the calibration score is not necessarily a good indicator of the model performance in simulation mode, as shown in Arsenault et al. (2018) and Shen et al. (2022), and few (if any) studies have looked at model calibration score impacts on forecasting skill. Furthermore, the fact that data assimilation is performed before each forecast means that cumulative model errors (in simulation) are corrected before each forecast, and only the forecasts themselves are affected by the calibration score. This thus changes the dynamics of the forecasting process, and poorer models in calibration can still outperform others when used for forecasting, although it can be seen that the top and bottom quartiles of the GR4J average profits are slightly higher than those of the other two models. For example, a model that has a lower calibration score overall but has a very accurate rainfall-runoff component during summer could still outperform a generally better model in these conditions. Furthermore, a model that is less constrained could accumulate more error over time (and thus perform worse in calibration), but once the initial states are corrected, the model could still generate better forecasts than other models. In the current study, NRCan-calibrated models outperform those calibrated with station data even though they were the worst in calibration/simulation mode. Likewise, CEQUEAU was the model with the best calibration performance in most combinations, but average profits are essentially identical to those generated from the GR4H model.

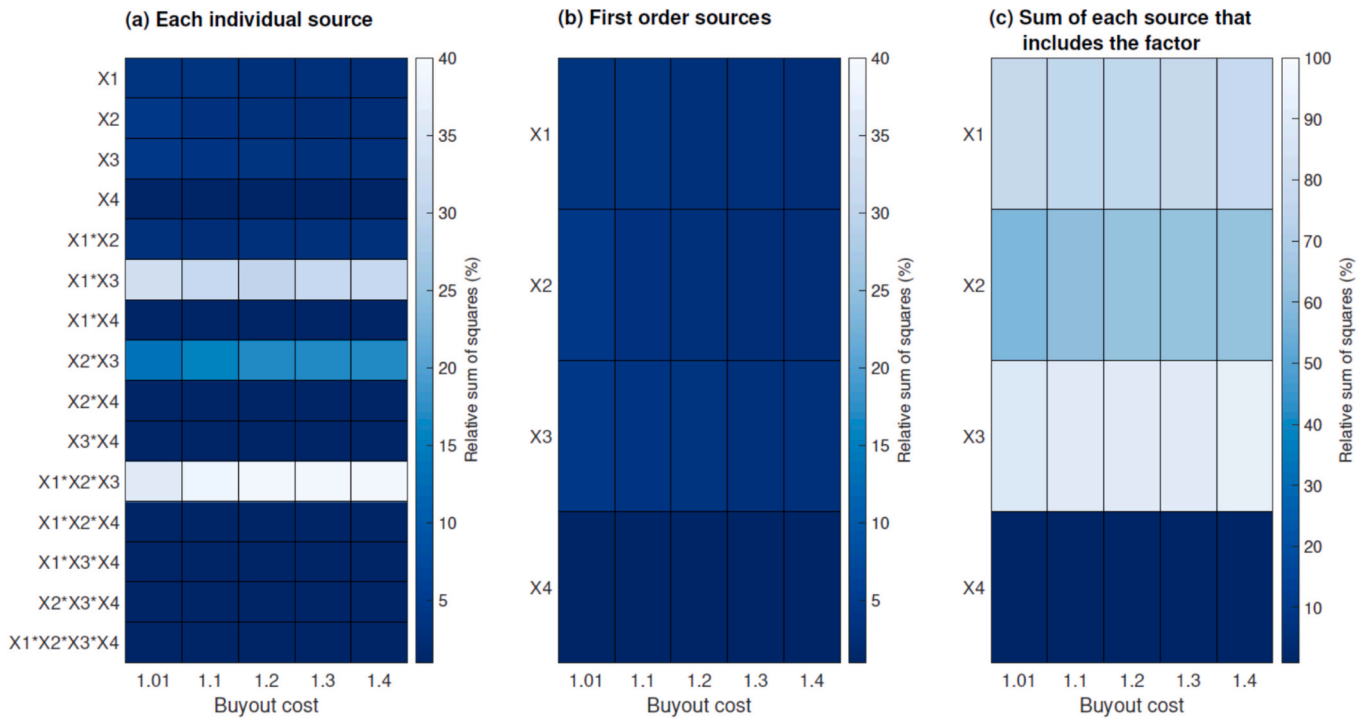


Fig. 5. Relative sum of squares on average profit generated by each source at multiple energy buyout cost for: (a) the individual sources; (b) the first order sources (factors); and (c) for every source in which the factor is included. Each factor is named from X1 to X4: hydrological model (X1), hydrometeorological dataset (X2), objective function for calibration (X3), and bias/dispersion (X4).

3.3. *Variance in the average profit*

The relative variance of the average profit, as defined by equation (1), is presented in Fig. 5 in the form of heatmaps for each source and for different energy buyout costs (BC). Fig. 5a) and 5b) show that, when taken individually, each factor has a relatively small contribution to the total variance in the average generated profit (under 4.5 %). The bias/dispersion is the factor with the smallest impact (maximum of 0.4 %), and any source that has the bias/dispersion involved as an interaction term leads to a negligible impact (under 0.2 %). This is possible because lower-order interactions already explain the variance of their interactions, and adding X4 to them does not help explain more variance on average profits. These results do not corroborate the previous findings of Arsenault and Côté (2019) and Cassagnole et al. (2021); this is further discussed in Section 3.5.

Results indicate that the BC of energy has little influence on the relative SS in the average profit of each individual source (Fig. 5a and 5b). Also, the higher order sources and the summation remain stable as it changes (Fig. 5a and 5c). The first order sources with the most change in variance as BC changes are the hydrolometeorological dataset (X2) and the objective function (X3); they vary from 2.6 % to 4.5 % and from 3.2 % to 4.4 %, respectively. Changes of similar amplitudes are observed for some higher order sources. The third order source which excludes bias/dispersion varies from 36.3 % to 39.6 % at most. As the BC goes from 1.01 to 1.40, the variance of average generated profit remains relatively steady (the changes are below 3.4 %). This shows that the findings of this study could be generalized to the fluctuating BC of energy.

Fig. 5a) indicates that, except for those combinations containing X4 (bias/dispersion), a higher order (the more interactions there are among the factors) leads to a larger percentage of relative variance of average profit. On average, first order terms contribute 10 % of the total variance, while second, third and fourth order terms include respectively 50 % and 40 % of the total variance, even though there are 6 2nd order terms and only 4 3rd order terms. Overall, interactions between the hydrological model, the hydrometeorological dataset, and the objective function explain a large part (89.4 %) of the total variance in the average profit. Considering that the sources of the first order have little impact (10.2 %), leaving [89.4 % − 10.2 % = 79.2 %] to the interaction terms of X1, X2 and X3, this suggests that these three elements of the hydrological forecasting chain are intricately intertwined and that the improvements of the forecasting chain cannot target a single element.

The previous results can be further investigated by looking at the sum of the relative variance for each factor (i.e., each combination with the same element). As shown in Fig. 5c), when summing up the variance from each source in which the individual factors are involved, bias/dispersion has a marginal influence (0.4 %), while the objective function plays a more important role (90.4 %) than the hydrological model (76.6 %) and hydrometeorological dataset (61.3 %). This supports the findings from Fig. 5a) and 5b). It is to be noted that the sum is larger than 100 % because the interaction terms count multiple times in the various combinations.

3.4. *Impacts of varying the reservoir storage capacity and MLC*

To investigate if other characteristics of the hydropower system have an impact on the value of the elements of the hydrological forecasting chain, the volume of the reservoir was modified by +/- 50 % and results are shown in Fig. 6. It can be seen that the average profit increases with larger reservoir size, which is a trivial and expected result. However, it can also be seen that the relative performance of the different elements in each panel are similar, with the KGE-calibrated models performing better, and with the ERA5 dataset being more profitable. It is also interesting to note that as the reservoir size increases, the differences between boxplots of the same category (e.g. dataset, or objective function) seem to become larger, as if the optimization model has more freedom to maximize performance and leverage better forecasts. It can also be seen that for the 50 % larger reservoir (Fig. 6c), the forecasts with the dry bias seem to generate fewer profits than the others, while it is the opposite for the smaller reservoir (Fig. 6a). The “original” reservoir shows no such impacts. This could explain the lack of importance of the

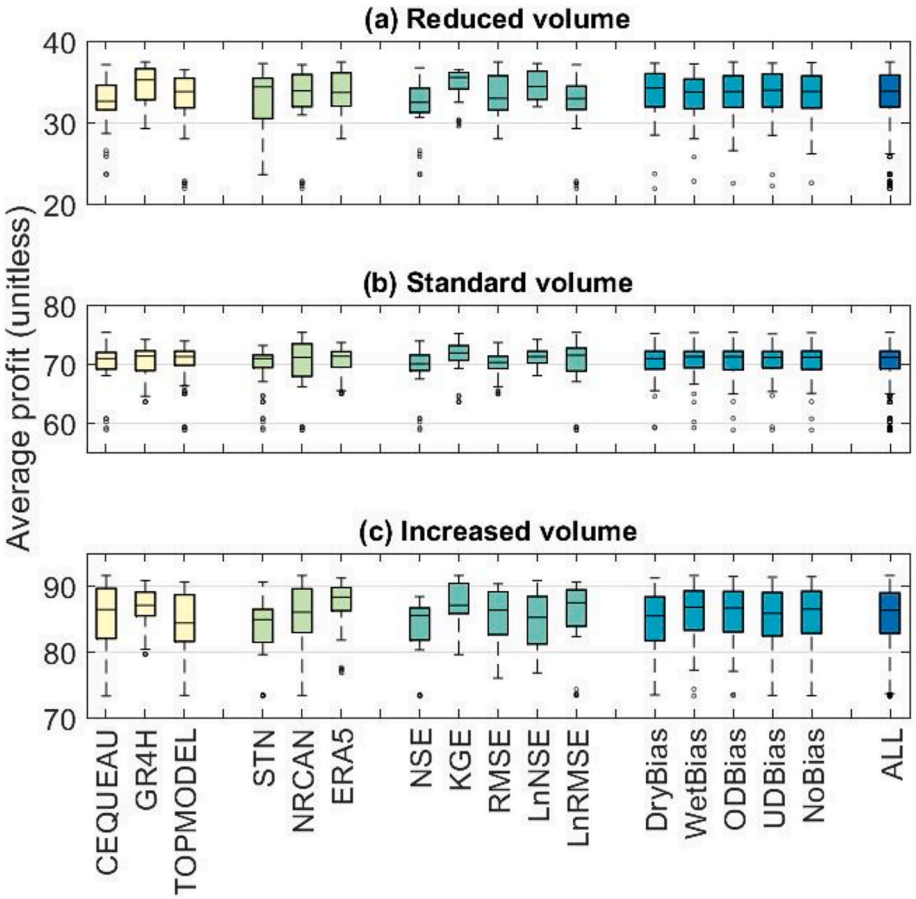


Fig. 6. Relative profit obtained from the various combinations of forecasting chain factors according to reservoir storage capacity. The standard (original) volume is presented in the center (b) panel, whereas the reservoirs with 50% less and 50% more storage are presented in panels (a) and (c) respectively.

bias and dispersion on the variability of results. It is possible that the sizing of the reservoir makes the optimization robust to this metric.

Next, the impact of an MLC on generated profit was assessed by removing that constraint from the test bed. Fig. 7 presents the results for the two cases, with MLC activated (Fig. 7a) and deactivated (Fig. 7b). Results show that, as expected, the deactivation of MLC clearly increases

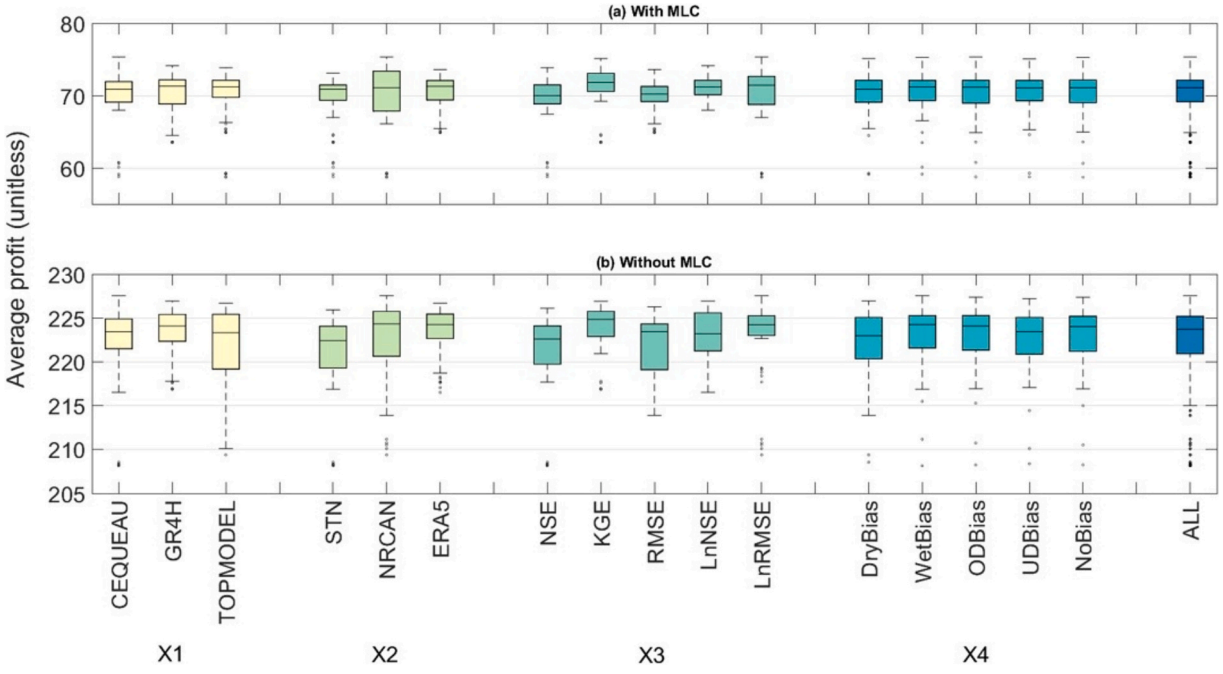


Fig. 7. Average profits from the hydropower system when Minimum Load Constraints (MLC) are activated (panel a) or deactivated (panel b).

average profits as the system can easily optimize for hydropower generation over long timespans instead of sacrificing long-term gains for short-term needs. There is also no penalty cost for missing the MLC constraints. These results must therefore be considered separately, looking at differences in behaviour between the two cases but without comparing them directly. It can be seen that the dry bias is slightly penalized without the MLC. This is expected since the MLC will still generate power given low forecasts to attain its goals, and then be "surprised" with extra water than anticipated during operation. The model without MLC will tubine less when drier forecasts are expected, thus leading to lower generation. It can also be seen that there is a larger spread in results within groups when MLC is deactivated. Again, this can be attributed to the fewer constraints that force the system to turbine given suboptimal circumstances when MLC is active. Without MLC, the optimization algorithm can make better use of the various forecast properties, which is reflected in the larger variance in average profits. This suggests that the degree of freedom of the algorithm is a factor in the value of the elements of the hydrological modelling chain.

3.5. Further investigations into the value of the bias and dispersion factor

The previous results suggest that the bias/dispersion factor has a marginal impact on the average profit with the standard (original) volume. This is not in agreement with Cassagnole et al. (2021) and Arsenault and Côté (2019), although the effects of bias were shown to be affected by reservoir storage capacity. It is plausible that these previous studies saw different results due to different ratios between inflows and reservoir sizes. However, for the catchment in this study, the low variance of the bias/dispersion factor could be explained by the watershed's properties and the optimization algorithm decision-making process. The LSJ is a large and relatively flat watershed characterized by a multi-day concentration time. Because of the lag between precipitation/snowmelt and the observed outflows, the ESPs generated for a watershed with a concentration time longer than the modelling time step tend to be composed of similar inflow members for the first few time steps (most generated ESP members have similar flows for the first two or three time steps). This makes the bias/dispersion irrelevant if the decisions are made on those time steps. To generate biased ESPs, Cassagnole et al. (2021) added noise to the observed inflow data at every time step, thus creating ESPs with diverging members on the first time step. Arsenault and Côté (2019) used the whole LSJ hydropower system, which includes four smaller size watersheds (1297 km2 to 11528 km2). In this case, the effect of bias on energy generation comes from the upstream power stations (not modelled and flows not considered in the present study) which are fed by watersheds with shorter concentration time. Further investigations are required to validate if the lack of spread between the ESP members in the first few time steps is responsible for the low impact of the bias/dispersion factor on variance. One possible pathway to explore is the optimization algorithms' information weighting strategy in a context of reservoir management.

3.6. Results analysis with respect to the test bed conditions

The results in this study display certain characteristics that can be explained by the test bed conditions and could be used to better understand the generalizability of the study results. First, the MLC conditions impose a strong constraint on the system, which must carefully balance energy generation for the current time step and the ability to generate energy at future time steps, knowing that purchasing energy is more costly than the sale of excess energy. This leads the optimization algorithm to generate just enough energy each day and preserving the rest for the future, unless large inflows are forecasted in the near future. This means that even though the forecasts are biased or over/under-dispersed, the overall trend in the forecasts is still dominated by all remaining ensemble members and these need to align with more extreme inflow conditions. The removal of the most extreme members from the ESP ensembles was also performed on a 28-member ensemble, meaning that removing a few members was probably not sufficient to alter the optimization algorithm's decision under these circumstances. It is likely that removing a larger number of members would have modified the results more significantly. However, without MLC, the model is more flexible and has more degrees of freedom to attempt to generate more profits, and is more aggressive (depending on the buyout cost), as described previously.

The fact that relative results (i.e. the order of importance of individual factors) were relatively stable when halving or doubling the reservoir size indicates that the results are likely generalizable to similar systems with similar hydroclimatological conditions, and under the same constraints. For hydropower systems with a single reservoir, as is the case in this study, it would seem that the optimization problem is relatively straightforward given the number of constraints and results would depend heavily on the interactions between the elements generating the hydrological forecasts (i.e. all except the dispersion/bias levels).

3.7. Limitations

The main limitation of the study is the fact that only the hydrological forecasting chain is examined. Since the optimization algorithm can have a large impact on the profit-generating capacity of a hydropower system, future studies should encompass the whole hydropower reservoir management chain including the decision-making algorithm.

Other limitations come from the test bench which is based on a simplified single watershed and a hydropower station system. Having a simplified system makes it difficult to generalize the findings to more complex, interdependent systems where the value of the various elements of the forecasting chain could be different.

In this study, data assimilation is used to prevent the hydrological model's state variables from going astray. This is not representative of real-world operations since typically hydrologists manually update model states using their expertise and knowledge of the watershed flow characteristics. The value of data assimilation into the modelling chain might be evaluated by using different data assimilation schemes.

Model calibration equifinality is a well-known problem in hydrological modelling that is partly addressed in the present work using calibration score degradation. To better assess the value of the calibration score by taking the equifinality principle into account, multiple calibration parameter sets could be generated as well. However, this would have multiplied the required computing power linearly and would have required more than a few variants to fully explore the associated variability.

The computing power required to run distributed models has limited the framework of this study to lumped hydrological models. The variance of distributed models could be investigated if computing power is available. Also, sub-daily time steps are not investigated, which could have an impact on hydropower systems fed by small and highly reactive watersheds. Different SA and PET routines are expected to have similar influence on the variance of generated profit; nevertheless, this can also be investigated in future studies. Finally, the findings of this study relate to a large-size watershed with a multi-day concentration time. The ESP members take multiple time steps to diverge, making the bias/dispersion irrelevant if the algorithm puts too much weight on the first time step when decision-making. Further investigations are necessary to validate if our findings can be generalized to smaller watersheds.

4. Conclusion

This study assessed the value of different elements of the hydrological forecasting chain through an analysis of the variance of the elements in profit generation. To do so, ESPs were generated by modifying four elements of the forecasting chain: the hydrological model, the hydrometeorological dataset, the objective function for model calibration, and

the bias/dispersion. The ESPs were then used as inputs to a test bench simulating the Lac-Saint-Jean Reservoir, a single energy generation system optimizing decisions with a linear programming algorithm.

The variance between the average profit generated by the 225 sets of ESPs from the various combinations was analyzed to identify which potential element allows the overall profit to be maximized. Contrary to the findings of Cassagnole et al. (2021) and Arsenault and Côté (2019), there is little variance in average profit caused by the changes in bias and dispersion, both at the individual level and higher orders of interactions. It is hypothesized that this is due to the long-term climatological forecasts that do not provide sufficient discrimination between members, as compared to short-term forecasts from climate models. It is also shown that the size of the reservoir can affect the impact of biases in forecasts. The variance of every source containing the bias and dispersion is low, including the highest order (four interactions). The bias and dispersion, at least for this climatology-based ESP forecasting system, do not lead to a significant improvement or degradation of the overall value of the hydrological modelling chain.

We show that the average profit generated by the combinations containing ERA5 as the hydrometeorological dataset yield better performance in terms of average profit compared to the other datasets; the combinations containing the KGE objective function also performed the best. This leads to the recommendation that water resources managers should test multiple datasets, models and objective functions for calibration for their operations, as there can be value in identifying the best options even for highly constrained optimization problems.

The modification of the reservoir capacity by ± 50 % gives similar results in terms of relative variance from the diverse sources; the same findings are obtained when comparing results with and without a minimum load constraint. This suggests that the results can be generalized to systems with different reservoir sizes and energy production obligations. However, only one type of optimization algorithm is analyzed and the lack of variance of the bias and dispersion on the average profit partly depends on the watershed size. Therefore, the results cannot be generalized to all optimization algorithms or sizes of watershed.

We recommend further investigations into the various elements composing the hydrological forecasting chain to determine, to which extent, the present results can be generalized. From the findings of this study, most elements of the hydrological forecasting chain need to be improved to increase the overall profits as no single element can achieve this on its own.

CRediT authorship contribution statement

Jonathan Davidson-Chaput: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Richard Arsenault:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Jean-Luc Martel:** Writing – review & editing, Writing – original draft, Validation. **Magali Troin:** Writing – review & editing, Writing – original draft, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Jonathan Davidson-Chaput was supported financially by the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as the Fonds de Recherche du Québec – Nature et Technologies (FRQNT). This project was also partially funded by the NSERC grant number CRDPJ-522126-17. The authors would like to thank Rio Tinto for access to their datasets used in the present study. Finally, the authors would like to extend their gratitude to the anonymous reviewers whose comments and suggestions were instrumental in shaping the manuscript into its current form.

Data availability

Data will be made available on request.

References

Arsenault, R., Poulin, A., Côté, P., Brissette, F., 2014. Comparison of stochastic optimization algorithms in hydrological model calibration. Journal of Hydrologic Engineering 19 (7), 1374–1384.

Arsenault, R., Côté, P., 2019. Analysis of the effects of biases in ensemble streamflow prediction (ESP) forecasts on electricity production in hydropower reservoir management. Hydrol. Earth Syst. Sci., 23, 2735–2750. https://doi.org/10.5194/hess-23-2735-2019.

Arsenault, R., Brissette, F., Martel, J.L., 2018. The hazards of split-sample validation in hydrological model calibration. Journal of Hydrology 566, 346–362.

Beven, K.J., Kirkby, M.J., Schofield, N., Tagg, A.F., 1984. Testing a physically-based flood forecasting model (topmodel) for 3 UK catchments. Journal of Hydrology 69, 119–143. https://doi.org/10.1016/0022-1694(84)90159-8.

Boucher, M.-A., Anctil, F., Perreault, L., Tremblay, D., 2011. A comparison between ensemble and deterministic hydrological forecasts in an operational context. Adv. Geosci., 29, 85–94. https://doi.org/10.5194/adgeo-29-85-2011.

Butts, M.B., Payne, J.T., Kristensen, M., Madsen, H., 2004. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. Journal of Hydrology 298, 242–266. https://doi.org/10.1016/j.jhydrol.2004.03.042.

Cassagnole, M., Ramos, M.-H., Zalachori, I., Thirel, G., Garçon, R., Gailhard, J., Ouillon, T., 2021. Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs – a conceptual approach. Hydrol. Earth Syst. Sci., 25, 1033–1052. https://doi.org/10.5194/hess-25-1033-2021.

Côté, P., Leconte, R., 2016. Comparison of Stochastic Optimization Algorithms for Hydropower Reservoir Operation with Ensemble Streamflow Prediction. Journal of Water Resources Planning and Management 142, 04015046. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000575.

Dion, P., Martel, J.L., Arsenault, R., 2021. Hydrological ensemble forecasting using a multi-model framework. Journal of Hydrology 600, 126537.

CPLEX User Manual, 1987. Ibm ilog cplex optimization studio. Version 12 (1987–2018), 1.

Evensen, G., 2003. The Ensemble Kalman Filter: theoretical formulation and practical implementation. Ocean Dynamics 53, 343–367. https://doi.org/10.1007/s10236-003-0036-9.

Georgakakos, A.P., Yao, H., Mullusky, M.G., Georgakakos, K.P., 1998. Impacts of climate variability on the operational forecast and management of the Upper Des Moines River Basin, Water Resour. Res., 34, 799–821. https://doi.org/10.1029/97WR03135.

Girard, G., Morin, G., Charbonneau., R., 1972. Modèle précipitations-débits à discrétisation spatiale. Cahiers ORSTOM, Série Hydrologie, 9, 35-52.

Guo, B., Zhang, J., Xu, T., Croke, B., Jakeman, A., Song, Y., Yang, Q., Lei, X., Liao, W., 2018. Applicability Assessment and Uncertainty Analysis of Multi-Precipitation Datasets for the Simulation of Hydrologic Models. Water 10, 1611. https://doi.org/10.3390/w10111611.

Gupta, H.V., Kling, H., 2011. On typical range, sensitivity, and normalization of Mean Squared Error and Nash-Sutcliffe Efficiency type metrics. Water Resources Research 47 (10).

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

Hansen, N., Ostermeier, A., 2001. Completely derandomized self-adaptation in evolution strategies. Evolutionary Computation 9 (2), 159–195.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D. & Thépaut, J-N., 2018. ERA5 hourly data on single levels from 1979 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (Accessed on 15 Nov 2020), doi: 10.24381/cds.adbb2d47.

Hopkinson, R.F., et al., 2011. Impact of aligning climatological day on gridding daily maximum-minimum temperature and precipitation over Canada. J. Appl. Meteorol. Climatol. 50, 1654–1665. https://doi.org/10.1175/2011JAMC2684.1.

Hutchinson, M.F., et al., 2009. Development and testing of Canada-Wide Interpolated Spatial Models of Daily Minimum-Maximum Temperature and Precipitation for 1961-2003. J. Appl. Meteorol. Climatol. 48, 725–741. https://doi.org/10.1175/2008JAMC1979.1.

Mai, J., Arsenault, R., Tolson, B.A., Latraverse, M., Demeester, K., 2020. Application of parameter screening to derive optimal initial state adjustments for streamflow forecasting. Water Resources Research 56, e2020WR027960. https://doi.org/10.1029/2020WR027960.

Mathevet, T., 2005. Quels modèles pluie-débit globaux au pas de temps horaire? Développements empiriques et comparaison de modèle sur un large échantillon de bassins versants. École Natl. du Génie Rural des Eaux et des For, Paris. PhD thesis.

McKenney, D.W., et al., 2011. Customized spatial climate models for North America. Bull. Am. Meteorol. Soc. 92, 1612–1622. https://doi.org/10.1175/2011BAMS3132.1.

Meng-Xuan, J., Hua, C., Chong-Yu, X., Qiang, Z., Xin-e, T., 2016. A comparative study of different objective functions to improve the flood forecasting accuracy. Hydrology Research 47 (4), 718–735. https://doi.org/10.2166/nh.2015.078.

Murphy, A.H., 1993. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting, Weather and Forecasting 8, 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through. Part I. A conceptual models discussion of principles. Journal of Hydrology 10, 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F., Loumagne., C., 2005. Which potential evapotranspiration input for a lumped rainfall-runoff model? part 2 - Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. Journal of Hydrology, 303, 290-306. doi: 10.1016/j.jhydrol.2004.08.026.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology* 279 (1–4), 275–289.

Sabzipour, B., Arsenault, R., Troin, M., Martel, J.L., Brissette, F., 2023. Sensitivity analysis of the hyperparameters of an ensemble Kalman filter application on a semi-distributed hydrological model for streamflow forecasting. Journal of Hydrology 626, 130251.

Santos, L., Thirel, G., Perrin, C., 2018. Technical note: Pitfalls in using log-transformed flows within the KGE criterion. Hydrol. Earth Syst. Sci., 22, 4583–4591. https://doi.org/10.5194/hess-22-4583-2018.

Schreiner-McGraw, A.P., Ajami, H., 2020. Impact of uncertainty in precipitation forcing data sets on the hydrologic budget of an integrated hydrologic model in mountainous terrain. Water Resources Research 56, e2020WR027639. https://doi.org/10.1029/2020WR027639.

Shen, H., Tolson, B.A. and Mai, J., 2022. Time to update the split-sample approach in hydrological model calibration. Water Resources Research, 58(3), p. e2021WR031523.

Tang, G., Clark, M.P., Papalexiou, S.M., Newman, A.J., Wood, A.W., Brunet, D., Whitfield, P.H., 2021. EMDNA: an Ensemble Meteorological Dataset for North America. Earth Syst. Sci. Data 13, 3337–3362. https://doi.org/10.5194/essd-13-3337-2021.

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

Thiboult, A., Seiller, G., Poncelet, C., & Anctil, F., 2019. HOOPLA: un outil multifonction pour la modélisation hydrologique, Colloque AQT/RHQ - La télédétection de l'eau, dans tous leurs états, Sherbrooke, Canada. (available at https://github.com/AntoineThiboult/HOOPLA).

Troin, M., Arsenault, R., Wood, A.W., Brissette, F., Martel, J.-L., 2021. Generating ensemble streamflow forecasts : a review of methods and approaches over the past 40 years. Water Resources Research 57, e2020WR028392. https://doi.org/10.1029/2020WR028392.

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, in Press. https://doi.org/10.1016/j.jhydrol.2022.127766.

Valéry, A., V. Andréassian, V., Perrin., C., 2014. As simple as possible but not simpler": What is useful in a temperature-based snow-accounting routine? part 2 - sensitivity analysis of the cemaneige snow accounting routine on 380 catchments. Journal of Hydrology, 517, 1176-1187. doi: 10.1016/j.jhydrol.2014.04.058.

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., Gailhard, J., 2012. Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. Adv. Sci. Res., 8, 135–141. https://doi.org/10.5194/asr-8-135-2012.