EPiC Series in Engineering
Volume 3, 2018, Pages 62–69
HIC 2018. 13th International Conference on Hydroinformatics

Multi-Structure Hydrological Ensemble to Improve Flow Daily Prediction in the Sumapaz River Basin, Colombia

Pedro Arboleda[1], David Zamora[1], Carolina Vega[1], Nicolás Duque[1], Erasmo Rodríguez[1]
[1] Universidad Nacional de Colombia
pfarboledao@unal.edu.co,dazamoraa@unal.edu.co,cvegav@unal.edu.co, nduqueg@unal.edu.co,earodriguezs@unal.edu.co

Abstract

Hydrological ensembles have gained importance for prediction and forecasting in water cycle variables. In spite of this, the relevance of the individual models in the ensemble is not usually established, in terms of the ensemble structure (i.e. their members) and the performance this structure exhibits through different climatic conditions (intrannual variability, for example). This analysis accounts for the uncertainty in the structure of the models and their responses (e.g. outputs), in comparison to the observed data. In this regard, the research here described attempts to determine the incidence of the ensemble members built for each month of the year, in the prediction of daily flows, through the use of the Bayesian Model Averaging (BMA) method. Moreover, using BMA calibrated parameters as inputs, an uncertainty analysis is carried out for the calibration period, and in monthly average terms, obtaining finer uncertainty bounds. This analysis was implemented in the Sumapaz River basin, part of the Magdalena Cauca Macro-Basin (MCMB) in Colombia. Results showed differences in ensemble structures and performance according to its original performance criteria, and better results when using a monthly BMA for the uncertainty analysis.

1 Introduction

Hydrological models are not perfect, and its associated uncertainty is revealed when their results diverge for a single hydrological event. This is due to the model's structure and its parameters, which correspond to a simplified and adjusted representation of the lumped behavior of complex and variable natural processes (Montgomery et al., 2015)(HEPEX, 2004).

A model ensemble permits to use heterogeneity on the structure of the models, on its inputs, and consequently on its responses. This approach has been implemented in different disciplines (for example, meteorology), adding flexibility and performance in the predictive and forecasting capacity of natural events. Nowadays, hydrological forecasting centres around the world are changing from single deterministic to multiple probabilistic forecast, using a wide representation of the uncertainty sources (Thielen-del Pozo et al., 2010). This means an extensive use of hydrological ensembles, which is important for the estimation of a range of possible future states (Brown, Demargne, Seo, & Liu, 2010).

Despite its best performance compared to single models, ensemble outputs from different sources (models) need to be processed, in order to give an ensemble unique response. Usual processing includes a simple mean approach, where each model (or output source) has the same weight into the final ensemble response. This approach causes the ensemble response to be under dispersive, i.e., to have lack of variability or present biased responses. Bayesian Model Averaging (BMA) tries to avoid these problems by statistically post-processing ensemble outputs, obtaining a calibrated and sharp prediction (Hamill, 2001; Raftery, Gneiting, Balabdaoui, & Polakowski, 2005).

Within the framework of the eartH2Observe research project (Jaap Schellekens et al., 2017), a hydrological set of ensembles was constructed using physical, lumped and semi-distributed hydrological models, applied in small catchments in Colombia. The ensembles were evaluated under different structures (quantity and type of members), and its performance was assessed using a deterministic criterion (Kling-Gupta Efficiency - KGE) and a probabilistic metric (Continuous Ranked Probabilistic Score - CRPS). In order to determine the individual model weights, the best ensembles (TBE) were evaluated with the BMA method [6], for all time-series in the calibration period. Afterwards, an additional daily evaluation per month with BMA allowed to determine the weights (influence) of the members to predict daily flows in each month. It means we built ensembles using weights by TBE and then we considered the evolution of the weights of the TBE across year.

2 Methods

2.1 Bayesian Model Averaging

BMA is a post-processing statistically approach used to infer a prediction based on different models, and to assess the inter and within model uncertainty, considering a whole ensemble of K members. Given an ensemble of statistical models, where there is no certainty on which one is the best model, the probability of the forecast y on the basis of training data y^T is given by equation (1), where $p(y)$ is the Probability Density Function (PDF) of the forecast y, $P(Y|M_k)$ is the PDF of the forecast y based on model Mk and $P(M_k|y^T)$ is a-posteriori PDF of model Mk being correct given the training data.

$$p(y) = \sum_{k=1}^{K} P(Y|M_k) * P(M_k|y^T) \quad (1)$$

The a posteriori model probability adds up to 1, and so it could be seen as weights, while the PDF of the forecast, based on the model, is always considered, for computational purposes, as normally distributed, but can be enlarged to other probability functions.

The BMA model for a dynamical ensemble forecasting is given in equation (2), where $p(y|f_1, f_2, \dots, f_k)$ is the average response of the ensemble based on K models, w_k are the models weights and $g(y|f_k)$ are conditioned PDF, associated to deterministic forecast f_k that can be interpreted as the PDF of y given f_k.

$$p(y|f_1, f_2, \dots, f_k) = \sum_{k=1}^{K} w_k * g(y|f_k) \quad (2)$$

Considering the hypothesis of normality, the PDF of the model forecast, $g(y|f_k)$ can be approximated to $g(y|f_k) \sim N(f_k, \sigma_k)$, where is the variance of the model prediction.

Calibration of the BMA model needs an objective function (OF). Original BMA implementation uses maximum likelihood as OF, but BMA can be adjusted using others OF, or even with a multiobjective approach (Dong, Xiong, & Yu, 2013).

Original BMA implementation uses the Expectation Maximization (EM) algorithm to adjust the parameter values (Dong et al., 2013; Qu, Zhang, Pappenberger, Zhang, & Fang, 2017; Raftery et al., 2005). These parameters are the weights for each model, and the corresponding variance for the normal PDF. EM is iterative, and composed of two steps; Expectation, when latent values (unobserved quantity) z_k^{st} are calculated given the current values of the parameters, and Maximization, when the parameters are calculated given the values of z_k^{st}.

2.2 Uncertainty assessment

Using the calibrated PDF through the EM algorithm is possible to assess the uncertainty implementing a Monte Carlo approach. The final result is a prediction uncertainty interval, based on BMA probabilistic prediction, for any time t.

This uncertainty analysis (given in (Dong et al., 2013)) choose a model from the ensemble, and generate a forecast given the normal function values for that model (normal PDF using the model forecast as the mean value, and the fitted variances as the variance of the PDF). This procedure is repeated as many times as necessary, and finally, the 5% and 95% quantiles are chosen as the boundaries of the uncertainty interval. This method can be used to generate the uncertainty intervals for every model, using the calibrated PDF from EM and the Monte Carlo sampling (Dong et al., 2013).

3 Materials

3.1 Study Case

The Sumapaz River basin is a tropical watershed part of the Magdalena-Cauca Macro Basin (MCMB) in Colombia. Compared to the MCMB, the Sumapaz is a small catchment with an area of 2,180 km2, mean annual precipitation and evapotranspiration of 1146 mm and 1102 mm, respectively. This watershed has high differences in elevation (and steep slopes), narrow canyons with deep rivers, and distinguishable mountainous and flat areas. Hydrometeorological information corresponds to in situ daily data provided by IDEAM for variables including precipitation, temperature and discharge. Information about soil texture originated from IGAC, while land use data comes from a land coverage map obtained by applying the Corine Land Cover methodology and elaborated by IDEAM, IGAC, and Cormagdalena (HEPEX, 2004). General information about the basin and the hydrological models implementation are shown in Figure 1.

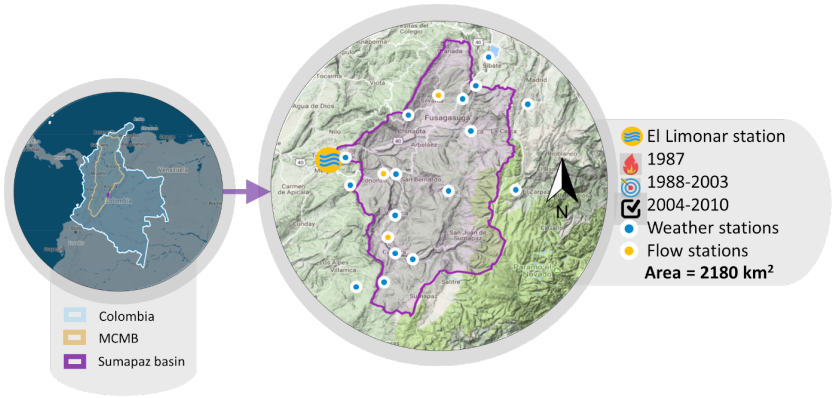


Figure 1: General location of the Sumapaz River basin in Colombia. In the legend the first line presents the discharge gauge station, second line the spin-up year, third line the calibration period and fourth line the validation period used for the implementation of each individual model in Table 1**.**

3.2 Hydrological models in the ensemble

In order to model the hydrological processes in the Sumapaz basin, physical, lumped and a semi-distributed model were implemented at a daily temporal resolution (see Table 1).

Models		General description		Calibration method	Objective function - Nash-Sutcliffe Efficiency	
		Type	Structure/Parameters		Calibration	Validation
RRL(Podger, 2004)	AWBM	CL	3T - 8Par	Pattern search multi-start	0.45	0.53
	Sacramento	CL	5T - 17Par		0.54	0.60
	SimHyd	CL	3T - 9Par		0.32	0.17
	SAMR	CL	3T - 9Par		0.41	0.57
	Tank model	CL	5T - 18Par		0.53	0.66
TOPMODEL(Beven, 1997)		PbSd	2L - 10Par	GLUE	0.34	0.25
IHACRES_1(Croke, Andrew, Spate, & Cuddy, 2005)		ECL	6Par - SES	Monte Carlo	0.61	0.33
IHACRES_3(Croke et al., 2005)			6Par - 2ESS-ISP		0.62	0.35
IHACRES_4(Croke et al., 2005)			6Par - ES-ISP		0.61	0.36
IHACRES_5(Croke et al., 2005)			6Par - 2ESP		0.63	0.34
MESH(Pietroniro et al., 2006)		PbSd	Land Surface (3 L)	Dynamically Dimensioned Search	0.14	0.11
WFLOW-HBV(J Schellekens, 2014)		ClSd	3T – 9Par	Particle Swarm Optimization	0.33	0.32

Table 1: Characteristics of the hydrological models in the ensemble, calibration techniques, objective function used and results during calibration and validation with daily data. CL: Conceptual-Lumped, PbSd: Physically based - Semi-distributed. ECL: Empirical- Conceptual – Lumped, ClSd: Conceptual – Semi distributed, Par: Parameters, T: Tanks, L: Layers, SES: Single Exponential Store, 2ESS-ISP: Two Exponential Stores and Instantaneous Stores in Parallel, ES-ISP: Exponential Store and Instantaneous Stores in Parallel, 2ESP: Two Exponential Stores in Parallel

3.3 Ensemble implementation

To assure that the assumption about normality of the PDF in every model forecast is valid, all data passed through a Johnson system of distribution, which permits to transform a variable to a standard normal distribution (see Table 2) (Chou, Polansky, & Mason, 1998).

Johnson family	Transformation	Conditions
Bounded (S_B)	$Z = \gamma + \eta \ln\left(\frac{X-\epsilon}{\lambda+\epsilon-X}\right)$	$\eta, \lambda > 0;$ $-\infty < \gamma, \epsilon, < \infty$ $\epsilon < X < \epsilon + \lambda$
Bounded from below (S_L)	$Z = \gamma + \eta \ln(X-\epsilon)$	$\eta, > 0;$ $-\infty < \gamma, \epsilon, < \infty$ $X > \epsilon$
Unbounded (S_U)	$Z = \gamma + \eta \sinh^{-1}\left(\frac{X-\epsilon}{\lambda}\right)$	$\eta, \lambda > 0;$ $-\infty < \gamma, \epsilon, < \infty$ $-\infty < X < \infty$

Table 2: Transformation associated with the Johnson system, based on (Chou et al., 1998)

The ensemble construction used all the possible combinations from the different members (i.e. the models). For example, ensemble 1 was formed by IHACRES_1 and IHACRES_5, ensemble 2 was formed by IHACRES_3, 4 and 5, TankModel and MESH. From all the combinations, the two best ensembles were selected; one based on the KGE (equation 3) and the other on the CRPS (equation 4) criteria (Kling, Fuchs, & Paulin, 2012; Vrugt, Clark, Diks, Duan, & Robinson, 2006):

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad (3)$$

Where r is the correlation coefficient between simulated and observed runoff, $\beta = \frac{\mu_s}{\mu_o}$ and $\gamma = \frac{\sigma_s}{\sigma_o}$, s meaning simulated values, o meaning observed values, μ is the mean value of the time series, and σ is the variance of the time series.

$$CRPS = \frac{1}{n}\sum_{s,t}\int_{-\infty}^{\infty}(H_{st}(x) - 1\{x \geq y\}) \quad (4)$$

Where $H_{st}(x)$ is the cumulative distribution function of x and $1\{x \geq y\}$ is the Heaviside function, giving 1 if and 0 otherwise.

Subsequently, these ensembles were evaluated using the BMA method, to determine the weights of each member in the ensemble. Inputs for the BMA correspond to daily discharges during the calibration period (1988-2003). As stated before, the PDF of the ensemble members is assumed to be normally distributed. Therefore, before using BMA, all time series were transformed with the Johnson algorithm. The parameters of the normal distribution were fitted with maximum likelihood through EM algorithm, as part of the computations associated with the BMA.

As for the next step, the best ensembles were evaluated using BMA by month (monthly weights ensembles). It means that the model weights were assessed using daily data clustered by month, for every month. On the other hand, KGE and CRPS metrics were calculated to determine the predictive performance of the EM algorithm results during the calibration and validation periods.

Uncertainty analysis was undertaken only for the two best ensembles. For each ensemble, two uncertainty bands were computed: (i) First one was calculated based on the total weights and variance of the ensemble, and (ii) second was calculated based on the monthly weights and monthly variance of the ensemble.

The results of the uncertainty assessment were evaluated using a performance criteria called: the containing ratio, defined as the percentage of observed data included in the prediction bounds (Dong et al., 2013).

4 Results and discussion

The ensemble structure, associated with the number of members, seems to be highly dependent on the performance metric used (see Figure 2). In the study area the use of the KGE metric leads to a two members ensemble [IHACRES_1 and IHACRES_5] (blue triangle in Figure 2), while the use of the CRPS score leads to a five members ensemble [IHACRES_3, IHACRES_4, IHACRES_5, TankModel and MESH] (gold triangle in Figure 2) as the best ones. Both ensembles featured an improvement in the prediction during the calibration period, with respect to the best individual model (TBM – gray dotted line in Figure 2-a. CRPS cannot be calculated for a single model).

In Figure 3, the weights (in percentage) for every member of the best ensembles (TBE) obtained with the BMA analysis are plotted, for the two metrics evaluated. Both ensembles share only one of their members, the IHACRES_5 model, but its weights present differences depending on the chosen ensemble.

When considering monthly weights, IHACRES_5 performance depends on the ensemble structure: for the KGE ensemble (two models), IHACRES_5 obtain higher weights from November to March, and on May, getting lower values on the other months; for the CRPS ensemble (five models), IHACRES_5 obtain the highest weights in four months (November, March, May and June).

The results in Figure 3 raise some questions: Why the two ensembles include a model with weights near to 1 on July, August and September? If there is one model that outperforms the other members on those months, why it is not the model the two ensembles are sharing (IHACRES_5)? This unbalanced weight distribution leads to better results on uncertainty?

These results could point to a poor ensemble performance in those months, which is solved by the BMA methodology using the model results with the best performance on those months. On the other hand, it is not clear why the two ensembles, do not share the same model from July to September, even when the weights are almost the same. This perhaps could be linked to the criteria used in every ensemble (KGE and CRPS) and its conceptual differences.

The results in Figure 3 raise some questions: Why the two ensembles include a model with weights

Figure 4 shows the multiannual evaluation of the performance of the best ensembles using daily data clustered by month. There are some irregularities on the metrics values in both ensembles throughout the year. Even if both

ensembles have similar performance on calibration and validation, the performance depends on the month of the year under assessment: best performances occur in January-March, May-June, August-September, and December (using KGE as metric) which are dry months, but this changes during the validation period. If the CRPS is used, the best performances occur from January to March and July to September, for both calibration and validation.

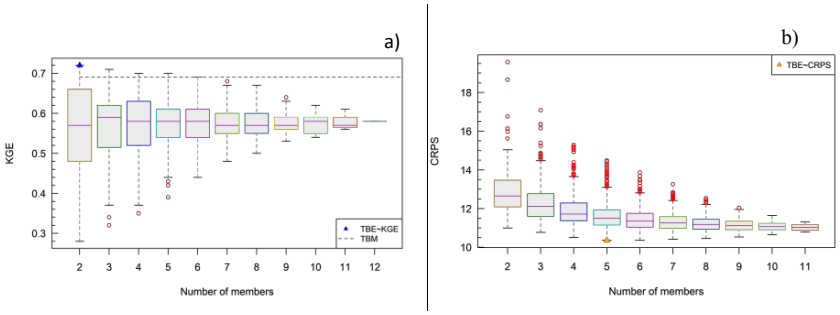


Figure 2: Performance of ensembles as a function of the number of members a) Results using KGE, and b) Results using CRPS – In both cases for the calibration period 1988-2003.

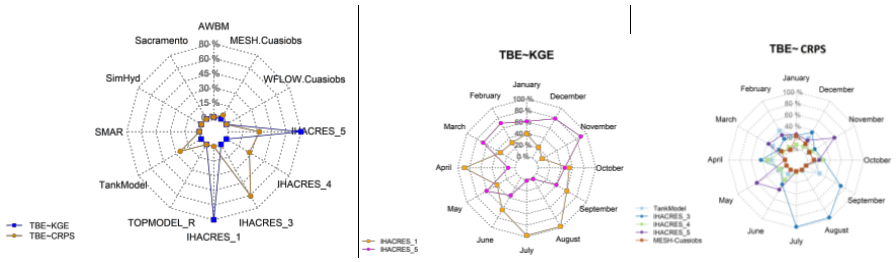


Figure 3: Model weights for the two best ensembles, calculated using the performance metrics KGE and CRPS (on the left), model weights for each month of TBE according to KGE (center) and CRPS (on the right) criteria.

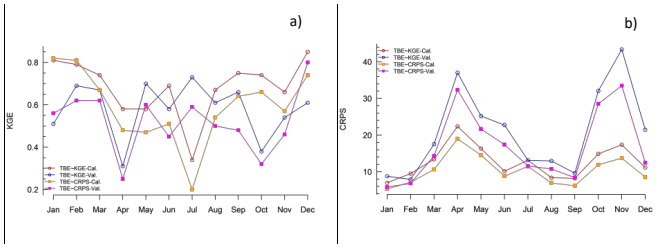


Figure 4: Monthly multi-year performance of the two best ensembles during calibration and validation periods (TBE-KGE and TBE-CRPS), a) performance using KGE, b) performance using CRPS.

The uncertainty analysis results, shown in the form of intervals, is plotted in Figure 5. Uncertainty intervals for the total BMA (BMA using all the data, without month classification) seem to perform well for the two ensembles (Figure 5), but the boundaries seem to dramatically increase for some peaks through the year 1989, especially for the CRPS ensemble. When considering the month of March 1989, the uncertainty intervals for the monthly BMA seem to work equally well even with a finer width. This is especially true for the peaks, and in this sense this could be an option to better forecast hydrological floods in the basin. Despite those differences between the total and

monthly BMA uncertainty intervals, the two approaches permit to construct consistent bands, where the observation data have great chances to be within.

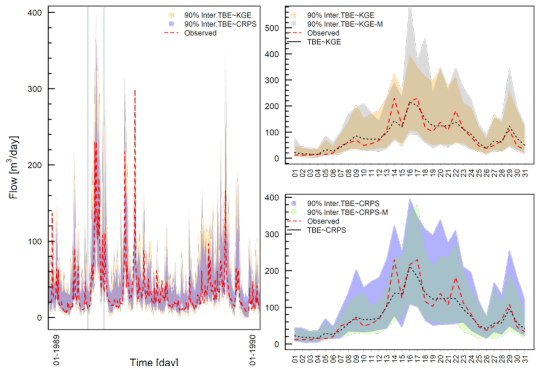
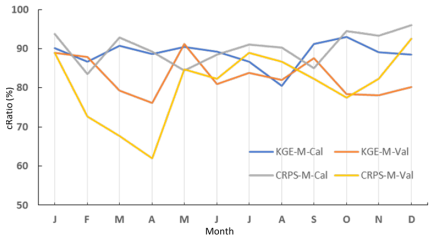


Figure 5: Uncertainty intervals of the two best ensembles with total weights for year 1989 (left) and uncertainty intervals with total values and monthly values (-M in the plot) in March-1989, for the KGE ensemble



When considering the other models, i.d. ensemble members, it is visible an unbalanced weight distribution for

Figure 6: Multiannual Mean monthly Containing ratio (cRatio) percentage of monthly KGE ensemble and monthly CRPS ensemble, for the calibration period (Cal) and the validation period (Val).

the CRPS model in July, August and September, with the IHACRES_3 model obtaining a weight near or equal to 1 (all the ensemble weight). The same happens for KGE model on the same months, but with the IHACRES_1 model, which was not included in the CRPS ensemble.
A monthly comparison of the containing ratio values between the two monthly ensembles is A monthly comparison of the containing ratio values between the two monthly ensembles is depicted in Figure 6. It is remarkable the loss of performance of the KGE ensemble on February, March and April, especially for the validation period. Meanwhile, the CRPS tends to be more constant in the cRatio values throughout the year, with a containing ratio consistently over 80%. When comparing results data from calibration and validation, differences in the KGE ensemble are higher than for the CRPS ensemble: KGE can present differences of more than 15 units for a single month (February and April), while the CRPS ensemble never shows a difference larger than 10 units.

5 Conclusions

Two model ensembles, obtained through two different performance scores (KGE and CRPS) and based on 12 individual hydrological models, have been successfully developed for the Sumapaz River basin in Colombia, using the Bayesian Model Averaging method. Results indicate that the number of members in the ensembles and their weights are highly dependent on the score selected, and that for this basin the ensembles are mainly comprised by lumped models. Results also show that although the ensembles performances are overall better than for the individual models, for some months the ensembles are not capable of adequately simulate flow discharges, even

when monthly weights are used, pointing up simulation issues for some of the members. This could mean that the ensemble itself could be used to assess the performance of its members in a deeper way, identifying moments in time when the simulations are not good (even when a global metric shows a good performance).

Some specific conclusions arise from this work: 1) It is evident that it is not possible to talk about a unique best ensemble, and that a modeller decision, like the Objective Function choice, will produce different ensemble structures. 2) The two ensembles here developed using different metrics (KGE and CRPS) seem to perform equally well in a global assessment, but they show differences in the performance throughout the year; CRPS ensemble has a reduced uncertainty, meanwhile KGE ensemble produces better forecast values. 3) Uncertainty analysis using the BMA outputs helps to better assess the ensemble performance with low computational cost; in addition, the use of monthly BMA weights permits to obtain even finer uncertainty bands, without losing performance. These three considerations are important for the use of ensembles on an operational level (for example, water management or forecasting). Finally, it is clear that those modeller decision must be taken according to the objectives the ensemble wants to fulfil: if a lower uncertainty is the final objective of the modelling, for local water management for example, a set-up ensemble based on the CRPS criteria could the best option; on the other hand, if better forecast values are needed, for regional water management, KGE based ensemble could be a better option.

References

Beven, K. (1997). TOPMODEL: a critique. Hydrological Processes, 11(9), 1069–1085. https://doi.org/10.1002/(SICI)1099-1085(199707)11:9<1069::AID-HYP545>3.0.CO;2-O

Brown, J. D., Demargne, J., Seo, D.-J., & Liu, Y. (2010). The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environmental Modelling & Software, 25(7), 854–872. https://doi.org/10.1016/j.envsoft.2010.01.009

Chou, Y.-M., Polansky, A. M., & Mason, R. L. (1998). Transforming non-normal data to normality in statistical process control. Journal of Quality Technology, 30(2), 133–141. https://doi.org/http://rube.asq.org/data/subscriptions/jqt_open/1998/april/jqtv30i2chou.pdf

CORMAGDALENA, IDEAM. (2001) Estudio ambiental de la cuenca Magdalena-Cauca y elementos para su ordenamiento territorial. Resumen Ejecutivo. [online] 13-02-2016. http://documentacion.ideam.gov.co/openbiblio/bvirtual/000051/EstudioAmbientalCMagdalena-Cauca.pdf

Croke, B., Andrew, F., Spate, J., & Cuddy, S. (2005). IHACRES User Guide. Retrieved from http://www.toolkit.net.au/ihacres

Dong, L., Xiong, L., & Yu, K. (2013). Uncertainty Analysis of Multiple Hydrologic Models Using the Bayesian Model Averaging Method. Journal of Applied Mathematics, 2013, 1–11. https://doi.org/10.1155/2013/346045

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. Monthly Weather Review, 129(3), 550–560. https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2

HEPEX (2004) Hydrologic Ensemble Prediction Experiment. Workshop. Reading, England. 2004. [online] https://hepex.irstea.fr/.

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

Montgomery, J., Hollenbach, F., and Ward, M. (2015). Calibrating ensemble forecasting models with sparse data in the social sciences. International Journal of Forecasting., vol. 31, no. 3, pp. 930-942.

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., … Pellerin, P. (2006). Using the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale. Hydrology and Earth System Sciences Discussions, 3(4), 2473–2521. https://doi.org/10.5194/hessd-3-2473-2006

Podger, G. (2004). Rainfall runoff library user guide. Cooperative Research Centre for Catchment Hydrology. Retrieved from www.toolkit.net.au/rrl

Qu, B., Zhang, X., Pappenberger, F., Zhang, T., & Fang, Y. (2017). Multi-model grand ensemble hydrologic forecasting in the Fu river basin using Bayesian model averaging. Water (Switzerland), 9(2). https://doi.org/10.3390/w9020074

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly Weather Review, 133(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Schellekens, J. (2014). OpenStreams wflow documentation release 1.0RC1. Deltares. Retrieved from http://wflow.readthedocs.org/en/latest/

Schellekens, J., Dutra, E., Martínez-De La Torre, A., Balsamo, G., Van Dijk, A., Sperna Weiland, F., … Weedon, G. P. (2017). A global water resources ensemble of hydrological models: The eartH2Observe Tier-1 dataset. Earth System Science Data, 9(2), 389–413. https://doi.org/10.5194/essd-9-389-2017

Thielen-del Pozo, J., Pappenberger, F., Salomon, P., Bogner, K., Burek, P., & de Roo, A. (2010). The state of the art of flood forecasting - Hydrological Ensemble Prediction Systems. EMS Annual Meeting Abstracts, 7(March 2016), 1.

Vrugt, J. A., Clark, M. P., Diks, C. G. H., Duan, Q., & Robinson, B. A. (2006). Multi-objective calibration of forecast ensembles using Bayesian model averaging. Geophysical Research Letters, 33(19), 2–7. https://doi.org/10.1029/2006GL027126