On the difficulty to optimally implement the Ensemble Kalman Filter: An experiment based on many hydrological models and catchments

A. Thiboult[a,*], F. Anctil[a]

[a]*Dept. of Civil and Water Engineering, Université Laval, Québec, Canada*

Published in Journal of Hydrology 529 (2015) 1147–1160
https://doi.org/10.1016/j.jhydrol.2015.09.036

Abstract

Forecast reliability and accuracy is a prerequisite for successful hydrological applications. This aim may be attained by using data assimilation techniques such as the popular Ensemble Kalman filter (EnKF). Despite its recognized capacity to enhance forecasting by creating a new set of initial conditions, implementation tests have been mostly carried out with a single model and few catchments leading to case specific conclusions. This paper performs an extensive testing to assess ensemble bias and reliability on 20 conceptual lumped models and 38 catchments in the Province of Québec with perfect meteorological forecast forcing. The study confirms that EnKF is a powerful tool for short range forecasting but also that it requires a more subtle setting than it is frequently recommended. The success of the updating procedure depends to a great extent on the specification of the hyper-parameters. In the implementation of the EnKF, the identification of the hyper-parameters is very unintuitive if the model error is not explicitly accounted for and best estimates of forcing and observation error lead to overconfident forecasts. It is shown that performance are also related to the choice of updated state variables and that all states variables should not systematically be updated. Additionally, the improvement over the open loop scheme depends on the watershed and hydrological model structure, as some models exhibit a poor compatibility with EnKF updating. Thus, it is not possible to conclude in detail on a single ideal manner to identify an optimal implementation; conclusions drawn from a unique event, catchment, or model are likely to be misleading since transferring hyper-parameters from a case to another may be hazardous. Finally, achieving reliability and bias jointly is a daunting challenge as the optimization of one score is done at the cost of the other.

Keywords:
Data assimilation, Uncertainty estimation, Ensemble Kalman filter

1. Introduction

Despite the modelling advances in representing hydrological processes and providing more accurate streamflow forecasts, there is still a need for reducing and quantifying uncertainty. Most hydrological prediction systems are still deterministic and provide only the most likely outcome without addressing estimates of their uncertainty. The sources of uncertainty stem from multiple places in the hydrometeorological chain such as in inputs, initial conditions, parameter estimation, model structure, and outputs (e.g. Ajami et al., 2007; Salamon and Feyen, 2010; Liu and Gupta, 2007; Liu et al., 2012) and these uncertainties should be deciphered to enhance model predictive abilities and reliability for efficient decision making (Ramos et al., 2010).

A broad range of techniques has been developed to control uncertainty at different levels such as the Generalized Likelihood Uncertainty Estimation (GLUE), Shuffle Complex Evolution Metropolis algorithm (SCEM) for parameter uncertainty (Beven and Binley, 1992; Vrugt et al., 2003) and BMA combination technique for structural uncertainty (Jeremiah et al., 2011; Duan et al., 2007; Parrish et al., 2012; Ajami et al., 2007). Proper initial conditions are frequently identified as one of the main factors that contributes to an accurate forecast (DeChant and Moradkhani, 2011; Lee et al., 2011). Among others, data assimilation (DA) is commonly used in hydrometeorology to reduce initial condition uncertainty and proved to be a useful tool for modelling. DA incorporates observations into the numerical model to issue an analysis, which is an estimation of the best current state of the system. This has not only been largely applied to remote sensing for snow (Kuchment et al., 2010), soil moisture estimates (Forman et al., 2012; Meier et al., 2011; Renzullo et al., 2014; Alvarez-Garreton et al., 2014) or hydraulic information (Bailey and Bau, 2012), but also to update radar forcing (Harader et al., 2012; Kim and Yoo, 2014). Many applications also use in situ observations such as catchment discharge, snowpack measurements, or soil moisture to update models (e.g., Seo et al., 2009; Clark et al., 2008; Thirel et al., 2010; DeChant and Moradkhani, 2011; Franz et al., 2014). In addition, DA may be coupled with parameter optimization (Vrugt et al., 2005; Moradkhani et al., 2005; Nie et al., 2011).

*Corresponding author
Email address: antoine.thiboult.1@ulaval.ca (A. Thiboult)

Sequential DA techniques such as particle filter and the Kalman filter family are frequently used for recursive updating of the states of a system, each time an observation is made available. Among them, the ensemble Kalman filter (EnKF, Evensen, 1994) proved to be a powerful tool for hydrological forecasting (DeChant and Moradkhani, 2012; Rakovec et al., 2012; Vrugt and Robinson, 2007; Weerts and El Serafy, 2006; Abaza et al., 2014) that is effective and reliable enough for operational use (Andreadis and Lettenmaier, 2006). Several studies claim that they developed techniques that improved upon traditional EnKF (e.g., Clark et al., 2008; Whitaker and Hamill, 2002) by focusing on the relaxation of constrains of traditional EnKF implementation, or by explicitly including time lag between the soil moisture and the discharge in the updating process (Li et al., 2013, 2014; McMillan et al., 2013).

A key feature of EnKF is the proper specification of hyper-parameters (perturbations of inputs and outputs) and model states to be updated (Moradkhani et al., 2005). In most studies, EnKF implementation is based on an a priori selection of the hyper-parameters and updated states combination, which is then scarcely justified. Noteworthy exceptions are Moradkhani et al. (2005) and Chen et al. (2013), but these studies are very specific as they are performed on a single model and one or two catchments. Accurate perturbations representing error estimates are crucial since the EnKF updating scheme is based on the weighting of the model and observation relative error. However this specification is complex in practice as the different sources of uncertainty experience strong interactions (Moradkhani et al., 2006; Hong et al., 2006; Kuczera et al., 2006). Several attempts to account explicitly for structural error have been reported, for example by directly adding perturbations to the state variables (Reichle et al., 2002; Vrugt et al., 2006; Clark et al., 2008), or by updating model parameters (Moradkhani et al., 2005; Vrugt et al., 2005; Naevdal et al., 2003).

Moreover, despite encouraging results, DeChant and Moradkhani (2012) point that little research has been done to examine the effectiveness and robustness of EnKF and that "studies need to provide a more rigorous testing of these techniques than has previously been presented". Another issue that needs consideration is that EnKF performance is mostly discussed as 'standalone', regardless of the influence of the coupling with the hydrological model. This is mainly due to the fact that EnKF is often tested on a single model. Thus, the question of adequacy between the DA technique and the model is rarely assessed.

The present study aims at identifying EnKF parametrization to reduce and quantify optimally the uncertainty related to initial conditions in a forecast mode. A second scope addresses the question of EnKF and hydrological model adequacy. In order to achieve this, the analysis is conducted on 20 structurally dissimilar lumped conceptual models, 38 catchments, 12 hyper-parameter sets, and all possible combinations of the state variables to strive for general results. Finally, the effectiveness of identifying the best EnKF parametrization without exploring all combinations is discussed.

Section 2 presents EnKF's basics, models, basins and scores. Section 3 presents the results of the DA techniques followed by a discussion and the conclusion statements are provided in section 4.

2. Material and methods

2.1. Hydrological models, snowmelt modules, and PET

The EnKF is tested individually on 20 lumped conceptual models, which differ by their structure. The selection was initially carried out by Perrin (2000) and revised by Seiller et al. (2012) for hydrological projection purposes. Because they are based on diverse hydrological concepts and present different degrees of complexity (4 to 10 calibrated parameters and 2 to 7 reservoirs to represent perceptual and conceptual hydrologic processes), they allow to test the EnKF in a comprehensive manner according to structure diversity (see Table 1). The models have been modified to match a common frame and they should not be directly compared to their original version. In the case where the original models included a module to compute evapotranspiration or snow accumulation and melting, the module has been omitted as these processes are computed externally beforehand.

The models exploit various conceptualizations and thus their parameters and state variables perform particular roles in simulating rainfall-runoff processes. Their reservoirs may describe systems ranging from precipitation interception to routing (or more conceptual functions). The role of state variables is not detailed in the article for concision purpose. For the same reason, the state variable values before and after the analysis step will not be discussed here but only the outputs of the models, i.e., simulated streamflow will be considered. For further details on state variable meaning, refer to Perrin (2000).

The lumped models are driven by potential evapotranspiration and precipitation. The potential evapotranspiration is computed from the formula proposed by Oudin et al. (2005), which relies on mean air temperature and the calculated extraterrestrial radiation. To partition snow accumulation, snowmelt, and liquid precipitation, the snow module (Cemaneige, Valery et al., 2014) is executed before hydrological models.

The snow module divides the watershed into 5 elevation bands and is based on a degree-day approach modulated by an energy balance index to simulate the dynamic of the snowpack.

The models are calibrated with the Shuffle Complex Evolution algorithm (Duan et al., 1992) and the $RMSE$ is used on square-rooted streamflows as objective function to ensure that calibration does not favor either low or high streamflows. Each of the 20 models is calibrated individually with the 2-parameter snow module — the parameter values of the snow

Table 1: Main characteristics of the 20 lumped models (Seiller et al., 2012)

Model acronym	No of param.	No of reserv.	Derived from
M01	6	3	BUCKET (Thornthwaite and Mather, 1955)
M02	9	2	CEQUEAU (Girard et al., 1972)
M03	6	3	CREC (Cormary and Guilbot)
M04	6	3	GARDENIA (Thiery, 1982)
M05	4	2	GR4J (Perrin et al., 2003)
M06	9	3	HBV (Bergström and Forsman, 1973)
M07	6	5	HYMOD (Wagener et al., 2001)
M08	7	3	IHACRES (Jakeman et al., 1990)
M09	7	4	MARTINE (Mazenc et al., 1984)
M10	7	2	MOHYSE (Fortin and Turcotte, 2007)
M11	6	4	MORDOR (Garçon, 1999)
M12	10	7	NAM (Nielsen and Hansen, 1973)
M13	8	4	PDM (Moore and Clarke, 1981)
M14	9	5	SACRAMENTO (Burnash et al., 1973)
M15	8	3	SIMHYD (Chiew et al., 2002)
M16	8	3	SMAR (O'Connell et al., 1970)
M17	7	4	TANK (Sugawara, 1979)
M18	7	3	TOPMODEL (Beven et al., 1984)
M19	8	3	WAGENINGEN (Warmerdam et al., 1997)
M20	8	4	XINANJIANG (Zhao et al., 1980)

module are consequently different for each hydrological model.

2.2. *Experimental design, state updating, and EnKF implementation*

EnKF addresses explicitly initial conditions uncertainty by creating an ensemble of possible model reinitializations by updating state variables according to a recursive Bayesian estimation scheme. It estimates the true probability density function of the model states conditioned by the observations.

The evolution of the model state variables vector x may be described through time with a non-linear forward operator M driven by the previous states, the deterministic forcing u that includes an error term ζ_t, and the (time-invariant) model parameters θ to which a model error η is added. The η error term does not include only state variable error but also implicitly other sources of error such as the structural and parameter error or forcing error.

$$x_t = M(x_{t-1}, u_{t-1}, \theta) + \eta_t \tag{1}$$

States and observations z are related through the following expression

$$z_t = H(x_t) + \epsilon_t \tag{2}$$

with H being the observation function and ϵ_t the observation error.

The EnKF relies on an approximation of Bayesian rule to identify the conditional density of the model states, $p(x_t|z_{1:t})$, given the previous time steps observations $z_{1:t}$, where x_t is the state vector that contains the model states. EnKF needs several realisations (N members) to derive the model error matrix. As the real true state is unknown, it is approximated by the ensemble mean :

$$\overline{x_t} = \frac{1}{N}\sum_{i=1}^{N} x_t^i \tag{3}$$

where i refers to the i$^{\text{th}}$ member. The model error matrix is thus defined as the difference between the true state and the single hydrological model realisations:

$$\mathbf{E}_t = (x_t^1 - \overline{x_t}, x_t^2 - \overline{x_t}, ..., x_t^N - \overline{x_t}) \tag{4}$$

Therefore, the model covariance matrix can be defined as:

$$\mathbf{P}_t = \frac{1}{N-1}\mathbf{E}_t\mathbf{E}_t^T \tag{5}$$

When an observation is available, model states are updated ($\mathbf{X}^+$) as a combination of the prior states $\mathbf{X}^-$ and the difference between the prior estimate $\mathbf{H}_t\mathbf{X}_t^-$ and observation.

$$\mathbf{X}_t^+ = \mathbf{X}_t^- + \mathbf{K}_t(z_t - \mathbf{H}_t\mathbf{X}_t^-) \tag{6}$$

The Kalman gain $\mathbf{K}_t$ represents the relative importance of the observation error with respect to the prior estimate (i.e. model simulation) and acts as a weighting coefficient. $\mathbf{R}_t$ denotes the covariance of the observational noise.

$$\mathbf{K}_t = \mathbf{P}_t\mathbf{H}_t^T(\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T + \mathbf{R}_t)^{-1} \tag{7}$$

Since the identification of the Kalman gain is complex, the term $\mathbf{P}_t\mathbf{H}_t^T$ is approximated by the forecasted covariance between the model states and the simulation estimates and $\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T$ by the variance of the estimate.

$$\mathbf{P}_t\mathbf{H}_t^T = \frac{1}{N-1}\sum_{i=1}^{N}(x_t^i - \overline{x_t})(\mathbf{H}_t x_t^i - \overline{\mathbf{H}_t x_t})^T \tag{8}$$

$$\mathbf{H}_t\mathbf{P}_t\mathbf{H}_t^T = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{H}_t x_t^i - \overline{\mathbf{H}_t x_t})(\mathbf{H}_t x_t^i - \overline{\mathbf{H}_t x_t})^T \tag{9}$$

A more detailed description of EnKF equations and mathematical background can be found in Evensen (2003). In this study, the filter has been implemented following Mandel's (2006) computational recommendations.

A critical point in the EnKF implementation is a proper mapping of the errors ϵ, ζ, and η because they will determine the observation predictive distribution. In a very vast majority

of cases, model error is not directly identifiable. Users need to estimate it by using stochastic perturbations through ϵ and ζ computation if no more direct estimation of model error η is made through perturbation of states or updating state variable and parameters conjointly. Note that adding perturbations to states and parameter updating are also subject to inaccuracies since they are "based on order-of-magnitude considerations, and may therefore be statistically unreliable" (Liu et al., 2012). In the present study, only ϵ and ζ are considered. Errors are assumed to be normally distributed with zero mean but their variances need to be put under scrutiny.

Hyper-parameters define the statistical properties of the forcing and observations ensembles. This study concentrates on the influence of the uncertainty in precipitation and temperature forecasts and in streamflow observation. Three precipitation perturbations (with a standard deviation corresponding to 25%, 50%, and 75% of the initial precipitation forecast magnitude), two streamflow perturbations (with a standard deviation corresponding to 10% and 25% of the observation), two temperature perturbations (standard deviation of 2°C and 5°C) are evaluated. These perturbations are centred around the perfect forecast or the observation. We thus obtain 12 sets of hyper-parameters. All errors are assumed to be uncorrelated. Note that the potential evapotranspiration is not directly perturbed but it is computed by the Oudin formula forced with a temperature ensemble creating a subsequent set of perturbed PET values.

The present updating scheme relies on the Markov property that asserts that the future of the system is dictated only by the present state, not on the anterior sequence of observations. Model states are consequently updated according to the instantaneous covariance between states and the current streamflow observation while observations that preceded it are not incorporated. Li et al. (2013) affirms that this assumption may harm updating performance of models that incorporate unit-hydrograph routing but do not affect models that include dynamic routing stores, which is the case for 19 of the 20 models used in this study. Only model 5 (GR4J) is based on a unit-hydrograph approach.

Prior to the hyper-parameters evaluation, the number of members composing the EnKF ensemble is investigated. Four sizes (25, 50, 100 and 200 members) are tested on two sets of hyper-parameters. The experiment concerning the number of members was not conducted on all hyper-parameter sets to reduce computational cost.

The number and the combination of states to be updated are next put under scrutiny. Batch testing is used to investigate all states (reservoirs) combinations for each model regardless of their physical meaning. The number of possible combinations thus depends on the model at hand, varying from 3 for the 2-reservoir models up to 127 for the 7-reservoir model. As all combinations of state variables and hyper-parameters are tested, some cases turned out to be unrealistic and prone to make the EnKF unstable. This difficulty was overcame by setting back unrealistic states within their theoretical boundaries identified during calibration.

The EnKF is used to update daily model's states whenever streamflow observations are available. The model is then forced with the perfect meteorological forecast to issue a 10-day hydrological forecast. This lead time is sufficiently long to be able to see the effect of DA vanishing. A series of tests (not shown here) indicated that after 10 days, the influence of data assimilation is almost negligible for almost every model and catchments. Thus extending the forecast ahead would bring no additional information. This framework is comparable to an operational one except for the perfect meteorological forcing.

Finally, every 20 models are tested on 38 catchments, 12 different hyper-parameter sets, and all possible reservoir combinations.

2.3. Scores

Probabilistic scores offer the possibility to evaluate more than individual member or ensemble mean and provide to the forecaster a better picture of the forecast probability distribution by expressing the uncertainty level. Probabilistic forecasts should be assessed both in terms of bias and reliability to assess where the verification is situated among the ensemble, how the frequency of forecasted events corresponds to the frequency at which events are observed, and the gain of ensemble over deterministic forecast.

The Normalized Root-mean-square error Ratio (*NRR*) is used to quantify the spread of the ensemble with regard to its predictive skills (Murphy, 1988). A value of 1 indicates an appropriate spread, while greater and smaller values than 1 reflect too narrow and wide ensembles respectively. The *NRR* is function of the observation y_t, the ensemble forecast average $\bar{\hat{y}}_t$, the number of members in the ensemble N, and time t.

$$NRR = \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\left[\frac{1}{N}\sum_{n=1}^{N}\hat{y}_t^n\right]-y_t^n\right)^2}}{\frac{1}{N}\left\{\sum_{n=1}^{N}\sqrt{\frac{1}{T}\left[\sum_{t=1}^{T}(\hat{y}_t^n-y_t^n)^2\right]}\right\}\sqrt{\frac{N+1}{2N}}} \tag{10}$$

A complementary view of the *NRR* is the Spread Skill Plot (SSP) which is a graphical assessment that represents at the same time the bias of the ensemble, its spread and therefore its reliability. The SSP relies on the fact that the Root-Mean-Square Error (RMSE) should match the spread to achieve reliability (Fortin et al., 2014). In the case where the RMSE is greater than the spread, the ensemble is overconfident regarding to its predictive skills and vice versa.

The commonly used Nash Sutcliffe efficiency (Nash and Sutcliffe, 1970) is used to assess the bias of the median of the

EnKF ensemble. A NSE value equals to 1 identifies a perfect prediction while a value below 0 indicates that the average observation is more skilful.

To evaluate the improvement or deterioration of the quality of the simulations, NSE and NRR gains are computed as:

$$G = \frac{S^{sim} - S^{ref}}{S^{opt} - S^{ref}} \tag{11}$$

with G the gain, S^{sim} the score after the state updating, S^{ref} the score without updating (the open loop), and S^{opt} the perfect score. In the case of the NRR, which is not a monotonic score (i.e. the minimization or maximization of the value does not systematically indicate an improvement or a decrease of the reliability), a substitution is performed to compute the gain. We consider that under and overdispersion should be penalized the same way which is reflected by the distance from the optimal score 1.

$$NRR^{*} = |NRR - 1| \tag{12}$$

NRR^{*} is then negatively oriented and bounded by 0 and can be used to compute the gain G.

2.4. Catchments and hydrometeorological data

38 watersheds are evaluated for this study. They are mainly situated in the south of the Province of Québec, but some extend over Ontario or the north of the states of New York and Vermont (Fig. 1). Their latitudes range from 43°15'N to 52°20'N and they exhibit important winter snow cover. Consequently, the hydrological regime is dominated by a spring freshet and a second peak in autumn is frequently observed.

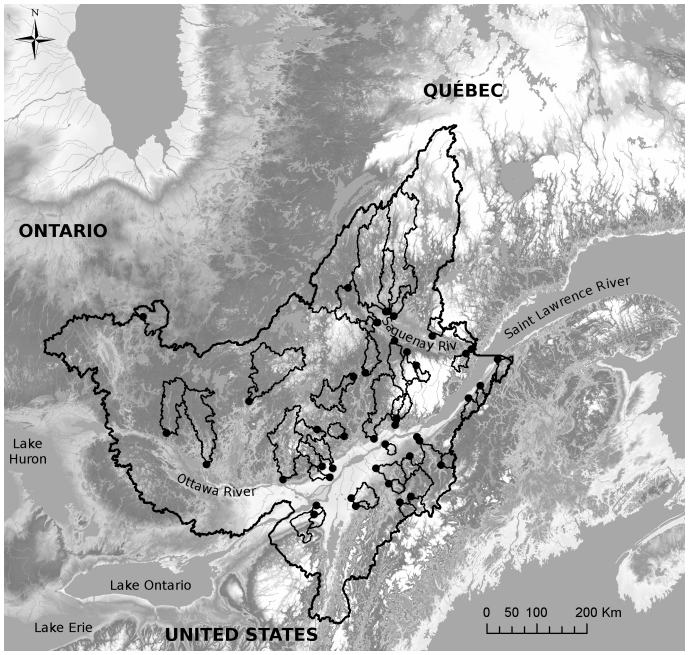

Figure 1: Spatial distribution of the watersheds

The size of the catchments ranges from 236 km^2 to 15342 km^2 and the median annual discharge varies from 5 m^3/s to 299 m^3/s. Maximal solid and total precipitation are respectively 501 mm and 1544 mm while minima are 218 mm and 985 mm. Solid precipitation is an estimation derived from a snowmelt model forced with rainfall and temperature observations.

16 years of daily streamflow, total precipitation, and maximum and minimum temperature are used. The meteorological dataset was created by the Centre d'Expertise Hydrique du Québec by kriging (interpolating) observations over a 1220-point grid at a 0.1° resolution. Temperature were then corrected by applying an elevation-based temperature gradient of -0.005°C/m. In this study, 10 years (1990-2000) are dedicated to model calibration, 3 years (October 2005 to October 2008) are used for model's warm up, and the period from October 2008 to December 2010 is dedicated to hydrological forecast assessment that is issued up to 10 days ahead.

2.5. Model performance in calibration

The study investigates EnKF implementation for many models i.e. the performance of the coupling of hydrological models and DA. It is not intended to compare model performances to each other. The 20 models are thus investigated separately and the comparison of models performance with or without EnKF updating is out of scope. However, one should recognize that models have different initial performance as shown in Figure 2. Their individual performance varies largely over the 38 catchments but no model consistently out performs or under performs the others in all situations. Best (or worst) results are frequently obtained by different models for different catchments. Therefore, performance after EnKF updating should be compared to each other only in terms of gain.

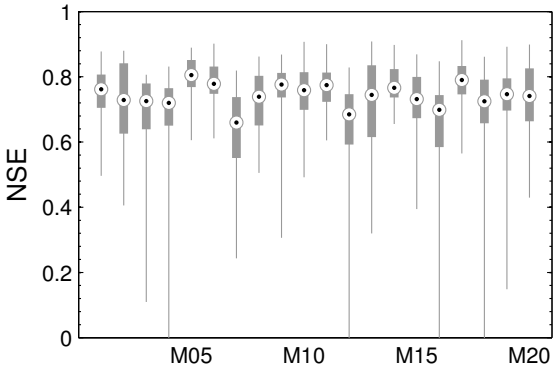

Figure 2: NSE of the 20 models over the 38 catchments. Each box plot correspond to a model.

2.6. Meteorological forecast

The first scope of this paper is to reduce and quantify the uncertainty related to the watershed initial conditions for hydrological forecasting. Thus, to focus on that specific aspect,

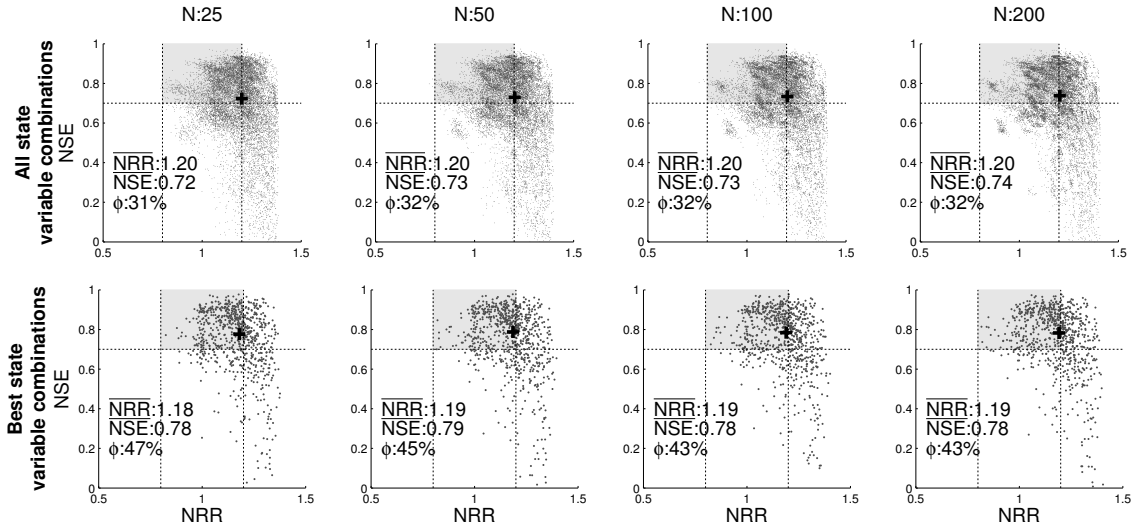

Figure 3: Influence of the number of members N on NSE and NRR in simulation

we do not use actual weather forecasts but meteorological observations to force the models. This ensures to minimize the error related to forcing as the remaining inaccuracies can be attributed to the representativeness of the measurements and the measurement errors rather than the many uncertainties related to weather forecasting. This forcing will be referred hereafter as 'perfect forecast'.

3. Results

3.1. An estimation of the required ensemble size

Ideally, one would propagate a large number of members to ensure to accurately sample the state variable probability density functions but this would increase drastically the computational cost. Thus, the first part of the study aims at identifying an approximation of the minimal number of member necessary to drive EnKF without performance loss. For this case, the hyper-parameters are fixed to $0.50P$ for precipitation, $0.1Q$ for streamflow, and 2^{o} for temperature for graphical convenience, and different numbers of members are tested ($N = 25, 50, 100, 200$). The influence of the number of members has been carried out with other sets of hyper-parameters and led to the same conclusions.

Figure 3 shows general behaviour according to the number of members N. NSE and NRR are used to asses forecast accuracy and reliability, respectively. On the upper sub-plots are displayed 12768 points corresponding to all simulations per set of hyper-parameters, i.e the results in simulation for every catchment, model and existing reservoir combination for a given model. As we want to quantify the effect of N on every simulation but also more precisely on best performing ones, the lower sub-plots display the best results by catchment and model. This implies to retain only the best state variable combination for updating. A single best simulation can be identified for a particular score for a catchment, but the simulation may be different whether it is assessed regarding its reliability or its bias. To overcome this selection issue, we retained the simulation offering the highest NSE among the three best NRR. This combined criterion ensures to keep the most reliable simulation in first place and then the lowest bias. Reliability is chosen as first order criterion to ensure to cover as well as possible the initial condition uncertainty without diminishing the EnKF spread.

Interest is set on forecasts with NSE and NRR close to 1, thus the vertical axis has been truncated for readability. Negative NSE are not shown even if they represent about 1.5% of the total number of simulations. Note that the NRR score is bounded for underdispersed distribution by $1/\sqrt{(N+1)/2N}$ and thus a NRR score below 0.8 or above 1.2 indicates doubtlessly a poorly reliable ensemble. For this purpose, an area is defined to delimit the ranges for deemed acceptable results ($0.8 < NRR < 1.2$ and $0.7 < NSE$) and is represented as a grey shade on Fig 3. The ratio ϕ of simulations having performance falling inside the aforementioned range to total number of simulations is displayed for every N. Additionally, median NSE and median NRR of simulations are depicted as a cross on each plot. The range of acceptable results may seem permissive, but it has to be wide enough to encompass a reasonable number of models and watersheds. By defining a more demanding range, there is a risk that all the points inside it belong to a small number of model and catchment pairs and that the variations inside this range are only due to the state variable choice which therefore harms the representativeness of the ϕ ratio.

Results are very similar for different values of N, for all the simulations or only the best ones. The sampling of the states variable probability density function is more subject to stochastic errors when the number of member is low, but

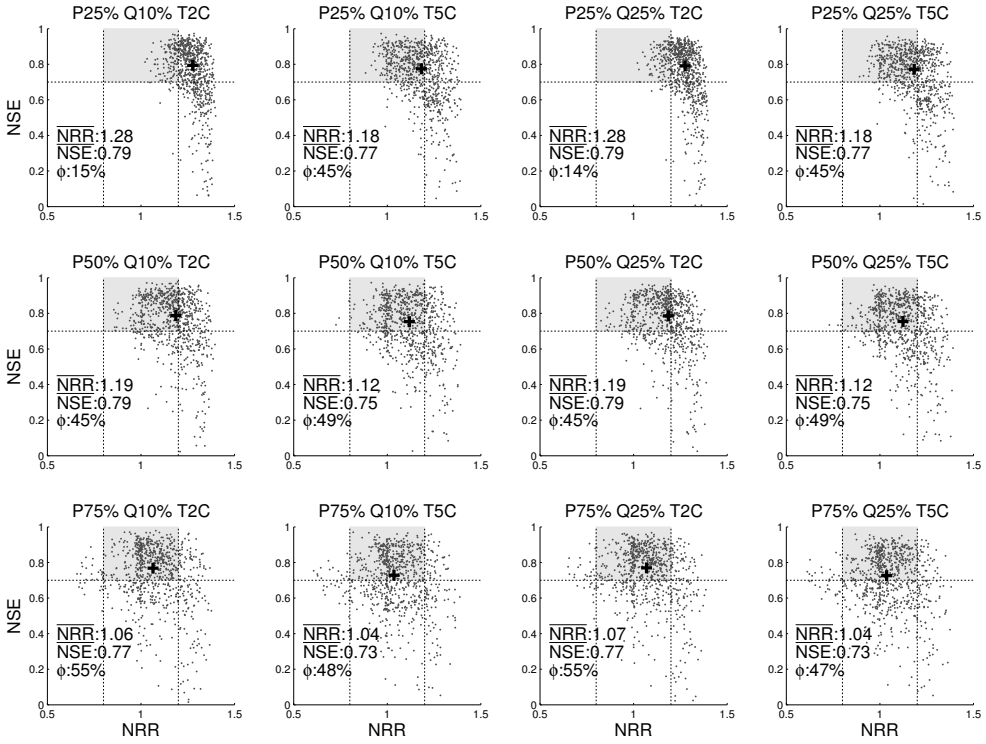

Figure 4: Influence of added perturbations to precipitations, temperatures and streamflows on NSE and NRR for lead time 1 and 50 members

results are largely similar with various values of N. The weak difference between the number of members indicates that it is reasonable to keep only 25 or 50 members to limit sampling error. Further results will be presented for 50 members.

3.2. Influence of hyper-parameters

Figure 4 depicts the bias and reliability for the 12 hyper-parameter sets for day 1. Unlike the number of members, the additional error to forcing and observation is a driving parameter. The precipitation perturbations have the greatest influence on performance followed by temperature and streamflow. Note that the importance of the temperature perturbations has to be considered regarding the dominant role of the spring freshet for the studied watersheds.

For a given hyper-parameter set, the cloud of simulations is greatly dispersed on both reliability and bias axes. The diversity of performance is important and depends on model and catchment for a specific hyper-parameter set.

Reliability is more often achieved by using important perturbations. The lowest perturbation set (P 25%, T 2°C, Q 10%) fails to encompass initial condition uncertainty as only 15% of simulations fall into the acceptable range of performance. Reliability best results are obtained with perturbations that are clearly unrealistically high to describe measurement uncertainty. However, few simulations using a hyper-parameter set that include low input perturbations, are close to perfect reliability. These simulations are very likely to become over-dispersive when the added noise magnitude increases, indicating that using too large perturbations also contributes to decrease reliability.

On the other hand, best NSE are found for smaller perturbations. Though, the calibration of hyper-parameters is more sensitive to reliability as the drop in bias is less severe than the drop in reliability.

These results indicate that in a vast majority of cases, EnKF should not be used with best estimates of real forcing uncertainty as it will lead to under-dispersive ensemble if no additional technique is used to explicitly decipher other sources of uncertainty. To achieve optimal implementation, the noisy forcing has to take into account not only real forcing uncertainty but it needs to compensate for the model structural and parameter uncertainty. This contributes to drastically increase the difficulty of identifying the correct covariances.

Figure 5 displays the same bias-reliability representation in forecasting mode, but for the 3-day-ahead lead time where a global loss of reliability is observed. Ensembles become overconfident with increasing lead time but the bias remains approximatively the same. Only a 75% perturbation of the precipitation manages to provide more than 30% of acceptable results. For the 7th day (result not shown here), the percentage of acceptable results never exceeds 13%. This clearly suggests that the uncertainty in initial conditions does not account for all sources of uncertainty, even if it encompasses more than real forcing and observation error. For instance, errors associated to

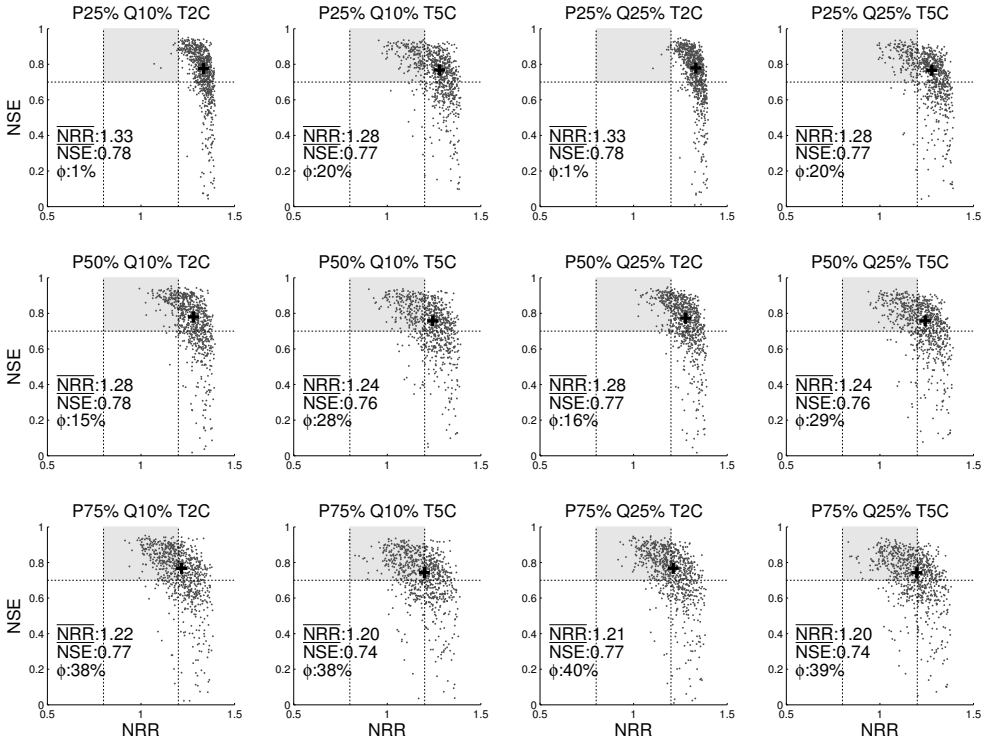

Figure 5: Influence of added perturbations to precipitations, temperatures and streamflows on NSE and NRR for lead time 3

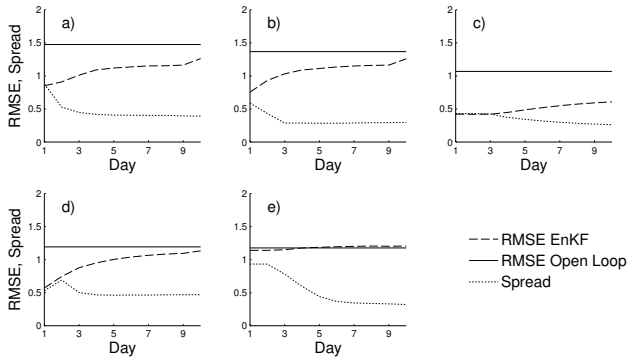

Figure 6: Typical spread-skill plots in forecasting. a) model 9 and watershed 28, b) model 5, watershed 28, c) model 1, watershed 36, d) model 5, watershed 26, e) model 7, watershed 13.

the meteorology also needs to be accounted for as the forecast horizon increases (Velázquez et al., 2011).

Relative performance of hyper-parameters remains unchanged with lead time as the best performing hyper-parameter sets for reliability or bias are the same for day 1 or day 9. The model's performance in forecasting thus depends on the quality of the DA initialization.

Typical results are plotted on Figure 6 to illustrate the spread and the bias ($RMSE$) of the EnKF ensemble and the gain which is related to the difference between EnKF $RMSE$ and the open loop $RMSE$. The most common situations are cases a) and b) in which the ensemble is reliable or close to reliability for the first forecasting day as the spread matches or is close to match the EnKF $RMSE$. The spread diminishes quickly after the first forecasting day while the error of the ensemble increases. Although the ensemble spread should grow to match the increasing error, it collapses and the ensemble becomes overconfident. However, the error remains lower than the open loop forecasting confirming the gain provided by the EnKF.

Less frequently, the spread may remain constant up to three days (case c) or for a very particular situation as in case d), which happens only for model number 5 (GR4J), the spread may increase for up to 2 days before eventually dropping. Finally, case e) reports a case where DA is unable to improve forecasting. In the latter case, EnKF simulation is quite similar to the open loop. This is explained by the state variable selection process where only the best state variable combination is kept. For this very particular model-watershed pair, DA works poorly and all state variable combinations deteriorate accuracy beyond the open loop performance. Consequently, the best simulation is achieved by the state variable combination influencing the least the streamflow and this combination is therefore selected. In an operational context and such case, the EnKF would not be used as it does not provide any gain and increases computational costs vainly.

In every cases, DA assimilation starts loosing its efficiency right after the spin up and the spread stops matching the EnKF $RMSE$ for longer lead times. This confirms that additional sources of uncertainty should be taken into account from the

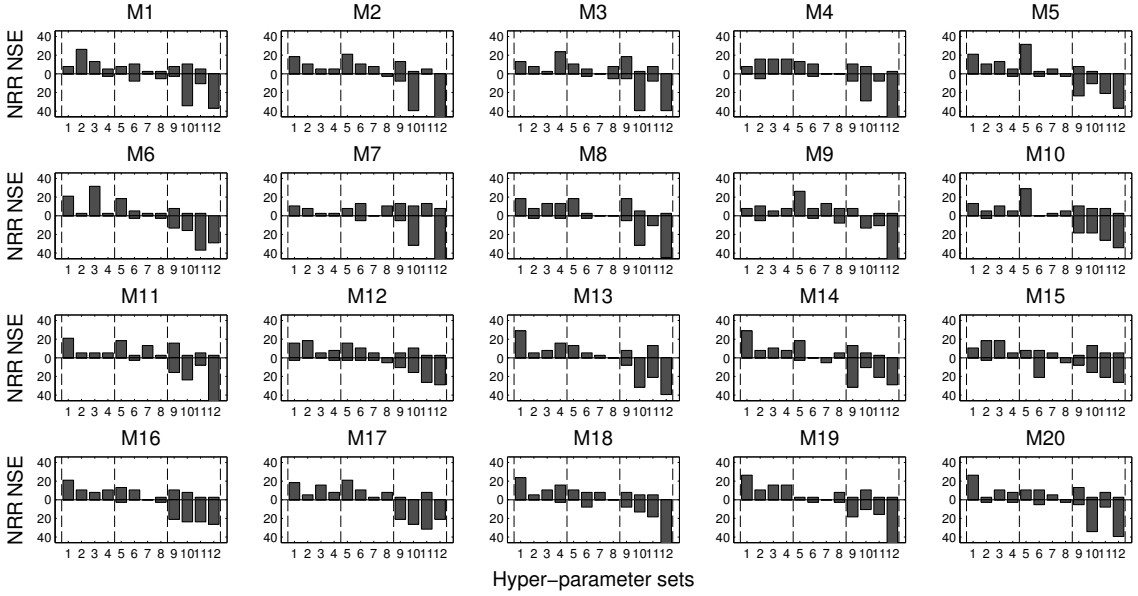

Figure 7: Frequency at which each of the 12 sets of hyper-parameters is better than others in term of accuracy (NSE) and reliability (NRR). Each boxplot corresponds to a model (see Table 1) and results are displayed for lead time 3

first forecasting day to achieve a reliable system and also implies that finding the best hyper-parameters only guarantees to find the optimal initialization without ensuring forecasting performances.

To investigate the possible relation between models and hyper-parameters, Figure 7 shows the frequency at which a hyper-parameter set outperforms the other for a given model. Each of the 20 sub-plot corresponds to a model. The 12 hyper-parameters are referred to by numbers as following:

1: P25%Q10%T2°C	2: P25%Q10%T5°C	3: P25%Q25%T2°C
4: P25%Q25%T5°C	5: P50%Q10%T2°C	6: P50%Q10%T5°C
7: P50%Q25%T2°C	8: P50%Q25%T5°C	9: P75%Q10%T2°C
10: P75%Q10%T5°C	11: P75%Q25%T2°C	12: P75%Q25%T5°C

The bars represent the frequency at which a hyper-parameter set outperforms the other. The upper part and lower part of the figure refer to the bias and reliability respectively. For instance, the hyper-parameter set number 10 is the best one for approximatively 10% of the catchments for the bias and 35% for reliability for the model 1.

The repartition of the best performing hyper-parameters confirms that no hyper-parameter set performs better than others systematically for the NSE or the NRR and exceeds rarely 40% for any model. Thus, to ensure to get optimal updating performance, an optimal hyper-parameter should be chosen according to model and catchment.

An additional difficulty arises from the fact that bias and reliability are optimized by different hyper-parameters. Optimal NSE values are often obtained by low to moderate noise magnitude while the best NRR are obtained with higher perturbations. This highlights the challenge to optimize bias and reliability collectively during EnKF updating, leaving to the modeller the burden of prioritizing one over the other.

3.3. *Influence of the choice of states variables*

For the the present section, results are shown for a particular hyper-parameters set P 50% Q 10% T 5% but similar conclusions could be drawn from the other tested sets.

This section addresses the question of identification of state variables that should be updated with the EnKF. For a N-state-variable-model, it exists $2^N - 1$ combinations and none is favored during testing. Thus, the number of possibilities depends on the model; see Table 1 for the number of states.

As the reservoirs are situated at different levels in the models (from interception to routing), their individual updating is expected to affect differently model outputs; more precisely they affect the time-lag between state perturbation and the change in simulated streamflow. Seo et al. (2003) suggest to not perturb reservoirs concerning soil moisture as it is a long-term component that has an influence which lasts much longer than the longest operationally used lead time. On the contrary, Wöhling et al. (2006) encourage soil moisture updating as it will act on all lead times. Physically based models offer the possibility to deal with values that are theoretically measurable. The knowledge about these values allows to estimate critical values that are the most subject to uncertainty. Conceptual models states values do not refer directly to a measurable value and the identification of variable states for updating is thus complex. The amount of uncertainty related to these variables is hardly definable and there is no apparent clue to update a

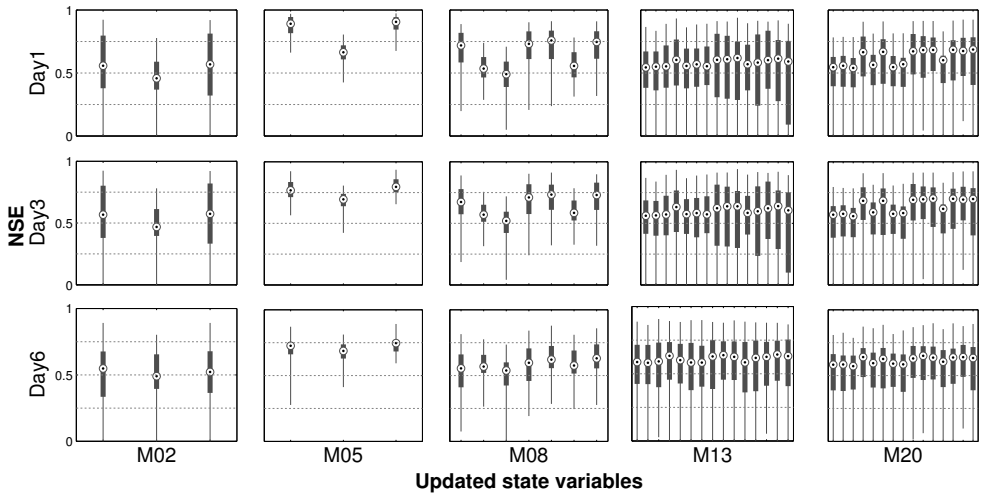

Figure 8: Distribution of NSE performance over the 38 catchments for model 2, 5, 8, 13, and 20 according to the updated states parameters for lead time 1, 3, and 6. The boxplot on the right of each sub-plot corresponds to the case where all states variable are updated.

certain reservoir and leaving others unperturbed.

Figure 8 presents the distribution of performance of every individual state combination per model, to illustrate the variability of NSE over the 38 catchments. Five models with different numbers of state variable are used to highlight the general 20 models behavior. Each box plot refers to a combination of updated state variable for a model. The box plot situated on the right of each sub-plot corresponds to the case where all model states are updated.

The main conclusion is twofold: the success of the updating procedure lays as much in the model as in the choice of the state variables. Most models best state combinations exhibit a median NSE higher than 0.75 for first lead time even if few models (model 2 in particular) seem to react poorly to the state updating. Best short-lead-time-model performance needs to be qualified as it frequently decreases with increasing lead time.

As median NSE values are frequently close to each other, it is possible to conclude that there is no obvious outperforming combination for any model – however, there are combinations that perform consistently poorly. Additional complexity in the choice of the best state combination to update arises from the performance variability over catchments for a specific updated set of states. As the median performance is close and the variability over catchments is important, it is very likely that one combination for a model on a catchment will be outperformed by another combination on another catchment.

As the state updating procedure is numerically implemented in the same way for all models, bad performance may be attributed to the suboptimal choice of updated model states or the potential inadequacy of the EnKF to a specific hydrologic model rather than the EnKF technique itself. On this subject, model number 2 open loop performances are often comparable with other models (see also Fig. 2) while its performance after updating are undoubtedly worse.

The question of best state set identification arises also as a function of the lead times (results not shown). In this study, we disregarded lead time specific states combinations since the use of different set of state combinations lead time dependant may improve performance for each lead time in average but it would imply to run in a parallel fashion several simulations for each lead time. An issue arising from such a technique would be the creation of discontinuities in the forecasting streamflows from one lead time to another.

Reservoirs which should be updated in priority are frequently –but not always– the closest reservoirs to the model outputs in the description of the rainfall-runoff process. The question of the number of reservoir to update is more complex as few global patterns emerge from the results. It is common practice to update all model states variable but this does not systematically lead to the best results (see also McMillan et al., 2013; Rakovec et al., 2015). In Fig 8, model 13 illustrates this since the updating of some state variable sub-ensembles shows improved performance for first, second and third quartiles. Therefore, for some models, optimal updating may be obtained by leaving some stores from the update, for instance the routing store for model 13. Generally, the number of updated states remain rather low, never exceeding 4 even for high dimensional states models (model 7, 12 and 14). All states should be updated for model 5, 6, 10 and 19 but other models states should be partially updated. Models with a large number of state variable (high degree of freedom) are more prone to encounter equifinality issues as many outcomes frequently end up close for a specific conditions. This lead to an already known problem that requires the user to take an arbitrary decision or possibly to retain several combination with the associated computational cost increase. Also, likewise for traditional model parameter estimation, the identification of best set depend on the score used as objective function. Selecting states

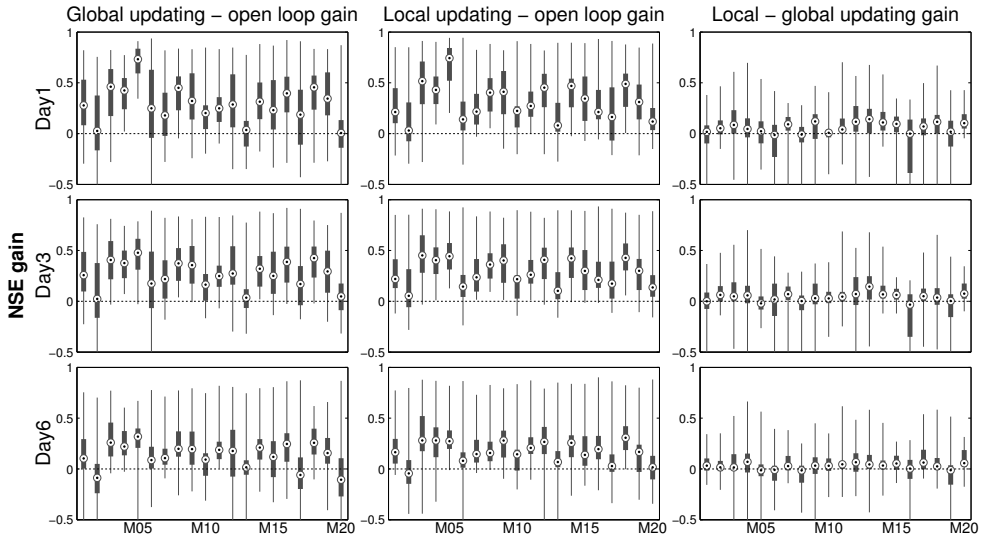

Figure 9: EnKF - open loop gains in NSE gains over the 38 catchments for global and locally defined hyper-parameters and state variable for lead time 1, 3 and 6.

set based on a NSE criterion does not guarantee to maximize other accuracy scores, and even less to achieve highest possible reliability. Thus, different sets of updated states may capture more or less accurately specificities of the hydrograph.

3.4. Global and local updating schemes

Setting EnKF catchment specificities is possible and may be operationally conceivable and worth considering. This case is more computationally demanding as states identification needs to be carried out for all watersheds. Thus, the gain of such approach needs to be quantified to justify the increase in commitment. In the opposite case, the forecaster takes a risk relying on optimal updated states set identified from only one catchment if this set is transferred onto another catchment.

Figure 9 displays gains obtained by EnKF over open loop. Two EnKF updating schemes are compared:

- A global scheme: updating is carried out with a single set of state variables and hyper-parameter per model, identified as best according to the combined criterion in average over all catchments. The updated states and the hyper-parameters are the same regardless of catchment.
- A local scheme : updating is carried out with a different set of state variables and set of hyper-parameters for each catchment identified as the best set of state variable per catchment. The approach is thus catchment specific.

The gain between the two updating schemes is also examined. In that case, the global performance is used as the reference in the gain equation (Eq. 11).

Overall, both EnKF schemes enhance open loop forecasts in the vast majority of cases, from short to longer lead time.

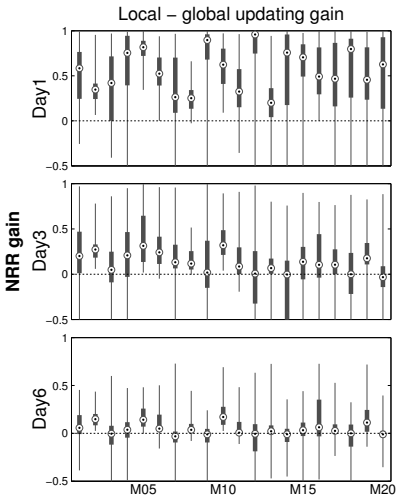

Figure 10: EnKF-open loop gains in *NRR* over the 38 catchments for global and locally defined hyper-parameters and state variable for lead time 1, 3 and 6

However, the gain in accuracy depends heavily on model and to a lesser extent on the global-local updating scheme. One can notice that models 2, 13 and 20 have a structure that react poorly to EnKF updating, especially for global states updating. The increase of computational resources may not be worth the potential gain in performance for the majority of catchments. Yet, these results are improved in the case where catchment specific state variable sets are used.

It is frequent –and normal– that the differences between the two updating schemes global/local for the same model are small. This is the case when a model has frequently the same best set of state variable over the 38 catchments which therefore turns to be the best in average over catchments and explains the frequent small dispersion of the local/global updating gain. However extremes are high as they are obtained

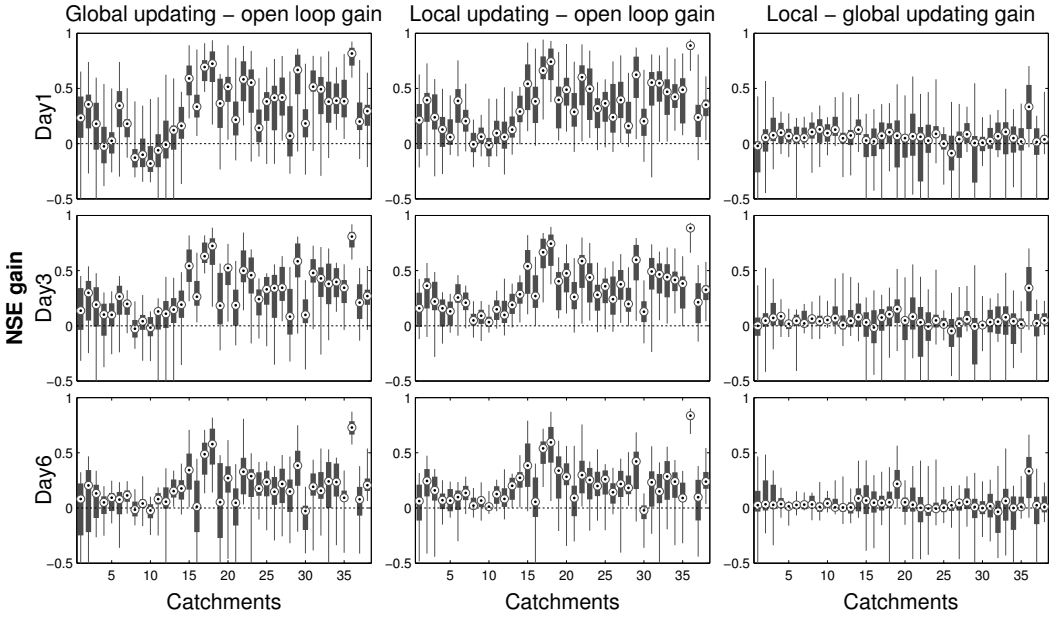

Figure 11: EnKF - open loop gains in NSE over the 20 models for global and locally defined hyper-parameters and state variable for lead time 1, 3 and 6 according to catchments

when the global updating fails largely on a catchment. The local scheme is logically better than global as it is designed to perform on all catchments but this does not ensure to be better than the global scheme in all cases. Indeed, even if the local updating is catchment specific, it is still averaged on lead times and thus the global state variable set may perform better for a particular lead time.

Figure 10 represents the gain between the local and the global updating procedure in reliability (no comparison is possible with the open loop as it is a deterministic forecast). The gain in reliability is consistently high for the first lead time as the second quartile is always positive and third quantile higher than 0.8 for most of the models. Some models, as the model 9, 12, 14 should be preferentially updated in a local way because their gain is substantial (third quartile is greater than 0.95). Interestingly, these models are among the most complex ones in the model pool and seem to require a more detailed setting to exploit optimally the EnKF for the first lead time. As with the NSE, the NRR gain decreases with lead time but stays mostly positive up to day 6 (see also Fig. 6). The gain may be negative for the reasons aforementioned with the NSE.

3.5. Influence of the catchments

To assess the importance of the catchments on forecasting performance, Figure 11 represents the models' NSE over the 38 watersheds. This complementary vision of Figure 9 reveals that catchments also have an influence on simulations that is as important as hyper-parameters, model structure, and the state variable selection.

The majority of the catchment can benefit from EnKF updating, especially in the case where local updating is used. Yet, there is a disparity in the gain as few catchments display a negative median gain, namely catchments 4, 8, 9, 10, 11, and 12 for the first lead time and global updating and catchments 9 and 11 for local updating.

The gain diminishes with increasing lead time except for the catchments that exhibit a negative gain from day one. The underlying reason is that EnKF is not able to update correctly the state variables, attributing erroneous values to the state variable, combined with the fact that the updated state variables have a greater influence for short lead times.

EnKF performance and gain were compared to the available climatic data and catchments characteristics. Specifically, the average annual total and liquid precipitation, the area and the estimated concentration time were put under scrutiny. No clear correlation between these values and EnKF performance has been identified.

4. Conclusion and recommendations

This paper discusses the performance and implementation of EnKF in forecasting over a wide variety of catchments and rainfall-runoff lumped models. An extensive testing was carried out to assess EnKF state updating and how it relates to model, catchment, and lead times.

The results show that an optimal implementation of the EnKF is more complex than frequently suggested and that a

detailed attention should be paid to the specification of hyper-parameters and updated state variable. While identification of the minimal number of members is relatively straightforward as a vast majority of models and catchment agree, there is no single and universal optimal EnKF implementation for any model. In practice, it is unlikely that the best state combination and hyper-parameter set in average are optimal for all watersheds. Unlike many case studies, it is not reasonable to recommend precise values, as the best EnKF settings are frequently case specific.

The hyper-parameters and more specifically, the perturbations of the inputs are frequently unintuitive to identify as there are often unrealistically high to implicitly account for other sources of uncertainty, especially parameter and structural uncertainty, and to eventually ensure model simulation reliability. An additional challenge arises from the difficulty to optimize reliability and ensemble median bias jointly as the improvement of one criterion is achieved at the expense of the other.

Models encounter important differences in their results and in the way they should be updated. Models with a high number of state variable (high degree of freedom) should receive an increased attention as they are more prone to encounter equifinality issues as many outcomes frequently end up close for a specific condition.

Regardless of the model, ensemble reliability decreases quickly with lead time as the ensemble spread drops from first days while the bias increases. This also underlines that taking into account explicitly initial condition uncertainty solely is not sufficient for medium range forecasting and that structural error and forcing error are dominant in modelling rainfall-runoff processes.

Despite these constrains, the gain that EnKF provides over open loop is substantial, especially if the optimization is carried out locally. The later implies a detailed testing of all combination to identify best performing EnKF implementation but is computationally more expensive. As the EnKF is not efficient with every model and catchment, we recommend to investigate data assimilation coupling with several models to go beyond EnKF - model structure compatibility issue.

Finally, we encourage EnKF users to perform a detailed analysis addressing the question of hyper-parameter and state variable selection of their system to ensure to make the most of EnKF. For further improvement, we also suggest to report explicitly the hyper-parameters and state variables they used to contribute to a better understanding of EnKF parametrization and to identify techniques that would allow to robustly identify the pertinent state variables that should be updated without the need to run all possible combinations.

5. Acknowledgements

The authors acknowledge the Centre d'Expertise Hydrique du Québec for providing hydrometeorological data. They also acknowledge financial support from the Chaire de recherche EDS en prévisions et actions hydrologiques and from the Natural Sciences and Engineering Research Council of Canada.

References

Abaza, M., Anctil, F., Fortin, V., Turcotte, R., 2014. Sequential streamflow assimilation for short-term hydrological ensemble forecasting. J Hydrol 519, 2692–2706.

Ajami, N. K., Duan, Q. Y., Sorooshian, Q., 2007. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour Res 43, 1–19.

Alvarez-Garreton, C., Ryu, D., Western, A. W., Crow, W. T.,Robertson, D. E., 2014. The impacts of assimilating satellite soil moisture into a rainfall-runoff model in a semi-arid catchment. J Hydrol 519, 2763–2774.

Andreadis, K.M., Lettenmaier, D.P., 2006. Assimilating remotely sensed snow observations into a macroscale hydrology model. Adv Water Resour 29, 872–886.

Bailey, R.T., Bau, D., 2012. Estimating geostatistical parameters and spatially-variable hydraulic conductivity within a catchment system using an ensemble smoother. Hydrol Earth Syst Sc 16, 287–304.

Bergström, S., Forsman, A., 1973. Development of a conceptual deterministic rainfall-runoff model. Nord Hydrol 4, 147–170.

Beven, K., Binley, A., 1992. The future of distributed models - model calibration and uncertainty prediction. Hydrol Process 6, 279–298.

Beven, K.J., Kirkby, M.J., Schofield, N., Tagg, A.F., 1984. Testing a physically-based flood forecasting model (TOPMODEL) for 3 uk catchments. J Hydrol 69, 119-143.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A generalized streamflow simulation system - Conceptual modelling for digital computers. US Department of Commerce, National Weather Service and State of California, United States, 204pp.

Chen, H., Yang, D.W., Hong, Y., Gourley, J.J., Zhang, Y., 2013. Hydrological data assimilation with the ensemble square-root-filter: Use of streamflow observations to update model states for real-time flash flood forecasting. Adv Water Resour 59, 209–220.

Chiew, F.H.S., Peel, M.C., Western, A.W., 2002. Application and testing of the simple rainfall-runoff model SIMHYD, in Mathematical Models of Small Watershed Hydrology and Applications. Water Resources Publication, Littleton, Colorado, United States, 335–367.

Clark, M.P., Rupp, D.E., Woods, R.A., Zheng, X., Ibbitt, R.P., Slater, A.G., Schmidt, J., Uddstrom, M.J., 2008. Hydrological data assimilation with the ensemble kalman filter: Use of streamflow observations to update states in a distributed hydrological model. Adv Water Resour 31, 1309–1324.

Cormary, Y., Guilbot, A., . Étude des relations pluie-débit sur trois bassins versants d'investigation, IAHS Madrid Symposium, IAHS Publication no. 108, 265–279.

DeChant, C.M., Moradkhani, H., 2011. Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. Hydrol Earth Syst Sc 15, 3399–3410.

DeChant, C.M., Moradkhani, H., 2012. Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. Water Resour Res 48, 1–15.

Duan, Q.Y., Ajami, N.K., Gao, X.G., Sorooshian, S., 2007. Multi-model ensemble hydrologic prediction using bayesian model averaging. Adv Water Resour 30, 1371–1386.

Duan, Q.Y., Sorooshian, S., Gupta, V., 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resour Res 28, 1015–1031.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte-carlo methods to forecast error statistics. J Geophys Res-Oceans 99, 10143–10162.

Evensen, G., 2003. The ensemble kalman filter: theoretical formulation and practical implementation. Ocean Dynam 53, 343–367.

Forman, B.A., Reichle, R.H., Rodell, M., 2012. Assimilation of terrestrial water storage from grace in a snow-dominated basin. Water Resour Res 48, 1–14.

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

Franz, K. J., Hogue, T. S., Bank, M., He, M. X., 2014. Assessment of SWE data assimilation for ensemble streamflow predictions. J Hydrol 519, 2737 – 2746.

Fortin, V., Turcotte, R., 2007. Le modèle hydrologique MOHYSE, Note de cours pour SCA7420. Report. Département des sciences de la terre et de l'atmosphère, Université du Québec à Montreal, Canada, 17pp.

Garçon, R., 1999. Modèle global pluie-débit pour la prévision et la prédétermination des crues. Houille Blanche 7/8, 88–95.

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

Harader, E., Borrell-Estupina, V., Ricci, S., Coustau, M., Thual, O., Piacentini, A., Bouvier, C., 2012. Correcting the radar rainfall forcing of a hydrological model with data assimilation: application to flood forecasting in the lez catchment in southern france. Hydrol Earth Syst Sc 16, 4247–4264.

Hong, Y., Hsu, K. L., Moradkhani, H., Sorooshian, S., 2006. Uncertainty quantification of satellite precipitation estimation and Monte Carlo assessment of the error propagation into hydrologic response. Water Resour Res 42, 1–15.

Jakeman, A.J., Littlewood, I.G., Whitehead, P.G., 1990. Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J Hydrol 117, 275–300.

Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., Sharma, A., 2011. Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive metropolis and sequential monte carlo samplers. Water Resour Res 47, 1–13.

Kim, J., Yoo, C., 2014. Use of a dual Kalman filter for real-time correction of mean field bias of radar rain rate. J Hydrol 519, 2785–2796.

Kuchment, L.S., Romanov, P., Gelfan, A.N., Demidov, V.N., 2010. Use of satellite-derived data for characterization of snow cover and simulation of snowmelt runoff through a distributed physically based model of runoff generation. Hydrol Earth Syst Sc 14, 339–350.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. J Hydrol 331, 161–177.

Lee, H., Seo, D.J., Koren, V., 2011. Assimilation of streamflow and in situ soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states. Adv Water Resour 34, 1597–1615.

Li, Y., Ryu, D., Western, A. W., Wang, Q. J., 2013. Assimilation of stream discharge for flood forecasting: The benefits of accounting for routing time lags. Water Resour Res 49, 1887 – 1900.

Li, Y., Ryu, D., Western, A. W., Wang, Q. J., Robertson, D. E., Crow, W. T., 2014. An integrated error parameter estimation and lag-aware data assimilation scheme for real-time flood forecasting. J Hydrol 519, 2722 – 2736.

Liu, Y. Q., Gupta, H. V., 2007. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. Water Resour Res 43, 1 – 18.

Liu, Y., Weerts, A. H., Clark, M., Franssen, H. J. H., Kumar, S., Moradkhani, H., Seo, D. J., Schwanenberg, D., Smith, P., Van Dijk, A., Van Velzen, N., He, M., Lee, H., Noh, S., Rakovec, O., Restrepo, P., 2012. Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. Hydrol Earth Syst Sc 16, 3863 – 3887.

Mandel, J., 2006. Efficient Implementation of the Ensemble Kalman Filter. Report. UCDHSC, Denver, Colorado, United States. 9pp.

Mazenc, B., Sanchez, M., Thiery, D., 1984. Analyse de l'influence de la physiographie d'un bassin versant sur les paramètres d'un modèle hydrologique global et sur les débits caractéristiques à l'exutoire. J Hydrol 69, 97–188.

McMillan, H.K., Hreinsson, E.O., Clark, M.P., Singh, S.K., Zammit, C., Uddstrom, M.J., 2013. Operational hydrological data assimilation with the recursive ensemble kalman filter. Hydrol Earth Syst Sc 17, 21–38.

Meier, P., Froemelt, A., Kinzelbach, W., 2011. Hydrological real-time modelling in the zambezi river basin using satellite-based soil moisture and rainfall data. Hydrol Earth Syst Sc 15, 999–1008.

Moore, R.J., Clarke, R.T., 1981. A distribution function approach to rainfall runoff modeling. Water Resour Res 17, 1367–1382.

Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state-parameter estimation of hydrological models using ensemble kalman filter. Adv Water Resour 28, 135–147.

Moradkhani, H.,Hsu, K., Hong, Y., Sorooshian, S., 2006. Investigating the impact of remotely sensed precipitation and hydrologic model uncertainties on the ensemble streamflow forecasting. Geophys Res Lett 33, 1–5.

Murphy, J.M., 1988. The impact of ensemble forecasts on predictability. Q J Roy Meteor Soc 114, 463–493.

Naevdal, G., Johnsen, L. M., Aanonsen, S. I., Vefring, E. H., 2003. Reservoir monitoring and continuous model updating using ensemble Kalman filter. SPE J 10, 66–74.

Nash, J. E., Sutcliffe, I., 1970. River flow forecasting through conceptual models. Part 1 - A discussion of principles. J Hydrol 10, 282–290.

Nie, S., Zhu, J., Luo, Y., 2011. Simultaneous estimation of land surface scheme states and parameters using the ensemble kalman filter: identical twin experiments. Hydrol Earth Syst Sc 15, 2437–2457.

Nielsen, S.A., Hansen, E., 1973. Numerical simulation of the rainfall-runoff process on a daily basis. Nord Hydrol 4, 171–190.

O'Connell, P.E., Nash, J.E., Farrell, J.P., 1970. River flow forecasting through conceptual models, part ii - the brosna catchment at ferbane. J Hydrol 10, 317–329.

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. J Hydrol 303, 290–306.

Parrish, M. A., Moradkhani, H., Dechant, C., 2012. Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation. Water Resour Res 48, 1–18.

Perrin, C., 2000. Vers une amélioration d'un modèle global pluie-débit. Thesis. Institut National Polytechnique de Grenoble - INPG, France. 289pp.

Perrin, C., Michel, C., Andreassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J Hydrol 279, 275–289.

Rakovec, O., Weerts, A.H., Hazenberg, P., Torfs, P., Uijlenhoet, R., 2012. State updating of a distributed hydrological model with ensemble kalman filtering: effects of updating frequency and observation network density on forecast accuracy. Hydrol Earth Syst Sc 16, 3435–3449.

Rakovec, O., Weerts, A.H., Sumihar, J., Uijlenhoet, R., 2015. Operational aspects of asynchronous filtering for flood forecasting. Hydrol Earth Syst Sc 19, 2911–2924.

Ramos, M.H., Mathevet, T., Thielen, J., Pappenberger, F., 2010. Communicating uncertainty in hydro-meteorological forecasts: mission impossible? Meteorol Appl 17, 223–235.

Reichle, R. H., Walker, J. P., Koster, R. D., Houser, P. R., 2002. Extended versus ensemble Kalman filtering for land data assimilation. J Hydrometeorol 3, 728–740.

Renzullo, L. J., Van Dijk, A., Perraud, J. M., Collins, D., Henderson, B., Jin, H., Smith, A. B., McJannet, D. L., 2014. Continental satellite soil moisture data assimilation improves root-zone moisture analysis for water resources assessment. J Hydrol 519, 2747–2762.

Salamon, P., Feyen, L., 2010. Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation. Water Resour Res 46, 1–20.

Seiller, G., Anctil, F., Perrin, C., 2012. Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. Hydrol Earth Syst Sc 16, 1171–1189.

Seo, D.J., Cajina, L., Corby, R., Howieson, T., 2009. Automatic state updating for operational streamflow forecasting via variational data assimilation. J Hydrol 367, 255–275.

Seo, D.J., Koren, V., Cajina, N., 2003. Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting. J Hydrometeorol 4, 627–641.

Sugawara, M., 1979. Automatic calibration of the tank model. Hydrolog Sci J 24, 375–388.

Thiery, D., 1982. Utilisation d'un modèle global pour identifier sur un niveau piézométrique des influences multiples dues à diverses activités humaines. IAHS-AISH Publication 136, 71–77.

Thirel, G., Martin, E., Mahfouf, J.F., Massart, S., Ricci, S., Regimbeau, F., Habets, F., 2010. A past discharge assimilation system for ensemble streamflow forecasts over france - part 2: Impact on the ensemble streamflow forecasts. Hydrol Earth Syst Sc 14, 1639–1653.

Thornthwaite, C.W., Mather, J.R., 1955. The water balance. Report. Drexel Institute of Climatology, Centerton, New Jersey, United States, 1–104.

Valery, A., Andreassian, V., Perrin, C., 2014. 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine?

part 1 - comparison of six snow accounting routines on 380 catchments. J Hydrol 517, 1166–1175.

Velázquez, J.A., Anctil, F., Ramos, M.H., Perrin, C., 2011. Can a multi-model approach improve hydrological ensemble forecasting? A study on 29 French catchments using 16 hydrological model structures. Adv Geosci 29, 33–42.

Vrugt, J.A., Diks, C.G.H., Gupta, H.V., Bouten, W., Verstraten, J.M., 2005. Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resour Res 41, 1–17.

Vrugt, J.A., Gupta, H.V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour Res 39, 1–14.

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and bayesian model averaging. Water Resour Res 43, 1–15.

Vrugt, J.A., Gupta, H.V., Nuallain, B. O., 2006. Real-time data assimilation for operational ensemble streamflow forecasting. J Hydrometeorol 7, 548–565.

Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol Earth Syst Sc 5, 13–26.

Warmerdam, P.M., Kole, J., Chormanski, J., 1997. Modelling rainfall-runoff processes in the hupselse beek research basin, in: IHP-V, Technical Documents in Hydrology, 155–160.

Weerts, A.H., El Serafy, G.Y.H., 2006. Particle filtering and ensemble kalman filtering for state updating with hydrological conceptual rainfall-runoff models. Water Resources Research 42, 17.

Whitaker, J.S., Hamill, T.M., 2002. Ensemble data assimilation without perturbed observations. Mon Weather Rev 130, 1913–1924.

Wöhling, T., Lennartz, F., Zappa, M., 2006. Technical note: Updating procedure for flood forecasting with conceptual hbv-type models. Hydrol Earth Syst Sc 10, 783–788.

Zhao, R.J., Zuang, Y.L., Fang, L.R., Liu, X.R., Zhang, Q., 1980. The xinanjiang model. IAHS-AISH Publication 129, 351–356.