Reply referee #1 on "Exploring hydrologic post-processing of ensemble streamflow forecasts based on Affine kernel dressing and Nondominated sorting genetic algorithm II"

Jing Xu[1], François Anctil[1], and Marie-Amélie Boucher[2]

[1]Department of Civil and Water Engineering, Université Laval, 1065 avenue de la Médecine, Québec, Québec, Canada;
[2]Department of Civil and Building Engineering, Université de Sherbrooke, 2500 Boul. de l'Université, Sherbrooke, Québec, Canada

Correspondence: Jing Xu (jing.xu.1@ulaval.ca)

Dear Prof. Solomatine and reviewers:

Thank you for your review comments that we received with respect to our paper. Those valuable comments have significantly enhanced our paper. We have carefully considered and addressed the reviewers' comments and suggestions, which led to significant revisions in many parts of the manuscript. Below we provide a point by point responses to each of the reviewer's comments. Please note that all the line numbers mentioned in our responses refer to the track-changes version of the manuscript.

1 Summary:

This study aims at improving the quality of probabilistic forecasts and facilitating the uncertainty communication by using two post-processing approaches within a hydrologic ensemble prediction system framework. The methods are clearly described and I find the results convincing. I think the article is ready for publication after some minor corrections.

2 Minor comments:

Specific question 1 : L53: It might be useful to give a short description of what "post-processing" means, something like "a correction of predictive distributions based on a comparison with past observations".

Response :

Thank you for your suggestion. We added a short description about "post-processing" here for better clarification.

"Line 54: By correcting the bias and adjusting the dispersion based on the comparison with past observations, statistical post-processing generally leads to a more accurate and reliable hydrologic ensemble forecast."

Specific question 2 : L68: A reference to "Another IMPREX project" is needed.

Response :

Thank you. We left this sentence out as the other reviewer suggested.

If you may be interested, the reference for this project is: *Cassagnole, M., Ramos, M.H., Zalachori, I, Thirel, G., Garçon, R., Gailhard, J., and Ouillon, T. : Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs−a conceptual approach, Hydrol. Earth Syst. Sci., 25(2), 1033−1052, https : //doi.org/10.5194/hess − 25 − 1033 − 2021 2021.*

Specific question 3 : L120: It might be useful to add the location of the gauging stations.

Response :

Thank you. The daily streamflow (m^3/s) time series entering the reservoirs were constructed by the electricity producer using a water balance equation for reservoirs and the turbine flow for run-of-the-river dams (identified as red thunder marks in Figure 1) and made available to the study along with spatially averaged minimum and maximum air temperature ($°C$) and precipitation (mm) for each sub-basin.

Specific question 4 : L148: "Five random hydrologic models from HOOPLA are exploited in this study.": what do you mean by "random" ? I find it very surprising that no other reason explains the choice of using those models.

Response :

Thank you for your question. For simplicity, we picked 5 models that are representatives among the 20 lumped models that were available to the study. The 5 representative models exploited here were selected from HydrOlOgical Prediction Laboratory (HOOPLA; Thiboult et al. (2020)) as typical examples. HOOPLA was able to provide a modular framework to perform calibration, simulation, and streamflow prediction using multiple hydrologic models (up to 20 models) (Perrin, 2000; Seiller et al., 2012). We rephrased the relative description in the manuscript for better clarification.

"Line 133: The HydrOlOgical Prediction Laboratory (HOOPLA; Thiboult et al. (2020)) provides a modular framework to perform calibration, simulation, and streamflow prediction using multiple hydrologic models (up to 20 lumped models) (Perrin, 2000; Seiller et al., 2012). The empirical two-parameter model CemaNeige (Valéry et al., 2014) simulates snow accumulation and melt. ***In this study, five representative models were selected from HOOPLA as typical examples.****Their main characteristics are summarized in Table 2."*

Specific question 5 : L225: It find it difficult to understand how the *NSE* is computed: with probabilistic streamflow forecast or with "average forecast and the average observation"? In addition, what do you mean by: "observed ones"? Did you use the same uncertainty estimation for observed values as in the *EnKF* (L160)? If not, please explain why.

Response :

Thank you. We rephrased this sentence using only "observations". We also added one more equation to show how we calculated the NSE score for better clarification. The x_t and y_t in the equation below represent the forecasted and observed values at time step t, respectively. We averaged the ensemble members for each time step.

"Line 220: Since here we are focused on probabilistic streamflow forecast, the accuracy could be measured by computing the distances between the forecast densities with the ***observations*** *(Wilks, 2011). Usually, hydrologists could rely on the Nash-Sutcliffe efficiency criterion (NSE, Nash and Sutcliffe (1970)) for measuring how well forecasts can reproduce the observed time series. Transforming the time series beforehand allows specializing it (i.e., NSE_{inv}, NSE_{sqrt}) for specific needs (e.g., Seiller et al., 2017).* ***NSE is attained by dividing the Mean square error (MSE) by the variance of the observations and then subtracting that ratio from 1.***

$$NES = 1 - \frac{MSE}{var(y)} = 1 - \frac{\sum_{t=1}^{T}(x_t - y_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{1}$$

where x_t and y_t are the forecasted and observed values at time step t, respectively. $\bar{y}$ and $var(y)$ represent the mean and variance of the observations. A perfect model forecast output would have an NSE value that equals to one."

No. The $EnKF$ uncertainty estimation is not our direct research subject. The $EnKF$ hyperparameters selection follows the work of Thiboult and Anctil (2015). Streamflow and precipitation uncertainties are assumed proportional; they are set to 10% and 50%, respectively. Temperature uncertainty is considered constant that it amounts to $2°C$. A Gaussian describes the streamflow and temperature uncertainty and a gamma law represents the precipitation uncertainty.

$Specific\ question\ 6$: L390: I find it a bit difficult to understand the link between the post-processing settings and the impact of some probabilistic scores. In particular, reliability is clearly improved, at the cost of dispersion, but I am not sure to understand how and why the weights obtained with NSGA-II can lead to those results. I would find it useful that you discuss this point in more depth.

$Response$:

Thank you for your question. We tested the post-processing performance of evolutionary multi-objective optimization (with NSGA-II) in this study. After the whole multiobjective evolutionary search, the un-repeated nondomination Pareto solutions can be identified. The solutions in the elbow region of the Pareto front are the compromise between both two objective functions. For example, in the study, the optimal NSE is inevitably accompanied with the highest $bias$ (e.g., NES= 0.846, $bias$=0.034), or vice versa. The solutions can be obtained daily by setting the sliding window. Specifically speaking, the NSGA-II post-processors were trained using only the past 30 days and day-4 forecast data and then re-trained every next day. Then the weight matrix can also be extracted from the solutions daily.

Hence, if y is assumed here as the target variable to streamflow forecast and Y_{obs}^{t}=[y^1, y^2,..., y^T] groups the time variant observations over the training period t = 1, ..., T. Also, let X_i^t=[x_1^t,x_2^t,...,x_K^t] represents K ensemble forecast members issued from the single-model H-EPSs daily. The weights w_k reflect how well the k^{th} prediction fits the training data at each time step. By assigning the weights to each candidate ensemble member, the bias and dispersion could be adjusted based on the comparison with past observations. This leads to a more reliable and skillful hydrologic ensemble forecast. For the detailed calculation steps, please refer to section 3.2 (Nondominated sorting genetic algorithm II) and section 3.4 (Experimental setup).

$Specific\ question\ 7$: L413: While I find the study convincing, I would appreciate a bit more of discussion, especially regarding the limits of the H-EPS and the various ways to improve them. It might be surprising that a H-EPS based on weather forecasts and data assimilation (EnKF) still needs some post-processing methods to be reliable. It might indicate that more work is needed to more appropriately define the data assimilation settings and inflate the ensemble within the data assimilation step. Or that a multimodel approach is required ?

$Response$:

Thank you for your comment. Yes, a comprehensive uncertainty analysis will be needed to track all sources of uncertainties in the hydro-meteorological forecasting chain. While, in practice, the forecasting performance of data assimilation fades away quickly as the lead time progresses. In addition, operational forecasts users may not be able to perfectly utilize all the forecasting tools (i.e., meteorological ensemble forcing, data assimilation, and multimodel) jointly. For the manuscript discussed here, we would like mainly focus on the single-model H-EPSs and did the uncertainty analysis for each model individually.

As for the other strategy of multimodel approach, We actually have another article focused on exploring "the hydrological post-processing of streamflow forecasts issued from multimodel ensemble prediction systems" published: Xu, $J.$, $Anctil$, $F.$ and $Boucher$, $M.A.$, 2019 $Hydrological\ post-processing\ of\ streamflow\ forecasts\ issued\ from\ multimodel\ ensemble\ prediction\ systems$, $J.\ Hydrol.$, 578, $p.$124002, $https://doi.org/10.1016/j.jhydrol.2019.124002$. In this above-mentioned paper, we took all sources of uncertainties into account since we tested on the grand multi-model ensemble forecast.

References

Movahedinia, F.: Assessing hydro-climatic uncertainties on hydropower generation, Université Laval, Québec city, 7 pp, https://corpus.ulaval.ca/jspui/handle/20.500.11794/25294, 2014.

Nash, J.E. and Sutcliffe, I.: River flow forecasting through conceptual models. Part 1-A discussion of principles. J. Hydrol., 10(3), 282-290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Perrin, C.: Vers une amélioration d'un modèle global pluie-débit, PhD diss., Institut National Polytechnique de Grenoble-INPG, Grenoble, 287 pp, 2000.

Seiller, G., Roy, R., and Anctil, F.: Influence of three common calibration metrics on the diagnosis of climate change impacts on water resources, J. Hydrol., 547, 280-295, https://doi.org/10.1016/j.jhydrol.2017.02.004, 2017.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth. Syst. Sc., 16, 1171-1189, https://doi.org/DOI : 10.5194/hess-1116-1171-2012, 2012.

Thiboult, A., Seiller, G., Poncelet, C., and Anctil, F.: The HOOPLA toolbox: a HydrOlOgical Prediction LAboratory to explore ensemble rainfall-runoff modeling, arXiv [preprint], Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-6, 28 January 2020.

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

Valéry, A., Andréassian, V., and Perrin, C.: As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 - Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, J. Hydrol., 517: 1176-1187, https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

Wilks, D.S.: On the Reliability of the Rank Histogram, Mon. Weather. Rev., 139, 311-316, https://doi.org/10.1175/2010MWR3446.1, 2011.

Reply referee #2 on "Exploring hydrologic post-processing of ensemble streamflow forecasts based on Affine kernel dressing and Nondominated sorting genetic algorithm II"

Jing Xu[1], François Anctil[1], and Marie-Amélie Boucher[2]

[1]Department of Civil and Water Engineering, Université Laval, 1065 avenue de la Médecine, Québec, Québec, Canada;
[2]Department of Civil and Building Engineering, Université de Sherbrooke, 2500 Boul. de l'Université, Sherbrooke, Québec, Canada

Correspondence: Jing Xu (jing.xu.1@ulaval.ca)

Dear Prof. Solomatine and reviewers:

Thank you for your review comments that we received with respect to our paper. Those valuable comments have significantly enhanced our paper. We have carefully considered and addressed the reviewers' comments and suggestions, which led to significant revisions in many parts of the manuscript. Below we provide our point by point responses to each of the reviewer's comments. Please note that all the line numbers mentioned in our responses refer to the track-changes version of the manuscript.

1 General Comments:

General question 1 : The body of the experiments and results is in my understanding a comparison of performance of two post-processing methods AKD and NSGA-II multi-objective optimisation. In the Introduction and Conclusion sections now a reflection has been added on user requirements of ensemble forecasts and benefits of being able to communicate trade-offs in which skill aspect(s) to increase with a hydrologic post-processor. The multi-objective optimisation has that benefit through presenting of pareto fronts and discussing these with end users. The current write-up of the Introduction and Conclusions, however, leaves the reflection on end user perspective still quite disconnected from the experimental set-up and results (and the reflection on end user perspective is not mentioned at all in the Abstract). A suggestion, for consideration, to further clarify the connection between objective and experimental set-up and results is to take as objective something like: "In this paper the performance of evolutionary multi-objective optimisation (with NSGA-II) as hydrological ensemble post-processor is tested and compared with a conventional state-of-the-art post-processor, AKD, because of the benefit of NSGA-II method in communicating trade-offs with end users on which performance aspects to improve." This is rather a long sentence and can be split-up in several sentences if that makes the message more clear. In the Conclusion sections the authors can then reflect on this objective, and objective and conclusions can be taken up in the abstract as well.

Response :

Thank you for your suggestions. We added more clarification in both ***Abstract*** and ***Conclusion*** sections.

"Line 1: Forecast uncertainties are unfortunately inevitable when conducting the deterministic analysis of a dynamical system. The cascade of uncertainty originates from different components of the forecasting chain, such as the chaotic nature of the atmosphere, various initial conditions and boundaries, inappropriate conceptual hydrologic modeling, and the inconsistent stationarity assumption in a changing environment. Ensemble forecasting proves to be a powerful tool to represent error growth in the dynamical system and to capture the uncertainties associated with different sources. ***In practice, the proper interpretation of the predictive uncertainties and model outputs will also have a crucial impact on risk-based decision. In this study, the performance of evolutionary multi-objective optimization (i.e., Non-dominated sorting genetic algorithm II, NSGA-II) as a hydrological ensemble post-processor was tested and compared with a conventional state-of-the-art post-processor, the Affine kernel dressing (AKD). Those two methods are theoretically/technically distinct, yet however share the same feature that both of them relax the parametric assumption of the underlying distribution of the data (the streamflow ensemble forecast). Both NSGA-II and AKD post-processors showed efficiency and effectiveness in eliminating forecast biases and maintaining a proper dispersion with increasing forecasting horizons. In addition, NSGA-II method demonstrated superiority in communicating trade-offs with end-users on which performance aspects to improve."***

"Line 422: In this paper, the performance of NSGA-II method is compared with a conventional post-processing method, the AKD. NSGA-II demonstrated its superior ability in improving the forecast performance as well as communicating trade-offs with end-users on which performance aspects to improve most. As selected objective functions here, neither NSE ***nor*** $bias$ ***could be improved more without negatively impacting the other. The use of NSGA-II opens up opportunities to enhance the forecast quality in line with the specific needs of end-users, since it allows for setting multiple specific objective functions from scratch. This flexibility should be considered as a key element for facilitating the implementation of H-EPSs in real-time operational forecasting."***

$General\ question\ 2$: The choice to train the post-processors on day-4 lead time and test/validate on days 1-3 and 5-7 has been clearly described, and the option to train for each lead time separately instead has been mentioned. The main point of discussion, however, is that by the approach taken, the observational data set for training and validation is the same. Question to be discussed in the paper by the authors is whether and how a validation of the methods to a period unseen would affect the results. Unless there is a miss-understanding on my side, because I do not quite understand the 30-day moving window mentioned as training period in lines 320-321. In other sections years 1985-2017 are referred to as the focus of this study (e.g. line 132), and 2011-2016 as "committed to forecasting" (line 150). It could be that the longer time period was used in earlier studies for calibrating and validating the hydrological models used, and then perhaps the years 2011-2016 have been used to emulate re-forecasting using the trained post-processing methods, and then perhaps the training of the post-processor is done only on the past 30-days (emulating an operational setting over the period 2011-2016 and re-training every next day within that period). After which the performance metrics reported in figures 7 onwards report the performance over 2011-2016. While these elements are all stated somewhere in the paper, I would request the authors to further clarify. This can be at

the lines referred to in my comment here, and in addition perhaps by adding analysis periods to the captions of the result figures.

Response :

Thank you for your comments and suggestions. After the whole multiobjective evolutionary search in the NSGA-II method, the un-repeated nondomination Pareto solutions can be identified. The solutions can be obtained daily by setting the 30-day sliding window. Specifically speaking, the NSGA-II post-processors were trained using only the past 30 days and day-4 forecast data and then re-trained every next day.

In addtion, we reformulated the relevant description of the dataset we used. The analysis periods were also added to each corresponding figure captions for better clarification.

"Line 133: The HydrOlOgical Prediction Laboratory (HOOPLA; Thiboult et al. (2020)) provides a modular framework to perform calibration, simulation, and streamflow prediction using multiple hydrologic models (up to 20 lumped models) (Perrin, 2000; Seiller et al., 2012). The empirical two-parameter model CemaNeige (Valéry et al., 2014) simulates snow accumulation and melt. In this study, five representative models were selected from HOOPLA as typical examples. Their main characteristics are summarized in Table 2.

The original observational time series extend from January 1950 to December 2017. While in terms of the input of HOOPLA, the observational period was limited to 33 years (1985-2017) to avoid the increased bias and variability caused by missing values within the record. The meteorological ensemble forecasts were retrieved from the European Center for Medium-Range Weather Forecasts (ECMWF; Fraley et al. (2010)). The time series extend from January 2011 to December 2016. The meteorological ensemble forecast used the reduced Gaussian transformation to the latitude-longitude system during the THORPEX Interactive Grand Global Ensemble (TIGGE) database retrieving by bilinear interpolation (e.g., Gaborit et al., 2013). The horizontal resolution was downscaled during retrieval from the 0.5° ECMWF grid resolution to a 0.1° grid resolution. This study resorts to the 12:00 UTC forecasts only, aggregated to a daily time step over a 7-day horizon. All data are aggregated at the catchment scale, averaging grid points located within each sub-catchments. All time series were split in two following the Split-Sample Test (SST) procedure of Klemeš (1986): 1986-2006 for calibration and 2013-2017 for validation. In both cases, three prior years were used for spin-up. January 2011-December 2016 is committed to hydrologic forecasting."

General question 3 : If there is any overlap between training/calibration and validation period in hydrological model development and/or re-forecast analysis and/or post-processor training, the potential impact of that overlap on the results and interpretation should be discussed.

Response :

Thank you for your comment. No, there is no overlap between training and validation period. The dataset used here can be considered having two components: the observations/forecasts that last from January 2011 to December 2016, and

the target ensemble for interpretation with a forecasting horizon that extends from day 1 to 7. In this study, a common calibration/validation procedure was conducted on the second component of the dataset. We added more clarification in section 3.4 Experimental setup as:

"Line 308: (1) Determine the training period. Subject to the dataset used in this study, it can be considered having two components: the observations/forecasts that last from January 2011 to December 2016, and the target ensemble for interpretation with a forecasting horizon that extends from day 1 to 7. Here, a common calibration/validation procedure was conducted on the second component of the dataset. We conducted the calibration on day-4 forecast and then tested it on other lead times to assess the robustness of the predictive models. The skill of hydrologic forecasts fades away with increasing lead time. The 4-day-ahead ensemble forecasts issued from each single-model H-EPSs and their corresponding observations are chosen as a training dataset, since located in the middle of the forecast horizon. The validation dataset then consists of the remaining forecasts: day 1-3 and 5-7 ahead raw forecasts issued from the associated H-EPSs. The procedure was selected as a specific example. *Yet one may decide otherwise, such as implementing the calibration/validation procedures separately for each day."*

2 Detailed comments and editorials:

Detailed comments 1 : l26-27: suggest to leave out colloquial sentence about insufficient single deterministic forecast.

Response :

Thank you for your comments. We removed this sentence.

Detailed comments 2 : l34: delete "national" (because ECMWF is not a national organisation)

Response :

Thank you. We deleted it.

Detailed comments 3 : l46-47: Edit, sentence is not correct.

Response :

Thank you. We rephrased the sentence.

"Line 48: the statistical hydrologic post-processors, which have been added in the H-EPS for rectifying biases and dispersion errors (i.e., too narrow/too large), are numerous as reviewed by Li et al. (2017)."

Detailed comments 4 : l57: Edit: "value of a hydrological forecasts"

Response :

Thank you. We edited this sentence.

"Line 59: Buizza et al. (2007) emphasized that both functional and technical qualities are supposed to be assessed for evaluating the overall forecast ***value of hydrological forecasts."***

Detailed comments 5 : Line 59: "She also demonstrated.." ? Colloquial. Provide specific name/reference.

Response :

Thank you. We added the reference.

"Line 61: Ramos et al. (2010) reported similar achievements from two studies obtained from a Member States workshop (Thielen et al., 2005) role-play game and another survey to explore the users' risk perception of forecasting uncertainties and how they dealt with uncertain forecasts for decision-making."

Detailed comments 6 : l64-69: this is an important paragraph leading-up to the objective of the paper, but it contains some unclear sentences. Please reformulate to clarify. (see suggestion in General Comment above)

Response :

Thank you. We reformulated this paragraph for better clarification.

"Line 71: Here, two hydrological post-processors, namely the Affine kernel dressing (AKD) and the evolutionary multi-objective optimization (Non-dominated sorting genetic algorithm II, NSGA-II), were explored. Compared to conventional post-processing method, such as AKD, NSGA-II opens up the opportunity of improving the forecast quality in harmony with the forecasting aims and the specific needs of end-users. Multiple objective functions (i.e., here, verifying scores) for evaluating the forecasting performances of the H-EPSs are selected to guide the optimization process."

Detailed comments 7 : l68: "..another IMPREX product conduct.." ? (not a correct sentence, but also IMPREX not introduced before I think. Leave out if possible)

Response :

Thank you for your suggestion. We left this sentence out.

Detailed comments 8 : l71: should be "This study is a contribution.."

Response :

Thank you. We deleted this sentence and reformulated this paragraph for better clarification.

"Line 71: Here, two hydrological post-processors, namely the Affine kernel dressing (AKD) and the evolutionary multi-objective optimization (Non-dominated sorting genetic algorithm II, NSGA-II), were explored. Compared to conventional post-processing method, such as AKD, NSGA-II opens up the opportunity of improving the forecast quality in harmony with the forecasting aims and the specific needs of end-users. Multiple objective functions (i.e., here, verifying scores) for evaluating the forecasting performances of the H-EPSs are selected to guide the optimization process."

Detailed comments 9 : l71: ".. to probe this topic.." Instead of "this topic" explicitly state the topic here. I assume it is the remaining challenge mentioned in l54-56: "how to improve the human interpretation of probabilistic forecasts and the communication of integrated ensemble forecast products to end-users (e.g., operational hydrologists, water managers, local conservation authorities, stakeholders and other relevant decision makers)." But then in l71 and further it should be explained how the testing of these two post-processing methods contributes to improving interpretation and communication. Or "this topic" refers to the paragraph l64-69, which needs reformulation to be more clear, but is directed towards harmonising forecast improvement and user-specific requirements and use in decision making, in which case l71 onwards also has to explain how the comparison of these two post-processing methods is contributing to this. In the present formulation of l71-75 it seems more as if the authors are addressing the benefits of distribution-free postprocessing methods. See my suggestion above under General Comment.

Response :

Thank you for your questions. Yes, this topic is referred to "how to improve the human interpretation of probabilistic forecasts and the communication of integrated ensemble forecast products to end-users (e.g., operational hydrologists, water managers, local conservation authorities, stakeholders and other relevant decision makers)." We reformulated the contents in the manuscript as:

"Line 71: Here, two hydrological post-processors, namely the Affine kernel dressing (AKD) and the evolutionary multi-objective optimization (Non-dominated sorting genetic algorithm II, NSGA-II), were explored. Compared to conventional post-processing method, such as AKD, NSGA-II opens up the opportunity of improving the forecast quality in harmony with the forecasting aims and the specific needs of end-users. Multiple objective functions (i.e., here, verifying scores) for evaluating the forecasting performances of the H-EPSs are selected to guide the optimization process. *The mechanisms of these two statistical post-processing methods are completely different. However, they share one similarity from another perspective, which is they can estimate the probability density directly from the data (i.e., ensemble forecast) without assuming any particular underlying distribution. As a more conventional method, Silverman (1986) firstly proposed the kernel density smoothing method to estimate the distribution from the data by centering a kernel function K that determines the shape of a probability distribution (i.e., kernel) fitted around every data point (i.e., ensemble members). The smooth kernel estimate is then the sum of those kernels. As for the choice of bandwidth h of each dressing kernel, Silverman's rule of thumb finds an optimal bandwidth h by assuming that the data*

is normally distributed. Improvements to the original idea were soon to follow. For instance, the improved Sheather Jones (ISJ) algorithm is more suitable and robust with respect to multimodality (Wand and Jones, 1994). Roulston and Smith (2003) rely on the series of "best forecasts" (i.e., best-member dressing) to compute the kernel bandwidth h. Wang and Bishop (2005) as well as Fortin et al. (2006) further improved the best member method. The later advocated that the more extreme ensemble members are more likely to be the best member of raw under-dispersive forecasts, while the central members tend to be more "precise" for over-dispersive ensemble. They proposed the idea that different predictive weights should be set over each ensemble member, given each member's rank within the ensemble. Instead of standard dressing kernels that act on individual ensemble members, Bröcker and Smith (2008) proposed the AKD method by assuming an affine mapping between ensemble members and observation over the entire ensemble. They approximate the distribution of the observation given the ensemble.

Given the single-model H-EPSs studied here, the hydrologic ensemble is generated by activating two forecasting tools: the ensemble weather forecasts and the EnKF. Henceforth, enhancing the H-EPS forecasting skill by assigning different credibility to ensemble members becomes preferred than reducing the number of members. The post-processing techniques, like the Non-dominated sorting genetic algorithm II (NSGA-II), are now common (e.g., Liong et al., 2001; De Vos and Rientjes, 2007; Confesor and Whittaker, 2007). Such techniques are conceptually linked to the multiobjective parameter calibration of hydrologic models using Pareto approaches. Indeed, formulating a model structure or representing the hydrologic processes using a unique global optimal parameter set proves to be very subjective. Multiple optimal parameter sets exist with satisfying behavior given the different conceptualizations, albeit not identical Beven and Binley (1992). For example, Brochero et al. (2013) utilized the Pareto fronts generated with NSGA-II for selecting the "best" ensemble from a hydrologic forecasting model with a pool of 800 streamflow predictors, in order to reduce the H-EPS complexity. ***Here, the expected output of NSGA-II method is a group of solutions, also known as Pareto front, that identify the trade-offs between different objectives, subject to the end-users' needs and requirements."***

Detailed comments 10 : l314: ".. for each day."

Response :

Thank you. We corrected this writing error

" Line 316: Yet one may decide otherwise, such as implementing the calibration/validation procedures separately ***for each day."***

Detailed comments 11 : l320-321: A training period of 30-days with moving window is mentioned here. Please kindly clarify, including whether that applies to re-training every next day both the AKD and NSGA-II post-processor, using only the past 30-days observational and forecast data (day-4 lead time).

Response :

Thank you for your suggestion. We added the clarification here:

"Line 323: Here a 30-day moving window is selected so it contains enough training samples with coherent consistency. ***Which is to say, the NSGA-II post-processors were trained using only the past 30 days and day-4 forecast data and then re-trained every next day.*** *Especially, from the operational perspective, a monthly moving window is more coherent and efficient in the real world, with limited length for time series."*

Detailed comments 12 : l354: I do not think the heading "Uncertainty analysis" covers what is presented and discussed here. I would suggest a separate heading for the NSGAII result, e.g. "4.2: NSGA-II convergence", and then for the comparison, e.g. "4.3 AKD and NSGA-II performance comparison".

Response :

Thank you for your suggestions. We added the separate headings.

"Line 353: 4.2 NSGA-II convergence"

"Line 365: 4.3 AKD and NSGA-II performance comparison"

Detailed comments 13 : l356: remove the space in "wi thout"

Response :

Thank you. We corrected this writing error.

Detailed comments 14 : l359-361: I do not understand what is done here and why, when referring to 'random selection from the pareto front'

Response :

Thank you for your question.

The un-repeated nondomination Pareto solutions is a set of optimal options for users to choose after the whole evolutionary multiobjective optimization search. The solutions in the elbow region of the Pareto front are the compromise between both two objective functions. For example, in the study, the optimal NSE is inevitably accompanied with the highest $bias$ (e.g., NES= 0.846, $bias$=0.034), or vice versa. As one representative multiobjective evolutionary search result shown in Figure 5, 35 (nondominated) Pareto solutions are identified. The solutions can be obtained daily by setting the sliding window. Specifically speaking, the NSGA-II post-processors were trained using only the past 30 days and day-4 forecast data and then re-trained every next day. Therefore, we decided to pick a random solution in the Pareto front at each time step since they were all optimal options.

Detailed comments 15 : l367-369: Remove. I suggest to start the new section 4.3 with "The reliability of the raw.."

Response :

Thank you. We removed these contents in the manuscript.

References

Beven, K. and Binley, A.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6(3), 279-298, https://doi.org/10.1002/hyp.3360060305, 1992.

Brochero, D., Gagné, C., and Anctil, F.: Evolutionary multiobjective optimization for selecting members of an ensemble streamflow forecasting model. Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference-GECCO, New York, United States, 6 July 2013, 13, 1221-1228, https://doi.org/10.1145/2463372.2463538, 2013.

Bröcker, J. and Smith, L.A.: From ensemble forecasts to predictive distribution functions, Tellus. A., 60(4), 663-678, https://doi.org/10.1111/j.1600-0870.2007.00333.x, 2008.

Buizza, R. Asensio, H. Balint, G. Bartholmes J, et al.: EURORISK/PREVIEW report on the technical quality, functional quality and forecast value of meteorological and hydrological forecasts, ECMWF Technical Memorandum, ECMWF Research Department: Shinfield Park, Reading, United Kingdom, 516, 1-21, http://www.ecmwf.int/publications/, 2007.

Crochemore, L., Ramos, M.H., Pappenberger, F., and Perrin, C.: Seasonal streamflow forecasting by conditioning climatology with precipitation indices, Hydrol.Earth Syst. Sci., 21, 1573-1591, https://doi.org/10.5194/hess-21-1573-2017, 2017.

Confesor, Jr.R.B. and Whittaker, G.W., 2007. Automatic Calibration of Hydrologic Models With Multi-Objective Evolutionary Algorithm and Pareto Optimization 1, JAWRA Journal of the American Water Resources Association, 43(4), 981-989, https://doi.org/10.1111/j.1752-1688.2007.00080.x, 2007.

De Vos, N.J. and Rientjes, T.H.M.: Multi-objective performance comparison of an artificial neural network and a conceptual rainfall—runoff model, Hydrolog. Sci. J., 52(3), 397-413, https://doi.org/10.1029/2007WR006734, 2007.

Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T.A.M.T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE. T. Evolut. Comput., 6(2), 182-197, https://doi.org/10.1109/4235.996017, 2002.

Fraley, C., Raftery, A.E., and Gneiting, T.: Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging, Mon. Weather. Rev., 138, 190-202, https://doi.org/10.1175/2009MWR3046.1, 2010.

Fortin, V., Favre, A.C., Saïd, M., 2006. Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member, Q. J. R. Meteorol Soc., 132, 1349-1369, https://doi.org/10.1256/qj.05.167, 2006.

Gaborit, É., Anctil, F., Fortin V., and Pelletier, G.: On the reliability of spatially disaggregated global ensemble rainfall forecasts, Hydrol. Process., 27(1), 45-56. https://doi.org/10.1002/hyp, 2013.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31(1), 13-24, https://doi.org/10.1080/02626668609491024, 1986.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z.: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting, Wires. Water., 4(6), e1246. https://doi.org/10.1002/wat2.1246, 2017.

Liong, S.Y., Khu, S.T. and Chan, W.T.: Derivation of Pareto front with genetic algorithm and neural network, J. Hydrol. Eng., 6(1), 52-61, https://doi.org/10.1061/(ASCE)1084-0699(2001)6:1(52), 2001.

Perrin, C.: Vers une amélioration d'un modèle global pluie-débit, PhD diss., Institut National Polytechnique de Grenoble-INPG, Grenoble, 287 pp, 2000.

Ramos, M.H., Mathevet, T., Thielen, J. and Pappenberger, F.: Communicating uncertainty in hydro-meteorological forecasts: mission impossible? Meteorol. Appl., 17(2), pp.223-235, https://doi.org/10.1002/met.202, 2010.

Roulston, M.S. and Smith, L.A.: Combining dynamical and statistical ensembles. Tellus. A., 55, 16-30, https://doi.org/10.3402/tellusa.v55i1.12082, 2003.

Seiller, G., Anctil, F., and Perrin, C.: Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions, Hydrol. Earth. Syst. Sc., 16, 1171-1189, https://doi.org/DOI : 10.5194/hess-1116-1171-2012, 2012.

Silverman, B.W.: Density estimation for statistics and data analysis, CRC press, London, 26, 1986.

Thiboult, A., Seiller, G., Poncelet, C., and Anctil, F.: The HOOPLA toolbox: a HydrOlOgical Prediction LAboratory to explore ensemble rainfall-runoff modeling, arXiv [preprint], Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-6, 28 January 2020.

Thielen, J., Ramos,M.H., Bartholmes, J., De Roo, A., Cloke, H., Pappenberger, F., and Demeritt, D.: Summary report of the 1st EFAS workshop on the use of Ensemble Prediction System in flood forecasting, European Report EUR, Ispra., 22118. http://floods.jrc.ec.europa.eu/efas-documents, 2005.

Valéry, A., Andréassian, V., and Perrin, C.: As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 - Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, J. Hydrol., 517: 1176-1187, https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

Wang, X. and Bishop, C.H.: Improvement of ensemble reliability with a new dressing kernel, 131(607), 965-986, https://doi.org/10.1256/qj.04.120, 2005.

Wand, M.P. and Jones, M.C.: Kernel smoothing, CRC press, vol. 60, 1 December 1994.