Water Resour Manage (2015) 29:3329–3342
DOI 10.1007/s11269-015-1001-3

Combining Stochastic Weather Generation and Ensemble Weather Forecasts for Short-Term Streamflow Prediction

Jie Chen[1] **· François P. Brissette**[1]

Received: 22 June 2014 / Accepted: 2 April 2015 /
Published online: 20 April 2015

Abstract Ensemble streamflow predictions (ESPs) offer great potential benefits for water resource management, as they contain key probabilistic information for analyzing prediction uncertainty. Ensemble weather forecasts (EWFs) are usually incorporated into ESPs to provide climate information. However, there is no simple way to combine both of them, since EWFs are generally biased and under-dispersed. This study presents a new short-term (1 to 7 lead days) probabilistic streamflow prediction system combining stochastic weather generation and EWFs. The bias and under-dispersion of EWFs were first corrected using a weather generator-based post-processing approach (GPP). The corrected weather forecasts were then coupled with a hydrological model for streamflow forecasts. The proposed GPP forecast was compared against two other forecasts, one using the raw EWF (GFS), and the other using a stochastic weather generator (WG). The comparison was carried out over two Quebec watersheds, using a set of deterministic and probabilistic verification metrics. The deterministic metrics showed that the GPP forecast is consistently the best at predicting the ensemble mean streamflow for both watersheds and for all the leads ranging between 1 and 7 days, followed by the WG forecast. The probabilistic metrics showed negative or near zero skill retained by the GFS forecast for the first 7 lead days. The WG system was much more skillful than the GFS. The GPP forecast consistently displayed the highest skill and reliability in terms of all the metrics applied. With increasing lead days, the skill and reliability of the GPP forecast tend to converge toward that of the WG forecast, indicating that the short-term GPP forecast could easily be linked to a pure WG forecast to extend the forecast horizon.

Keywords Ensemble streamflow prediction · Ensemble weather forecast · Stochastic weather generator · Hydrology

✉ Jie Chen
jie.chen@etsmtl.ca

[1] Department of Construction Engineering, École de technologie supérieure, Université du Québec, 1100 rue Notre-Dame Ouest, Montreal, QC H3C 1K3, Canada

1 Introduction

Water resource management relies extensively on hydrological modeling, which in turns depends on meteorological forecasts for flow predictions on a short-term basis. For medium to long-term forecasts, hydrology models generally rely on some form of resampling of the climatological record. In either case, hydrologists need to bridge the gap between the water sciences and the fields of meteorology and climatology. Recent initiatives such as the Hydrologic Ensemble Prediction EXperiment (HEPEX) program bring together hydrological and meteorological communities worldwide to build a research project focused on advancing probabilistic hydrological forecast techniques (Schaake et al. 2006).

Generally speaking, there are two types of hydrological forecasts: deterministic and probabilistic. In a deterministic forecast, the outcome is precisely determined, usually using a deterministic weather forecast, without any consideration for uncertainty in the forecast. Forecast uncertainty provides potentially important information that may be critical to the decision-making process. Probabilistic forecasts are designed to provide this critical information to decision makers (Cloke and Pappenberger 2009; Boucher et al. 2011). For example, Days (1985) presented a probabilistic streamflow forecast procedure using the National Weather Service River Forecast System for predicting streamflow. This system assumes that each year of the historical data is a possible representation of the future. One streamflow trace was simulated for each historical year using the current watershed conditions as the initial conditions for each simulation. Several studies (Velazquez et al. 2009; Boucher et al. 2011) have compared deterministic and probabilistic hydrological forecasts and consistently demonstrated that ensemble forecasts have more skill than their deterministic counterparts.

Numerical weather predictions are usually incorporated into ensemble streamflow prediction (ESP) to provide climate information (e.g., Clark and Hay 2004; Roulin and Vannitsem 2005; Ghile and Schulze 2010; Tang et al. 2010; Şensoy and Uysal 2012; Dutta et al. 2012). A hydrological forecasting system based on coupling the ensemble weather forecast (EWF) with hydrological models is able to capture the uncertainties associated with the weather forecast and thus to better predict river flows. For example, Roulin and Vannitsem (2005) showed that ESP systems based on ensemble precipitation forecasts were much better than those based on the historical precipitation input for two Belgian catchments. However, other studies have pointed out that a post-processing step is required before using EWFs with hydrological models. Clark and Hay (2004) assessed the possibility of using 8-day atmospheric forecasts from the National Centers for Environmental Prediction for streamflow prediction over the contiguous United States and found that the model output statistics (MOS)-based downscaling procedure increased the skill of streamflow predictions for snowmelt-dominated watersheds where daily variations in streamflow were strongly forced by temperature, while the skill of MOS-based forecasts in the rainfall-dominated basin were equivalent to that of the climatic ESP.

Short-term hydrological prediction is highly dependent on the quality of the meteorological output variable, especially the precipitation forecast (Roulin and Vannitsem 2005), since the hydrological prediction is usually based on daily meteorological inputs. However, EWFs generally suffer from bias and tend to underestimate the ensemble spread (Hamill and Colucci 1997; Eckel and Walters 1998), which limits their application to hydrological forecasts. The bias of EWFs can be introduced by systematic model errors, such as the imperfect model formulation and parameterization of some climate processes, as well as by the use of imperfect initial conditions, which can be amplified by the chaotic nature of the evolution equations of the dynamical system. To address the bias and under-dispersion of EWFs, a number of post-processing methods have been proposed. Wilks (2006) classified these methods into rank histogram techniques, ensemble dressing, Bayesian model averaging (BMA), logistic regression, analog techniques and nonhomogeneous Gaussian regression and compared their performances for post-processing EWFs.

All of these methods were developed to determine the underlying probabilistic distribution of the forecasted precipitation and temperature. However, for ESP, continuous time series over several days are needed to run hydrological models and it is not a simple task to go from daily probabilistic distributions to the coherent sequences of daily variables needed to run hydrology models. Chen et al. (2014a) presented a new method for post-processing EWFs using a stochastic weather generator. This method has the significant advantage of readily generating continuous time series that are fully consistent with the underlying distribution of EWFs.

However, one of the ultimate goals of using EWFs is for hydrological forecasting. The performance of any new post-processing method should be validated for hydrological forecasts at the watershed scale. Accordingly, this study evaluates the effectiveness of post-processing EWFs for short-term (1 to 7 lead day) ESPs. The EWFs referred to herein are the Global Forecast System (GFS) reforecast datasets. The 7-day EWFs are first subjected to post-processing using a weather generator-based approach (referred to as the GPP) (Chen et al. 2014a). The use of the 7th day as a limit of post-processing was based on the horizon for which the post-processed GFS reforecasts displayed skill levels above those of the resampling approaches for precipitation. In other words, past 7 days, the precipitation forecasts were no better than historical resampling results (Chen et al. 2014a). The GPP system is further compared with two other ESP systems over two Canadian watersheds in the province of Quebec. One directly links raw GFS EWFs to the hydrological model (referred to as GFS), while the other is achieved by resampling the historical meteorological record using a stochastic weather generator (referred to as WG).

2 Study Area and Dataset

2.1 Study Area

The proposed ESP system is evaluated over two Canadian catchments (Peribonka and Yamaska) located in the Province of Quebec (Fig. 1). These two catchments were also used by Chen et al. (2014a) for evaluating the effectiveness of post-processing EWFs. They are only briefly described as follows. More details can be found in Chen et al. (2014a). Both the Peribonka and Yamaska catchments are composed of several tributaries draining basins of approximately 27,000 and 4843 km^2, respectively. The southern parts of the Peribonka and Yamaska catchments, known as the Chute-du-Diable (CDD) and the Yamaska (YAM) watersheds, respectively, are used in this study. The CDD and YAM watersheds differ in terms of their size (9700 vs 3300 km^2) and location, thus allowing the evaluation of basin characteristics on ESP. The CDD watershed is a snow-dominated forested basin located in Central Quebec. Its hydrographs are strongly characterized by the snowmelt season. The smaller YAM watershed is located in southern Quebec (at the US border) in a mixed agricultural-forest setting.

2.2 Dataset

The dataset consists of both observed data and EWFs. The observed daily precipitation, maximum and minimum temperatures (Tmax and Tmin) over the two watersheds were taken

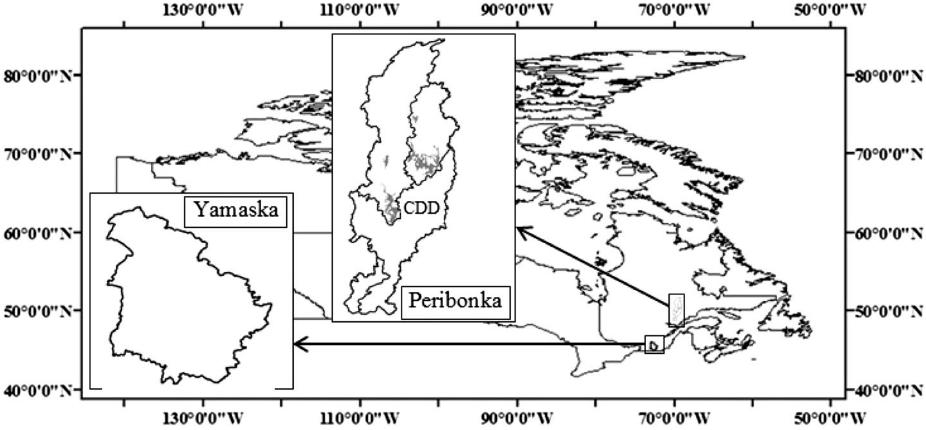

Fig. 1 Location map of the two catchments

from the National Land and Water Information Service (www.agr.gc.ca/nlwis-snite) dataset covering the period 1979–2003. This dataset was created by interpolating station data to a 10-km grid using a thin plate smoothing spline surface fitting method (Hutchinson et al. 2009). All the grid points within the watershed were averaged to obtain a single precipitation and temperature time series for each basin. The daily discharge at the watershed outlet was obtained from Environment Canada (http://www.ec.gc.ca/).

EWFs (daily total precipitation and mean temperature) with a global grid of 2.5° were taken from the GFS reforecast dataset (http://www.esrl.noaa.gov/psd/forecasts/reforecast/) (Hamill et al. 2006). Several previous studies have discussed the benefit of calibrating probabilistic forecasts using ensemble reforecast datasets (e.g., Hamill et al. 2004, 2006; Hamill and Whitaker 2006). Forecasts for each day since 1979 have been produced with GFS, comprised of a 15-member run out to 15 days. The grid boxes covering the watershed were selected; in this study, two grid boxes were selected and averaged for the CDD watershed and one grid box was selected for the YAM watershed.

3 Methodology

3.1 Stochastic Weather Generator

Stochastic weather generators are algorithms that produce climate time series of arbitrary lengths that possess statistical properties very similar to those of the input data. The Weather Generator of École de Technologie Supérieure (WeaGETS) (Chen et al. 2012a), which is a WGEN-like (Richardson 1981; Richardson and Wright 1984) three-variate (precipitation, Tmax and Tmin) single-site stochastic weather generator programmed in Matlab, was used in this study. In WeaGETS, the parameters are estimated biweekly.

The precipitation component of WeaGETS is a Markov chain for occurrence and a gamma distribution for quantity. A first-order two-state Markov chain is used to generate the precipitation occurrence. For a predicted rain day, a gamma distribution is used to generate daily precipitation amounts. The daily Tmax and Tmin are generated using a first-order linear autoregressive model. Readers are referred to Chen et al. (2012a) and Richardson (1981) for complete details. WeaGETS has been tested extensively at several locations under various climates and found to be adequate at reproducing precipitation and temperature characteristics (Chen et al. 2012a; Chen and Brissette 2014). The WeaGETS code is freely available on the Mathworks File exchange site.

Stochastic weather generation as used in this study is similar to resampling the historical database. However, it has several advantages over historical resampling. For example, weather generators can generate events that are outside the historical record, either by drawing large values from a properly fitted distribution or by allowing sequences of longer dry/wet series. Weather generators also allow for the use of an infinite number of time series, thus producing streamflow probability distributions that are much smoother than resampling the historical record.

The generation of ensemble climate time series (WG-based resampling) for system verification is based on a cross-validation approach (Wilks 2005) to ensure the independence of the training and evaluation data. Given 25 years of available forecasts, when making forecasts for a particular year, the remaining 24 years can be used as training data. One thousand time series were generated for precipitation, Tmax and Tmin, to represent 1000 ensemble members for a particular year. One thousand members are used to obtain the true expectancy of a weather generator. A small number of samples could result in biases due to the random nature of the stochastic process.

3.2 Weather Generator-Based Post-Processing Approach

For the GPP, the EWFs (precipitation and temperature) are first subjected to the post-processing method of Chen et al. (2014a). This method uses a weather generator whose parameters are skewed to the EWFs. The method is briefly summarized below. Full details can be found in Chen et al. (2014a).

The post-processing of ensemble precipitation forecasts is achieved by estimating relationships between forecasted precipitation classes and probability of observed precipitation occurrence or observed wet-day precipitation for each specific season. Specifically, for any given lead day, a precipitation class is first determined according to the forecasted ensemble mean precipitation. The probability of observed precipitation occurrence corresponding to this class is used as the probability of precipitation for this day. A gamma function is fitted for the observed precipitation amounts within each class (one gamma function for one class). One thousand random numbers drawn from a uniform distribution are then generated to represent 1000 members. If random numbers are less than or equal to the probability of observed precipitation occurrence within this class, the corresponding members are predicted to be wet, otherwise, they are predicted to be dry. If a member is deemed wet, the fitted gamma distribution in the corresponding class is used to generate precipitation amounts with uniform random numbers. All the procedures are repeated for all 7 lead days. In the end, one thousand well-calibrated 7-day time series of precipitation are generated.

A linear regression method is used to correct the ensemble mean temperature. Linear equations are fitted between the observed and ensemble mean temperature anomaly, using a 31-day window centered on the day of interest. To obtain Tmax and Tmin needed to run the hydrology model, the corrected temperature anomaly is added to the mean of the observed Tmax and Tmin. The spread of the ensemble temperature is added using a two-parameter (mean and standard deviation) normal distribution. The corrected ensemble mean temperature for each day is used as the mean of the normal distribution. The standard deviation for each season is obtained using an optimization algorithm proposed by Chen et al. (2014a). For a given day, the temperature is found by multiplying the standard deviation by a normally-distributed random number, and adding the bias-corrected ensemble mean temperature for that day. As with precipitation, one thousand 7-day series of calibrated temperatures are generated.

Differently from other post-processing methods (e.g., BMA and logistic regression) that only produce the underlying probability distribution of forecasted variables, the GPP method directly generates the continuous time series, which makes it possible to be linked to hydrological models for hydrological forecasts. The biases of precipitation occurrence are also specifically considered. Additionally, the use of parametric distribution (e.g., gamma distribution) to simulate precipitation amounts allows simulating the extreme values outside the range of the observed data. This is one of the major advantages of this method over the analog technique. In this study, the post-processed precipitation and temperature time-series are un-biased and fully consistent with the EWFs for that 7-day period. The generation of the EWFs for the ESP system verification is also based on the cross-validation approach. When making forecasts for a particular year, the remaining 24 years are used as training data.

3.3 Hydrological Modelling

The hydrological model HSAMI, developed by Hydro-Québec, is used for the hydrology forecast. HSAMI is a 23-parameter, lumped, conceptual, rainfall-runoff model. It has been used to forecast natural inflows for over 20 years (Fortin 2000). It is used by Hydro-Québec for hourly and daily forecasting of natural inflows on nearly 100 watersheds, with surface areas ranging from 160 km^2 to 69,195 km^2. It was also used as an impact model in several climate change impact studies (Chen et al. 2011, 2012b; Minville et al. 2009). Of the 23 parameters, 2 account for evapotranspiration, 6 for snowmelt, 10 for vertical water movement, and 5 for horizontal water movement. Vertical flows are simulated with four interconnected linear reservoirs (snow on the ground, surface water, unsaturated and saturated zones). Horizontal flows are routed through two unit hydrographs and one linear reservoir. The model takes snow accumulation, snowmelt, soil freezing/thawing and evapotranspiration into account. Model calibration is done automatically using the Covariance Matrix Adaptation Evolution Strategy (CMAES) (Hansen and Ostermeier 1996, 2001).

The basin-averaged minimum required daily input data for HSAMI are: Tmax, Tmin, liquid and solid precipitations. Cloud cover fraction and snow water equivalent can also be used as inputs, if available. A natural inflow or discharge time series is also needed for proper calibration/validation. For this study, 10 years (1979–1988) of daily discharge data were used for model calibration and another 10 years (1989–1998) of data were used for validation over the CDD watershed. The optimal combination of parameters was chosen based on Nash-Sutcliffe criteria. The selected parameter set yielded Nash-Sutcliffe criteria values of 0.913 and 0.835 for calibration and validation periods, respectively. For the YAM watershed, 8 years (1995–2002) of daily discharge data were used for model calibration and 7 other years (2003–2009) were used for validation, with the Nash-Sutcliffe criteria values of 0.749 and 0.712 for the calibration and validation periods, respectively.

It should be noted that since raw GFS reforecasts only provide mean temperature, Tmax and Tmin forecasts for running HSAMI were recovered by adding 50 % of the differences between observed Tmax and Tmin to GFS temperature forecasts and subtracting 50 % of the differences as observed by GFS temperature forecasts. To validate this choice, a test using station-based hourly temperature (24 values per day) shows that the daily mean temperature estimated by averaging daily Tmax and Tmin is just as good as that estimated by averaging hourly temperature (results not shown).

3.4 Verification of the Ensemble Streamflow Predictions

The three ESP systems are first compared in terms of their ability to reproduce the mean annual hydrographs for CDD and YAM watersheds. A set of verification metrics derived from the Ensemble Verification System (EVS) (Brown et al. 2010) are then used to quantify the quality of these systems. The metrics consist of both deterministic and probabilistic metrics. The deterministic metrics measure the difference between the ensemble mean forecast and the reference (observed discharge in this study) for a specific variable. They include the correlation coefficient (r^2), relative mean error (RME), mean absolute error (MAE) and root mean square error (RMSE). The probabilistic metrics measure the forecast probability of an ensemble forecast system. They include the mean continuous ranked probability skill score (MCRPSS), Brier score (BS) and reliability diagram. The BS and reliability diagram only verify discrete events within the continuous forecast distributions. One or more thresholds have to be defined to represent the cutoff values from which discrete events are computed. Six quantiles with probability of streamflow exceeding 10 % (lower decile), 33 % (lower tercile), 50 % (median), 67 % (upper tercile), 90 % (upper decile) and 95 % (95th percentile) are used as thresholds in this study. All metrics are calculated for lead times ranging between 1 and 7 days. The details of these metrics can be found in Brown et al. (2010), Demargne et al. (2010) and the User Manual of the EVS (Brown 2012) (http://amazon.nws.noaa.gov/ohd/evs/evs.html).

4 Results

The three ESP systems are first compared in terms of reproducing the mean annual hydrographs for two watersheds. Since similar results are obtained for all 7 leads, only the results for the 1- and 3-leads are presented in Fig. 2 for illustration. To avoid bias from the hydrological modeling process, streamflows for the reference period are represented by the modeled streamflows and not by observations. The GFS system over-predicts the spring, summer and winter discharges for the CDD watershed, and the entire year discharge for the YAM watershed at both lead days, indicating the over-prediction of precipitation amounts in the raw EWF. However, the winter discharge of the CDD watershed is reasonably reproduced, because CDD is a snow-dominated watershed. The snowfall accumulates for a few months in the winter. Hence, the over-prediction of precipitation has little effect on winter streamflow. However, the overestimation of snowmelt peak discharge still indicates an over-prediction of winter snowfall (Fig. 2a and b). In fact, the snowmelt peak discharge in snow-dominated watersheds is mainly affected by temperature. The relatively good timing of the average peak indicates that the ensemble mean temperature is not too biased.

The WG system reasonably reproduces the mean annual hydrograph for the CDD watershed with the exception of a slight over-prediction of the peak discharge. However, it under-predicts the winter discharge and over-predicts the spring discharge for the YAM watershed. This may be due to a slight cold bias in the historical resampling temperatures. The cold bias of temperatures does not have much influence on the winter discharge for the snow-dominated

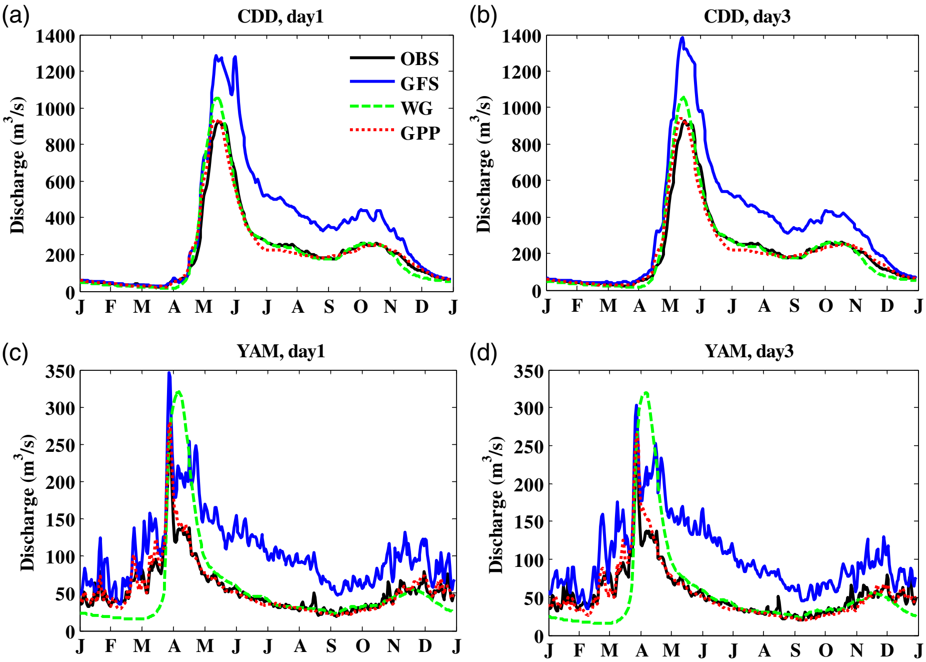

Fig. 2 Mean annual hydrographs of ensemble mean streamflow predicted by three ESP systems (GFS, WG and GPP) for 1- and 3-lead days over the CDD and YAM watersheds. The mean annual hydrographs simulated using observed meteorological data (OBS) are also plotted

watershed, but it significantly affects the winter discharge for the rainfall-dominated watershed.

The GPP system reproduces the mean annual hydrographs very well for both watersheds and all leads. Overall, this preliminary test clearly show that the post-processing of the EWFs is necessary, and that historical resampling may be adequate for snow-dominated watersheds, but it is more problematic for smaller or warmer watersheds. This further confirms the ability of the post-processing method presented by Chen et al. (2014a) to remove the biases of ensemble forecasts.

Figure 3a–h present the quality of the ensemble mean streamflows in terms of r^2, RME, MAE and RMSE. The statistics were calculated for all forecast-observation pairs. As indicated in Fig. 3a and b, the observed discharge and ensemble mean discharge predicted by the GFS system show a strong correlation for the CDD watershed and a relatively lower correlation for the YAM watershed. However, they are consistently better than those predicted by the WG system for all leads, especially for the YAM watershed. This performance can be attributed to the fact that the GFS system consistently over-predicts the discharge overall, as indicated in Fig. 2, while the WG system over-predicts the discharge for some seasons and under-predicts it for others, especially for the YAM watershed. The GPP system slightly improves the correlation for the CDD watershed and markedly improves the correlation for the YAM watershed. Moreover, there is a progressive decline in the correlation with increasing lead time, especially for the YAM watershed. However, the degradation in the correlation is more rapid for the GPP system than it is for the GFS system. This is especially noteworthy, when after 6 lead days, the GFS system's correlation is close to or even slightly better than that of the GPP system, indicating the use of the 7th day as the limit of post-processing is reasonable in terms of the correlation between observed and predicted streamflow.

The ensemble mean discharge is considerably over-predicted by the GFS for both watersheds, as shown in Fig. 3c and d. The bias for the CDD watershed is much greater than that of the YAM watershed, with mean RMEs of 60.5 and 88.1 %, respectively, across all 7 lead days.

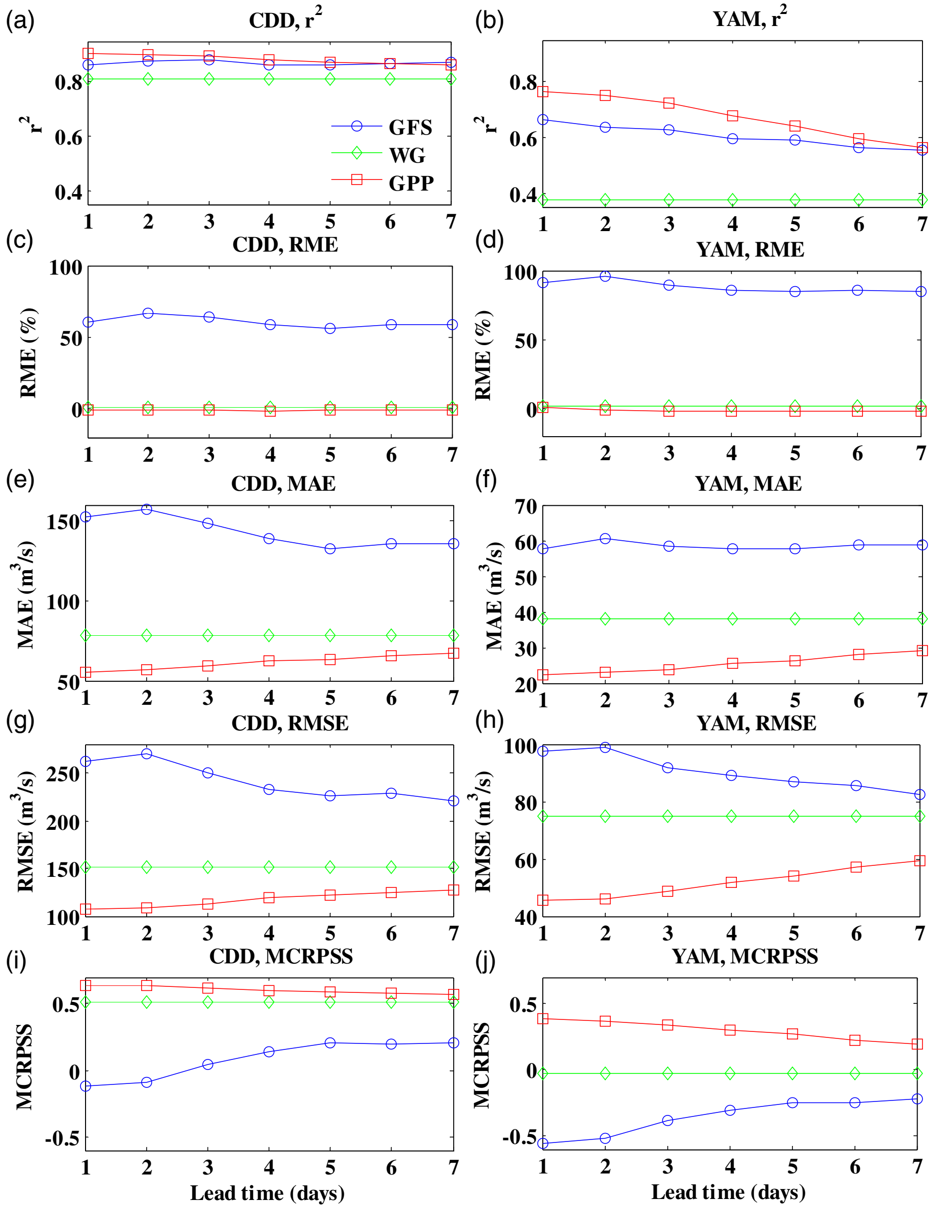

Fig. 3 Correlation coefficients (r^2), relative mean error (RME), mean absolute error (MAE), root mean square error (RMSE) and mean continuous ranked probability skill score (MCRPSS) of ensemble mean streamflow predicted by three ESP systems (GFS, WG and GPP) for 1- to 7-lead days over the CDD and YAM watersheds

However, little bias is observed for the ensemble mean discharge predicted by both WG and GPP systems, with the latter being slightly better.

Similarly to RMEs, the ensemble mean streamflows predicted by the GFS display a large error for both watersheds and all 7 leads in terms of the MAE and the RMSE (Fig. 3e–h). The WG system consistently performs better than the GFS system, while GPP is consistently the best. With increasing lead time, the quality of ensemble mean streamflow predicted by raw GFS slightly degrades for both watersheds. This may be due to the lack of spread in shorter lead ensemble forecasts and the larger spread in longer lead forecasts (Chen et al. 2014a). Meanwhile, the GPP's performance gradually decreases with increasing the lead time, outlining the diminishing skill of the EWF.

As indicated by MCRPSS (Fig. 3i and j), the ensemble streamflow predicted by the GFS system shows negative or near zero skill relative to the reference. It is worth observing that the MCRPSS is always less than −0.2 for the YAM watershed. Meanwhile, the forecast skill consistently increases with the forecast leads. Again, this may be caused by the lack of spread of shorter lead-time ensemble forecasts and the larger spread of longer lead-time forecasts. The ensemble streamflows predicted by WG display zero skill for the YAM watershed and positive skill for the CDD watershed, relative to the references. The GPP system consistently increases the skill of the ESP, especially for the YAM watershed. Although the skill of the GPP system progressively deteriorates with an increase in lead time, it is still greater than that of the WG system for the 7-lead day. Even though Chen et al. (2014a) showed that the post-processed EWFs retained little skill after 7 days for precipitation relative to the climatology, the combination of precipitation and temperature (which has a higher skill level) provides more skill than that of the climatic ESP at 7 days. Additionally, the ESP skill is watershed-size dependent. The predictive power of ESP is consistently higher for the CDD watershed than for the YAM watershed.

For probabilistic metrics computed for discrete events, such as for BS and reliability diagrams used in this study, a number of thresholds have been defined. Since similar trends are obtained for all six thresholds, only the results with a probability exceeding the median (50 %) are presented to illustrate the trends for both metrics.

BS can be decomposed into three components: reliability, resolution and uncertainty (Brown 2012). Figure 4 presents the BS and its components of ensemble streamflows predicted by all three ESP systems for both watersheds and 7 forecast leads. The reliability term of a BS measures how close the forecast probabilities are to the true probabilities, with smaller values indicating a better forecasting system. The resolution term measures how much the predicted probabilities differ from the hydrological average, with larger values indicating a better forecast system. According to its definition, the value of an uncertainty component is 0.25 (0.5×(1–0.5)) when using median as the threshold.

In terms of the BS and its reliability and resolution components, the GPP system is consistently more accurate than the other two systems at predicting ensemble streamflows for both watersheds and all 7 lead days and the WG system is more accurate than the GFS system. Additionally, all of the systems are more accurate with their predictions for the CDD watershed than for the YAM watershed.

As displayed by the BSS levels, the GFS system shows negative skill for all 7 forecast leads at the YAM watershed. The WG system is consistently more skillful than the GFS system with positive BSS values for the YAM watershed. The GFS system is very skillful for the CDD watershed when using the median as the threshold for all 7 leads. However, it is still consistently worse than the other two systems. The GPP system is consistently more skillful

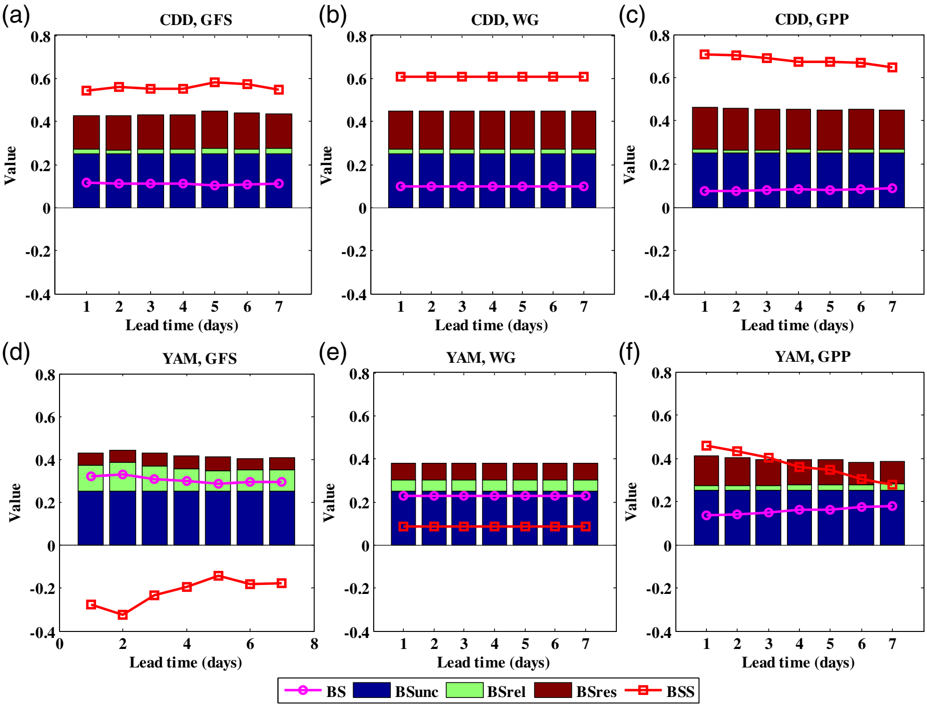

Fig. 4 BS and its decomposed components (Brier score uncertainty (BSunc), Brier score reliability (BSrel), and Brier score resolution (BSres)) of (*left*) GFS, (*middle*) WG, and (*right*) GPP ensemble streamflow predictions for 1–7 lead days over the (**a**)–(**c**) CDD and (**d**)–(**f**) YAM watersheds. Probability exceeding 50 % is used as the threshold. The BSS of the ensemble forecasts relative to the observed discharge is also presented

than the other two systems. Even though the skill of the GPP system decreases with the increase of forecast leads, the smallest BSS value of GPP is greater than the largest BSS value of the other two systems for both watersheds.

Figure 5 presents the reliability diagrams of the three ESP systems using probabilities of streamflow exceeding the median (50 %) as a threshold, for 1- and 3-lead days over the two watersheds. The sharpness histograms are presented alongside the reliability diagrams. Overall, the probabilities of streamflow are typically over-predicted by the GFS system for two watersheds, with the exception of the 1-lead day forecast with probability being lower than 50 % for the CDD watershed. This forecast has little resolution, especially at the lower forecast probabilities, as indicated by its reliability curves that are below the 1:1 reference line for most cases. The GFS is overconfident, as reflected in the U-shape sharpness (too sharp but with little reliability). Overall, the WG system is much more reliable than the GFS system, with little or no decrease in the sharpness for two watersheds. However, the observed probabilities are typically under-forecasted for the CDD watershed when forecast probability is lower than 50 %. In particular, the observed lower probabilities (<20 %) are considerably under-forecasted for the YAM watershed. However, the higher probabilities (>20 %) are reasonably well predicted. The GPP system dramatically improves the reliability and resolution of ESP for both watersheds and for 1- and 3-lead days, as displayed by reliability curves close to the 1:1 reference line. More importantly, there is little or no decline in the sharpness.

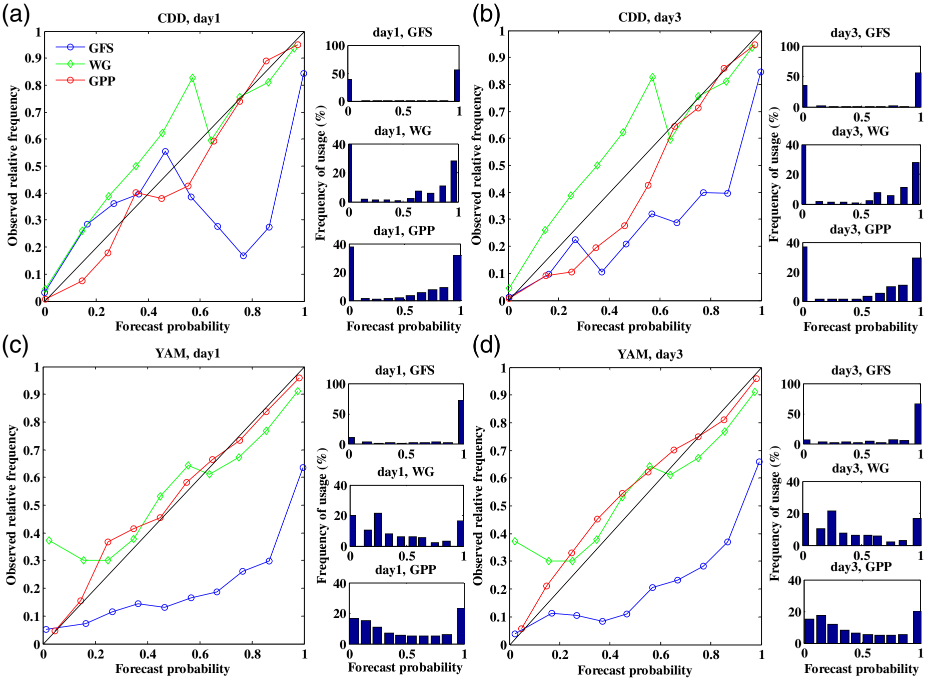

Fig. 5 Reliability diagrams of ensemble streamflow predicted by three ESP systems (GFS, WG and GPP) for 1- and 3-lead days over the (**a**)–(**b**) CDD and (**c**)–(**d**) YAM watersheds. The probability of streamflow exceeding 50 % (Q>50 %) is used as a threshold

5 Discussion and Conclusion

This study presented an ESP system for short-term hydrological forecasts using the post-processing method for EWFs. The proposed ESP system (GPP) was further compared to the other two systems (GFS and WG) for hydrological predictions over two Canadian watersheds in the province of Quebec. Overall, the WG system is generally better than the GFS system at predicting the ensemble mean streamflow for both watersheds and all leads ranging between 1 and 7 days. The GPP system is consistently better than the other two systems in terms of the overall deterministic metrics. Probabilistic metrics indicate that little skill is retained for the ESP produced by the GFS system, indicating that post-processing the EWFs is necessary for hydrological predictions. The WG system, which only takes into account the historical record of climate events, is consistently better than the raw EWFs. However, this system assumes that the past climatology is adequate at representing the current climatology, which is a weakness in the context of an evolving climate. Post-processing the EWFs considerably improves the predictive power of the ESP for all leads of between 1 and 7 days.

The performance of ESP systems is watershed-dependent in this study. All forecast systems consistently perform better over the CDD watershed, most likely due to the difference in watershed size. The CDD watershed is much larger than the YAM watershed and it is more consistent with the scale of the numerical weather model. The post-processing procedure used for calibrating the EWF also acts as a statistical downscaling method, and because the spatial resolution of the GFS reforecasts is much larger than the area of the YAM watershed, the post-processing method faces a more difficult task. Chen et al. (2014a) indicated that the ensemble precipitation forecasts retained little skill past 7 days. However, this study showed that the GPP system was still more skillful than the WG system at the 7th lead day with the exception of using r^2 as a metric. There are two likely reasons why the skill of the ESP system exceeds that of the EWF. The first reason is that the ESP reflects the synergic effect of precipitation and temperature. Even though there was little skill for precipitation forecast, the temperature retained a significant skill level after 1 week. The second reason is the lag between precipitation and streamflows, as the precipitation for 1 week may still affect the runoff for the following week. However, with increasing lead days, the skill and reliability of the GPP forecast converges toward that of the WG forecast, indicating the possibility of combining GPP and WG forecasts for short- to medium-term ESP. The lead time at which the both forecasts converge can be used as a transition point between the two forecast systems.

This study coupled EWFs to a lumped hydrological model for ESP. To use the lumped model, all observed grid points within the watershed were averaged. Averaging precipitation and temperature may lead to a loss of information on the spatial variability, and particularly for precipitation and for larger watersheds. In such cases, it may be more reasonable to use a distributed hydrological model for the flow forecasts. To properly achieve this, the raw EWFs would have to be post-processed using a multi-site stochastic weather generator which adds a complexity layer to the forecast, even though multi-site weather are now freely available (i.e., Chen et al. 2014b).

The weather generator-based approach post-processed precipitation and temperature independently. However, a correlation usually exists between precipitation and temperature. A previous study (Chen et al. 2014a) showed that the GFS forecasts overestimated the precipitation–temperature correlation, while the GPP forecasts slightly underestimated the correlation. These results are not entirely surprising since any post-processing method is expected to alter the precipitation–temperature correlation, unless it is specifically taken into account.

Finally, the proposed GPP system was only tested over two Quebec watersheds rather than watersheds from different climate regions. For a broader use of this system, a more comprehensive test, including using more datasets from different climate regions and comparisons with other approaches such as BMA would be required in further studies.

Acknowledgments This work was funded in part by the Projet d'Adaptation aux Changements Climatiques (PACC26) of the Province of Quebec, Canada. The authors would like to thank the Centre d'Expertise Hydrique du Québec (CEHQ) and the Ouranos Consortium on Regional Climatology and Adaptation for their contributions to this project. The authors also wish to thank Dr. James D. Brown of NOAA/national Weather Service, Office of Hydrologic Development, for providing the latest version of Ensemble Verification System (EVS), and the Earth System Research Laboratory, Physical Sciences Division for providing reforecast products.

References

Boucher MA, Anctil F, Perreault L, Tremblay D (2011) A comparison between ensemble and deterministic hydrological forecasts in an operational context. Adv Geosci 29:85–94

Brown DJ (2012) Ensemble Verification System (EVS). Version 4.0, User's Manual, pp. 107

Brown DJ, Demargne J, Seo DJ, Liu Y (2010) The Ensemble Verification system (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. Environ Model Softw 25:854–872

Chen J, Brissette FP (2014) Comparison of five stochastic weather generators in simulating daily precipitation and temperature for the Loess Plateau of China. Int J Climatol 34:3089–3105

Chen J, Brissette FP, Poulin A, Leconte R (2011) Overall uncertainty study of the hydrological impacts of climate change for a Canadian watershed. Water Resour Res 47, W12509. doi:10.1029/2011WR010602

Chen J, Brissette FP, Leconte R (2012a) Downscaling of weather generator parameters to quantify the hydrological impacts of climate change. Clim Res 51:185–200

Chen J, Brissette FP, Leconte R, Caron A (2012b) A versatile weather generator for daily precipitation and temperature. Trans ASABE 55(3):895–906

Chen J, Brissette FP, Zhang XC (2014a) A multi-site stochastic weather generator for daily precipitation and temperature. Trans ASABE 57(5):1375–1391

Chen J, Brissette FP, Li Z (2014b) Post-processing of ensemble weather forecasts using a stochastic weather generator. Mon Weather Rev 142:1106–1124

Clark MP, Hay LE (2004) Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. J Hydrometeorol 5:15–32

Cloke HL, Pappenberger F (2009) Ensemble flood forecasting: a review. J Hydrol 375:613–626

Day GN (1985) Extended streamflow forecasting using NWSRFS. J Water Res Pl-ASCE 111(2):157–170

Demargne J, Brown J, Liu Y, Seo DJ, Wu L, Toth Z, Zhu Y (2010) Diagnostic verification of hydrometeorological and hydrologic ensembles. Atmos Sci Lett 11:114–122

Dutta D, Welsh WD, Vaze J, Kim SSH, Nicholls D (2012) A comparative evaluation of short-Term streamflow forecasting using time series analysis and rainfall-runoff models in eWater source. Water Resour Manag 26(15):4397–4415

Eckel FA, Walters MK (1998) Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. Weather Forecast 13:1132–1147

Fortin V (2000) Le modèle météo-apport HSAMI: historique, théorie et application. Institut de Recherche d’Hydro-Québec, Varennes, p 68

Ghile YB, Schulze RE (2010) Evaluation of three numerical weather prediction models for short and medium range agrohydrological applications. Water Resour Manag 24(5):1005–1028

Hamill TM, Colucci SJ (1997) Verification of Eta–RSM short-range ensemble forecasts. Mon Weather Rev 125:1312–1327

Hamill TM, Whitaker JS (2006) Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. Mon Weather Rev 134:3209–3229

Hamill TM, Whitaker JS, Wei X (2004) Ensemble reforecasting: improving medium-range forecast skill using retrospective forecasts. Mon Weather Rev 132:1434–1447

Hamill TM, Whitaker JS, Mullen SL (2006) Reforecasts: an important dataset for improving weather predictions. Bull Am Meteorol Soc 87:33–46

Hansen N, Ostermeier A (1996) Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In Proceedings of the 1996 I.E. international conference on evolutionary computation 312–317

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

Hutchinson MF, McKenney DW, Lawrence K, Pedlar JH, Hopkinson RF, Milewska E, Papadopol P (2009) Development and testing of Canada-wide interpolated spatial models of daily minimum-maximum temperature and precipitation for 1961–2003. J Appl Meteorol Climatol 48:725–741

Minville M, Brissette F, Krau S, Leconte R (2009) Adaptation to climate change in the management of a Canadian water-resources system. Water Resour Manag 23(14):2965–2986

Richardson CW (1981) Stochastic simulation of daily precipitation, temperature, and solar radiation. Water Resour Res 17:182–190

Richardson CW, Wright DA (1984) WGEN: a model for generating daily weather variables. US Department of Agriculture, Agricultural Research Service, ARS-8, pp 83

Roulin E, Vannitsem S (2005) Skill of medium-range hydrological ensemble predictions. J Hydrometeorol 6:729–744

Schaake J, Franz K, Bradley A, Buizza R (2006) The Hydrologic Ensemble Prediction EXperiment (HEPEX). Hydrol Earth Syst Sci Discuss 3:3321–3332

Şensoy A, Uysal G (2012) The value of snow depletion forecasting methods towards operational snowmelt runoff estimation using MODIS and numerical weather prediction data. Water Resour Manag 26(12):3415–3440

Tang G, Zhou H, Li N, Wang F, Wang Y, Jian D (2010) Value of medium-range precipitation forecasts in inflow prediction and hydropower optimization. Water Resour Manag 24(11):2721–2742

Velazquez JA, Petit T, Lavoie A, Boucher MA, Turcotte R, Fortin V, Anctil F (2009) An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting. Hydrol Earth Syst Sci 13:2221–2231

Wilks DS (2005) Statistical methods in the atmospheric sciences, 3rd edn. Academic Press, New York, p 467

Wilks DS (2006) Comparison of ensemble-MOS methods in the Lorenz’96 setting. Meteorol Appl 13:243–256