ATMOSPHERIC SCIENCE LETTERS
Atmos. Sci. Let. **11**: 114–122 (2010)
Published online 6 April 2010 in Wiley InterScience
(www.interscience.wiley.com) **DOI:** 10.1002/asl.261

RMetS
Royal Meteorological Society

Diagnostic verification of hydrometeorological and hydrologic ensembles†

Julie Demargne,[1,2]* James Brown,[1,2] Yuqiong Liu,[1,3] Dong-Jun Seo,[1,2] Limin Wu,[1,4] Zoltan Toth[5] and Yuejian Zhu[6]

[1]*National Oceanic and Atmospheric Administration, National Weather Service, Office of Hydrologic Development, Silver Spring, Maryland, USA*
[2]*University Corporation for Atmospheric Research, Boulder, CO, USA*
[3]*Riverside Technology, Inc., Fort Collins, CO, USA*
[4]*Wyle Information Systems, McLean, Virginia, USA*
[5]*National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Boulder, CO, USA*
[6]*National Oceanic and Atmospheric Administration, National Weather Service, National Centers for Environmental Prediction, Camp Springs, Maryland, USA*

**Correspondence to:*
Julie Demargne, National Oceanic and Atmospheric Administration, National Weather Service, Office of Hydrologic Development, Silver Spring, Maryland, USA.
E-mail: julie.demargne@noaa.gov

†*This article is a U.S. Government work and is in the public domain in the U.S.A.*

Received: 30 September 2009
Revised: 18 December 2009
Accepted: 26 January 2009

Abstract

This paper presents a strategy for diagnostic verification of hydrologic ensembles, based on the selection of summary verification metrics (which could be extended to more detailed metrics) and the analysis of the relative contribution of the different sources of error. Such diagnostic verification could be conducted with the Ensemble Verification System (EVS) and is illustrated with a verification case study of experimental precipitation and streamflow ensemble reforecasts over a 24-year period. The EVS is proposed as a flexible and modular tool for the HEPEX verification test-bed to evaluate existing and emerging verification methods that are appropriate for hydrologic applications. Published in 2010 by John Wiley & Sons, Ltd.

Keywords: **ensembles; hydrological forecasting; probabilistic verification; uncertainty**

1. Introduction

Atmospheric and hydrologic forecasts are subject to uncertainty, which needs to be systematically quantified and effectively communicated to users (NRC, 2006). A common approach to providing such information in an operational forecasting context is to generate ensemble forecasts from which probabilistic statements are issued [e.g. see examples of operational ensemble flood forecasting systems in (Cloke and Pappenberger, 2009)]. Hydrologic ensembles and their corresponding hydrometeorological forecasts need to be *routinely* verified to improve both research and operations (Welles *et al*., 2007). Although hydrologic forecast verification has been limited to date, a number of verification case studies with hydrologic ensembles have been published (see references quoted in Cloke and Pappenberger, 2009). Furthermore, the meteorology and hydrology communities need to closely collaborate to define verification metrics and practices that are appropriate for hydrologic applications (Pappenberger *et al*., 2008). Such forecast verification needs to include two activities (Demargne *et al*., 2009): (1) diagnostic verification performed by scientists and operational forecasters to monitor forecast quality over time, analyze the different sources of error and skill across the entire river forecasting process, and evaluate forecast skill improvement from new science and technology; and (2) real-time verification, which aims to communicate along with real-time forecasts (and before the corresponding observations occur), verification information relative to historical analogue forecasts to assist operational forecasters and end users in their decision making.

The Office of Hydrologic Development (OHD) of the National Oceanic and Atmospheric Administration (NOAA) National Weather Service (NWS) has developed various capabilities for the Hydrologic Ensemble Forecast System (HEFS) to provide river ensemble forecasts for a wide range of spatiotemporal scales, from hours for flash flood forecasts at local scale, to months for water supply forecasts at regional scale. The Ensemble Verification System (EVS) developed by Brown *et al*. (2010) is the diagnostic verification component of the HEFS, designed to verify ensemble forecasts of any continuous numeric variables, produced at discrete locations and for any forecast horizon and time step. The OHD and the National Centers for Environmental Prediction (NCEP) are currently collaborating to improve the climate, weather, and river forecasts at the catchment scale for the HEFS and define standard verification metrics and products that are meaningful for hydrologic applications.

In this paper, a strategy for diagnostic verification is proposed for hydrologic ensembles, based on the selection of summary verification metrics and the analysis

of the different sources of error in the forecasting system. This approach is illustrated with a verification case study of experimental HEFS precipitation and streamflow ensembles from a 24-year period using the EVS software. Finally, future work and on-going collaborations to advance ensemble verification in operational river forecasting are described.

2. Diagnostic verification of hydrologic forecasts

The quality of forecast ensembles includes several attributes (Wilks, 2006), such as reliability, resolution, discrimination, and skill. Therefore, a variety of verification metrics need to be analyzed concurrently. The following verification metrics are proposed as key summary metrics to describe the main aspects in forecast quality and are briefly described hereafter (see further details in Jolliffe and Stephenson, 2003, Wilks, 2006, and Brown *et al*., 2010). Such summary verification information could be used by forecasters and scientists, as a screening tool before analyzing further specific quality attributes for events of interest, as well as by managers for tracking forecast performance.

To analyze first the quality of the ensemble mean, which is commonly used in operational forecasting as a convenient choice for single-valued representation of the ensemble forecast, two metrics from single-valued forecast verification are used: the mean error to measure how the ensemble mean agrees with the observed outcome on average, and the correlation coefficient to describe the relationship between the ensemble mean and the corresponding observation. Then the overall quality of the probabilistic forecast is described with the Continuous Ranked Probability Score (CRPS), which measures the integrated squared difference between the cumulative distribution function (cdf) of a forecast, $F_Y(y)$, and the corresponding cdf of the observation, $F_X(x)$ (which has probability 1.0 for values greater than or equal to the observation and probability 0.0 otherwise). The CRPS is given by:

$$\mathrm{CRPS} = \int_{-\infty}^{+\infty} [F_Y(y) - F_X(y)]^2 dy$$

In practice, the CRPS is averaged across a set of observed-forecast pairs. To evaluate the skill of the forecasting system relative to a reference system, the associated skill score, the Continuous Ranked Probability Skill Score (CRPSS), is computed from:

$$\mathrm{CRPSS} = 1 - \frac{\overline{\mathrm{CRPS}_{\mathrm{forecast}}}}{\overline{\mathrm{CRPS}_{\mathrm{reference}}}}$$

the overbar referring to the CRPS averaged across the sample of events. The CRPSS ranges from $-\infty$ to 1, with perfect skill of 1 and negative value when the forecast has worse CRPS than the reference. The CRPSS is a useful companion to the CRPS, because events with small probabilities of occurrence are associated with very small (squared) probabilities of error, and this attribute will be shared with the reference forecasting system, allowing errors in the tails of the probability distribution to be better identified.

To provide further details about the forecast performance, the CRPS decomposition (Hersbach, 2000) is given by:

$$\mathrm{CRPS} = \mathrm{Reliability} + \mathrm{Potential\ CRPS}$$

The reliability component of the CRPS, called CRPS Reliability, measures the average reliability of the ensemble forecasts similarly to the rank histogram (see references in Wilks, 2006). Specifically, it tests whether the fraction of observations that fall below the k-th of n ranked ensemble members is equal to k/n on average. The second component of the CRPS, called the Potential CRPS, represents the CRPS one would obtain when the forecasting system would become perfectly reliable (i.e. CRPS Reliability = 0). It is sensitive to the average ensemble spread and the frequency and magnitude of the outliers. For best Potential CRPS, the forecasting system needs narrow ensemble spread on average without too many and too high ensemble outliers (Hersbach, 2000). The CRPS, the CRPS Reliability, and the Potential CRPS are all negatively oriented, with perfect score of 0.

Finally, the Relative Operating Characteristic (ROC) score is used as a summary score to describe the ability of the forecasts to discriminate between events and non-events. The ROC curve plots the probability of detection against the probability of false detection for a range of probability levels (each one corresponding to a threshold at which a probability forecast leads to a binary decision), and for a given event (such as flooding). The ROC score is defined as the area below the ROC curve and above the diagonal, with a perfect score of 1, and measures the overall gain in discrimination over climatological forecasts for all probability levels.

All these verification metrics are proposed as key summary verification metrics because they are thought to convey the main attributes of forecast quality. For further analysis, more detailed verification statistics could be examined. For example, further information may be required on the reliability of the forecast probabilities in different parts of the forecast distribution, which could be analyzed with the cumulative rank histogram (a measure closely related to the CRPS). Similarly, the ROC score may be extended to the ROC diagram, which identifies the discriminatory power of the forecasting system for different decision situations. Because the ROC metrics measure only discrimination relative to specific observed events, their analysis may be accompanied by metrics that specifically measure reliability for each observed threshold, such as the reliability component of the Brier Score or the reliability diagram (Wilks, 2006). However, discussions of more

detailed verification metrics are not included in this paper.

In addition, hydrologic ensemble forecasts need to account for the atmospheric uncertainty and the hydrologic uncertainty, which includes uncertainty in the initial conditions, the model parameters and the model structure (Gupta *et al*., 2005). Forecasters and modelers need to analyze how the different sources of error affect the quality of hydrologic forecasts and which parts of the forecasting system represent the main sources of skill and error in these forecasts. Therefore, all the forcing input forecasts and hydrologic output forecasts should be verified, potentially at various temporal and spatial scales and for different forecast horizons depending on the forecast applications. The forecast performance needs to be analyzed under different conditions by stratifying the forecast-observed dataset and reporting verification statistics for subsets of events. To assess the skill in the atmospheric and hydrologic ensemble forecasts, skill scores should be computed with a reference probabilistic forecast that is meaningful for the application of interest. For example, to show how much skill the use of weather and/or climate forecasts may add to the atmospheric and hydrologic ensemble forecasts, reference forecasts could be defined using only climatology information for the atmospheric ensembles and the corresponding climatology-based hydrologic ensembles that are produced by the same hydrologic prediction system.

Furthermore, to analyze the relative importance of the atmospheric and hydrologic uncertainties, flow ensemble forecasts should be verified with the observed flows and with the simulated flows that are produced from the observed hydrometeorological inputs using the same model and the same initial conditions. The verification of flow ensembles with observed flows leads to the computation of the total error, including the contribution of the atmospheric uncertainty and the hydrologic uncertainty. The verification with simulated flows allows for the contribution of the atmospheric uncertainty (in the hydrometeorological forecasts) to be diagnosed, assuming that uncertainties in the observed hydrometeorological inputs are much smaller than the hydrologic uncertainty.

Such a diagnostic verification strategy could be conducted with the EVS software. The main features of the EVS are summarized below; a detailed description is provided in Brown *et al*. (2010). The EVS can perform temporal aggregation (e.g. daily total flows aggregated from 6-hourly instantaneous flows) and data stratification to verify subsets of forecast-observed pairs (e.g. for winter months, above an exceedance threshold). The EVS can aggregate the verification statistics produced across different locations to easily report forecast quality on larger areas. Finally, the EVS produces a range of graphical and numerical outputs of the verification statistics. The EVS software has been made available on line (http://www.nws.noaa.gov/oh/evs.html) to support collaborative work such as the Hydrological Ensemble Prediction Experiment (HEPEX) verification testbed project (http://hydis8.eng.uci.edu/hepex/testbeds/Verification.htm).

3. Verification case study

To illustrate the proposed diagnostic verification strategy, a case study is presented for experimental ensemble hindcasts of precipitation and flow generated with the current HEFS prototype. The precipitation ensembles (as well as temperature ensembles) are generated from single-valued forecasts by the NWS Ensemble Preprocessor (EPP) (Schaake *et al*., 2007). The EPP aims to remove the bias in the NWP single-valued forecasts while capturing the skill and uncertainty therein. The EPP estimates the joint distribution of single-valued forecasts and observations based on historical pairs. Ensemble members are sampled from the conditional probability distribution of the observations given a particular single-valued forecast. The Schaake Shuffle technique (Clarke *et al*., 2004) is applied to approximately reconstruct the space-time statistical properties of the precipitation and temperature variables for multiple lead times and locations based on historical observations. When no single-valued forecast is available, EPP estimates the climatological distribution from the historical observations and applies the Schaake Shuffle to the values sampled from the distribution. The resulting ensembles, termed resampled climatological ensembles, are used as reference forecasts to analyze the skill in the ensembles derived from the NWP single-valued forecasts.

The hydrometeorological ensemble hindcasts produced by the EPP are ingested into the Hydrologic Ensemble Hindcaster (HEH) (Demargne *et al*., 2007) to produce corresponding flow ensemble hindcasts based on various hydrological models. The HEH retrospectively generates the initial conditions of the hydrological models for each hindcast date. These retrospective initial conditions may not reflect the initial conditions used in real-time forecasting, which are usually modified by the forecasters based on their expertise, or by data assimilation techniques. However, this hindcast process supports the analysis of the impact of the atmospheric ensembles on the quality of hydrologic ensembles. Two sets of flow ensembles are generated: one using the EPP ensembles derived from the NWP single-valued forecasts, the other using the EPP resampled climatological ensembles, to analyze the skill in the flow forecasts when incorporating information from the NWP single-valued forecasts. These two sets of hydrologic ensembles account only for the atmospheric uncertainty, the hydrologic uncertainty being quantified by other components of the HEFS.

The verification study is presented for the North Fork of the American River above the North Fork Dam

(USGS stream gauge station ID 11427000) near Sacramento in north-central California. This is a headwater basin of 875 km^2 for which precipitation is the main forcing input. The NWP single-valued forecasts were obtained from the ensemble means of the precipitation and temperature reforecasts from the frozen version (circa 1998) of the NCEP's Global Forecast System (GFS) at T62 resolution for 14 days into the future (Hamill *et al.*, 2006). The EPP produced 6-hourly mean areal precipitation and mean areal temperature ensemble hindcasts at 12 : 00 UTC, from which the HEH generated 6-hourly flow ensembles for 14 days of forecast horizon. These hindcasts were produced for a period of almost 24 years from 1 January 1979 to 30 September 2002, each hindcast containing 55 ensemble members. The EPP resampled climatological ensembles and the corresponding climatology-based flow ensembles were also produced as reference forecasts. The EPP was calibrated using the forecasts and observations from the same period; independent verification analysis is currently being conducted. The precipitation forecasts were aggregated in EVS to be verified for each lead time as daily totals using precipitation observations. The precipitation verification statistics were also aggregated across two precipitation subareas. The 6-hourly flow forecasts were aggregated to daily averages to be verified with the USGS flow measurements that were available only at daily time step. To assess the relative contribution of the atmospheric and hydrologic uncertainties in the flow forecasts, the daily flow forecasts were also verified with the daily averages of the 6-hourly flow simulations generated from the observed hydrometeorological inputs using the same hydrologic model and initial conditions.

Verification statistics were computed using the whole 24-year period to verify, with sufficiently large sample sizes, the forecast performance for high events (defined by thresholds on the observed sample), which is critical for operational forecasting. Work is underway to estimate the confidence intervals of the verification metrics based on a bootstrapping approach to account for the sampling uncertainty. A preliminary assessment of confidence intervals for this case study (not shown) showed that sampling uncertainty becomes significant after Day 12 (i.e. a lead time of 12 days) (especially for the higher thresholds), rendering it difficult to draw any meaningful conclusions regarding the differences in forecast quality between the climatology-based ensembles and the GFS-based ensembles for these long forecast horizons.

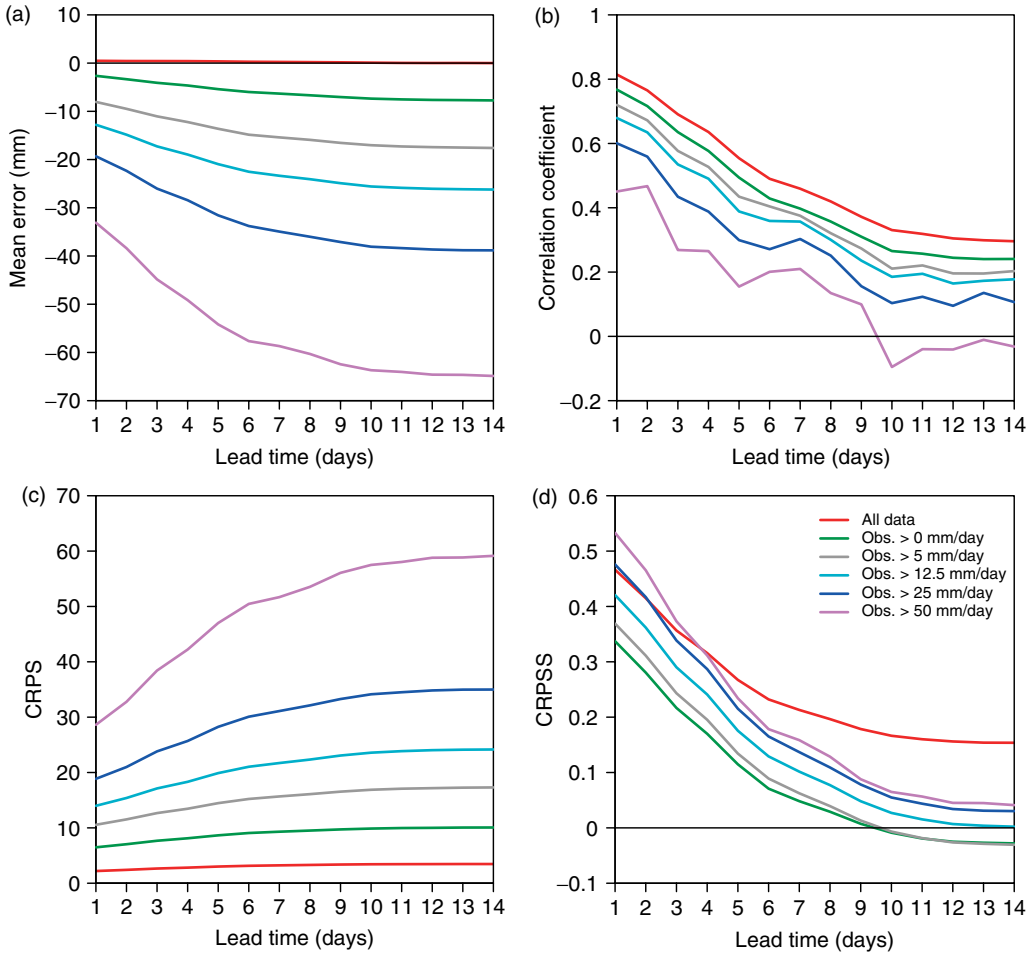

Figure 1. Mean Error (a) and Correlation Coefficient (b) of the ensembles means, as well as CRPS (c) and CRPSS (d) (in reference to resampled climatological ensembles) for the GFS-based precipitation ensembles.

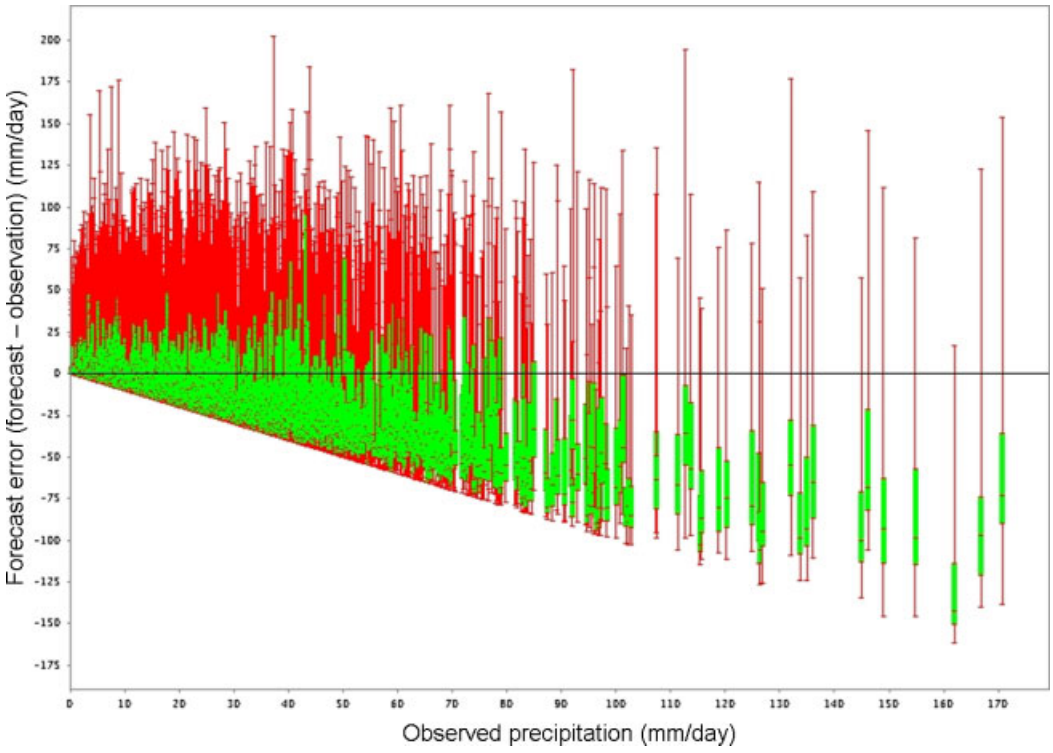

Figure 2. Box-and-whisker plot for the 0, 25, 50, 75, and 100 percentiles of the forecast error distribution for the GFS-based precipitation ensembles and for the first 24-h lead time.

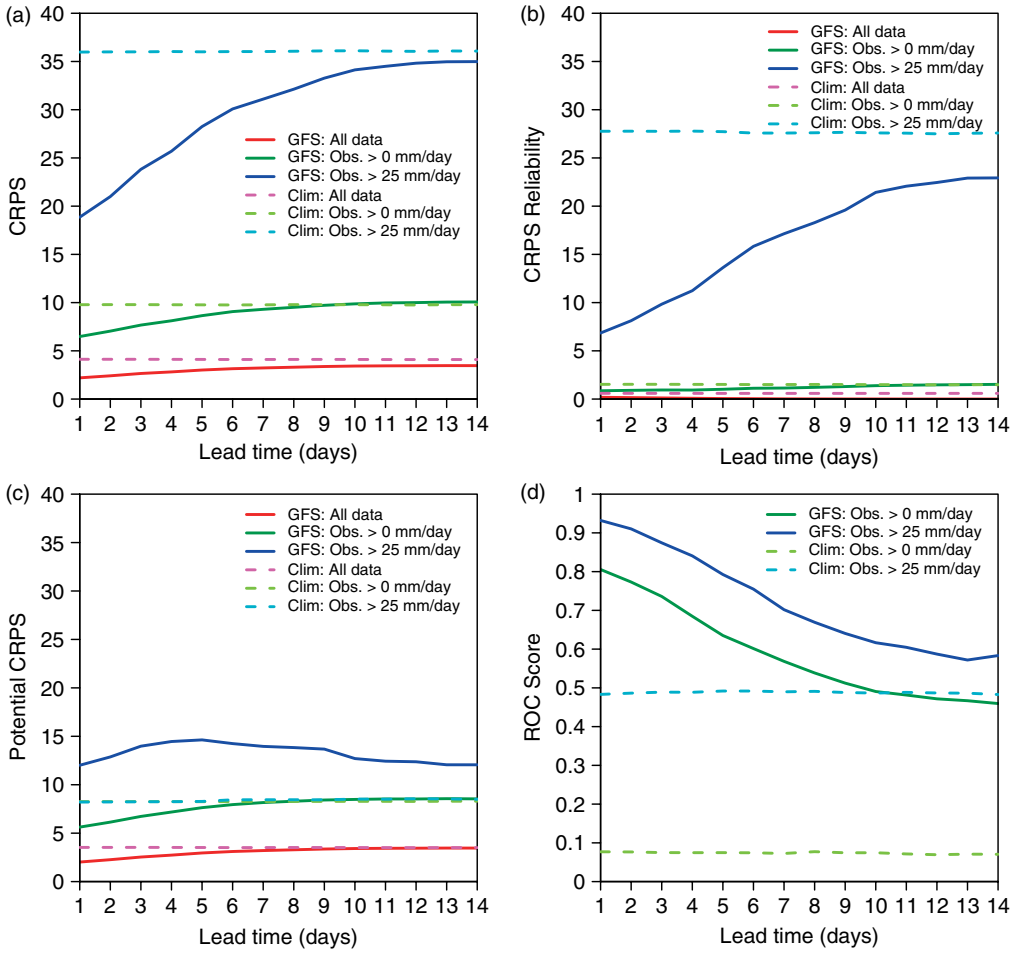

Figure 3. CRPS (a), CRPS Reliability (b), Potential CRPS (c), and ROC Score (d) for the GFS-based precipitation ensembles ('GFS') and the resampled climatological ensembles ('Clim').

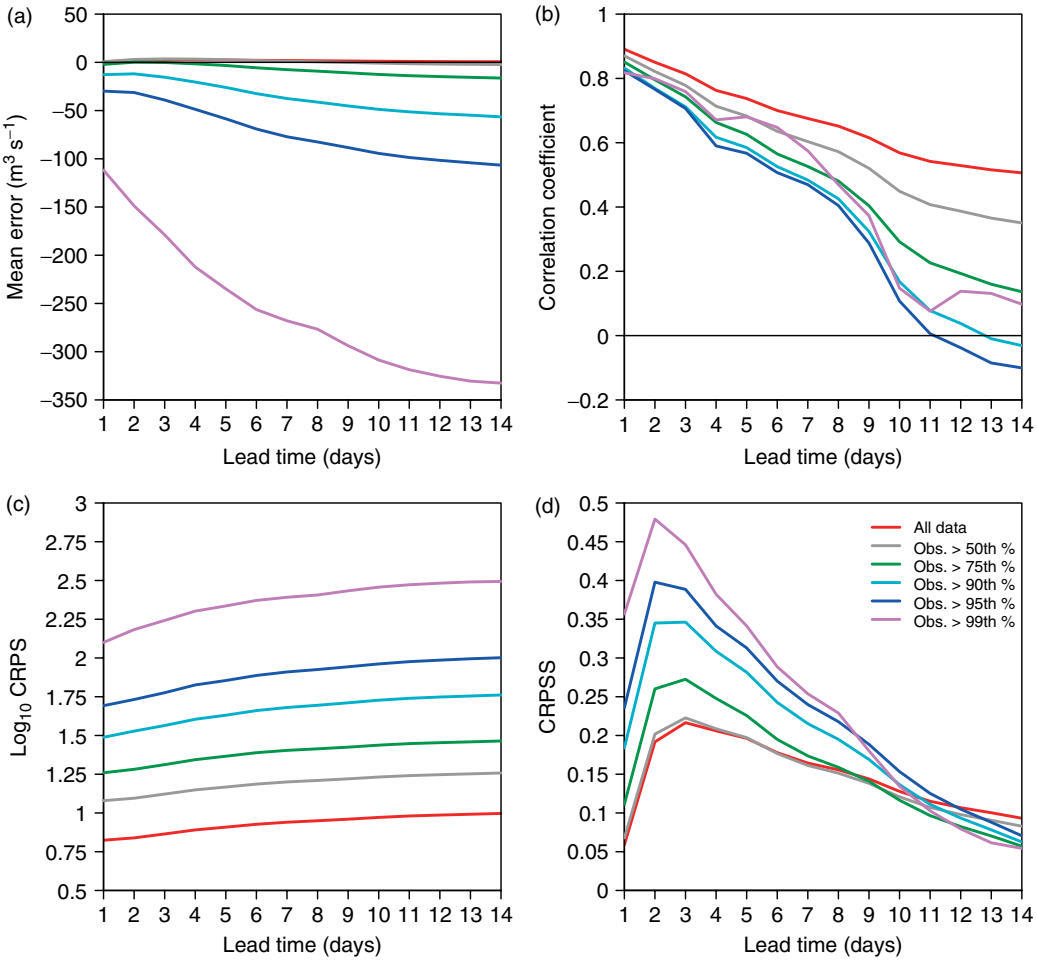

Figure 4. Mean Error (a) and Correlation Coefficient (b) of the ensembles means, as well as CRPS (c) and CRPSS (d) (in reference to climatology-based flow ensembles) for the GFS-based flow ensembles.

4. Results

The daily precipitation totals are verified for all the forecast-observed pairs (8660 pairs for the first 24-h lead time) and for different subsets of pairs defined by the observation exceeding 0 mm/day (i.e. probability of precipitation, PoP), 5 mm/day, 12.5 mm/day, 25 mm/day, and 50 mm/day. The last three thresholds correspond to non-exceedance probabilities in the climatological probability distribution of approximately 0.9, 0.94, and 0.98, respectively.

In Figure 1(a)–(c), the mean error and the correlation coefficient for the ensemble means, as well as the CRPS reflect the decreasing forecast quality with increasing lead time and with increasing observed precipitation amount for the GFS-based precipitation ensembles. Regarding the CRPSS [Figure 1(d)], the GFS-based ensemble forecasts have more skill than the resampled climatological ensembles at all lead times, with a larger gain for high precipitation events compared to low precipitation events. The skill score is slightly negative for the lower thresholds (when excluding the no-rain events) after Day 9, showing that the GFS-based ensembles are not skillful for the small precipitation events beyond this forecast horizon. However, the GFS-based ensembles clearly outperform the resampled climatological ensembles for the prediction of PoP at all lead times.

The modified box plot given in Figure 2 for the 24-h lead time shows the distribution of the errors in the ensemble members against the corresponding observed amount, arranged by increasing observed amount to help detect potential conditional bias. The forecast error (ensemble member – observation) is represented with a box-and-whisker diagram for the 0, 25, 50, 75, and 100 percentiles of the forecast error distribution, where the box corresponds to the interquartile range. The GFS-based precipitation ensembles exhibit a conditional bias, which increases with forecast lead time, as the mean error on Figure 1(a) also indicates: they tend to over-forecast small precipitation amounts and under-forecast large precipitation amounts. However, this conditional bias is much reduced than the bias in the resampled climatological ensembles (not shown), especially for high precipitation events.

In Figure 3, verification statistics for the GFS-based precipitation ensembles and the resampled climatological ensembles are compared against each other for all forecast-observed pairs and two subsets of pairs. Figure 3(b) shows that the GFS-based ensembles improve the CRPS Reliability compared to the resampled climatological ensembles, as expected from

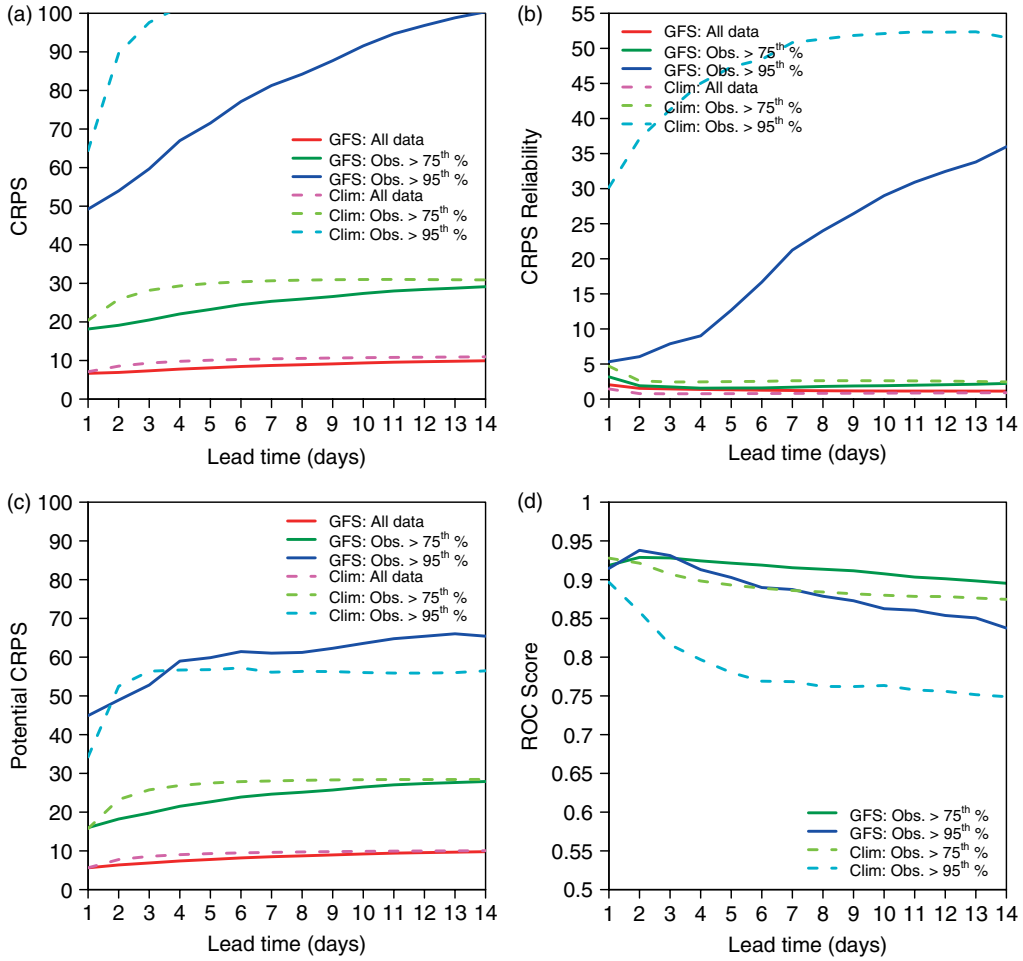

Figure 5. CRPS (a), CRPS Reliability (b), Potential CRPS (c), and ROC Score (d) for the GFS-based flow ensembles ('GFS') and the climatology-based flow ensembles ('Clim').

the reduced conditional bias in the GFS-based ensembles. For the intermittency threshold, the reliability is improved up to Day 12. For the >25 mm/day precipitation threshold, this relative improvement starts from 75% at Day 1 to reach 17% at Day 14. Regarding the Potential CRPS [Figure 3(c)], the GFS-based ensembles outperform the resampled climatological ensembles at all lead times for all the forecast-observed pairs, and until Day 7 when excluding the no-rain events, due to their narrower spread for small precipitation events. However for the >25 mm/day precipitation threshold, the GFS-based ensembles have worse Potential CRPS due to their larger ensemble spread. Therefore, for the high precipitation events, Figures 3(b) and (c) show that GFS-based ensembles exhibit better CRPS than climatology-based ensembles due to improved reliability. Regarding the ROC Score [Figure 3(d)], the forecast discrimination is much improved with the GFS-based ensembles compared to resampled climatological ensembles, especially for the prediction of PoP; this gain decreases with lead time, as expected.

Daily average flow ensembles are verified for all forecast-observed pairs (8660 pairs for the first 24-h lead time) and subsets of pairs based on the following non-exceedance probability thresholds (defined from the 24-year observation record): 0.5 (7 m^3 s^{-1}), 0.75 (30 m^3 s^{-1}), 0.9 (60 m^3 s^{-1}), 0.95 (84 m^3 s^{-1}), and 0.99 (210 m^3 s^{-1}). As indicated in Figure 4(a)–(c), the forecast quality decreases with increasing flow thresholds and with lead time. The GFS-based flow ensembles exhibit a conditional bias consistent with the conditional bias of the precipitation ensembles: over-forecasting of small events and under-forecasting of large events. Regarding the CRPSS [Figure 4(d)] in reference to the climatology-based flow ensembles, the GFS-based flow ensembles are more skillful at all forecast horizons and their skill at individual lead times increases with the flow thresholds until Day 10. The sharp increase in skill between Day 1 and Day 2 is due to the basin response time to precipitation amount.

The influence of the atmospheric ensembles on the flow forecasts is more pronounced after Day 1 as indicated in Figure 5. Because these flow ensembles do not capture any hydrologic uncertainty, both sets of flow ensembles are less reliable at Day 1 except for the very high flows [Figure 5(b)]. Figures 5(b)–(d) show that the GFS-based flow ensembles outperform the climatology-based ensembles in terms of the CRPS Reliability, the Potential CRPS, and the ROC Score for all lead times and all flow thresholds, except the Potential CRPS for the >0.95 non-exceedance

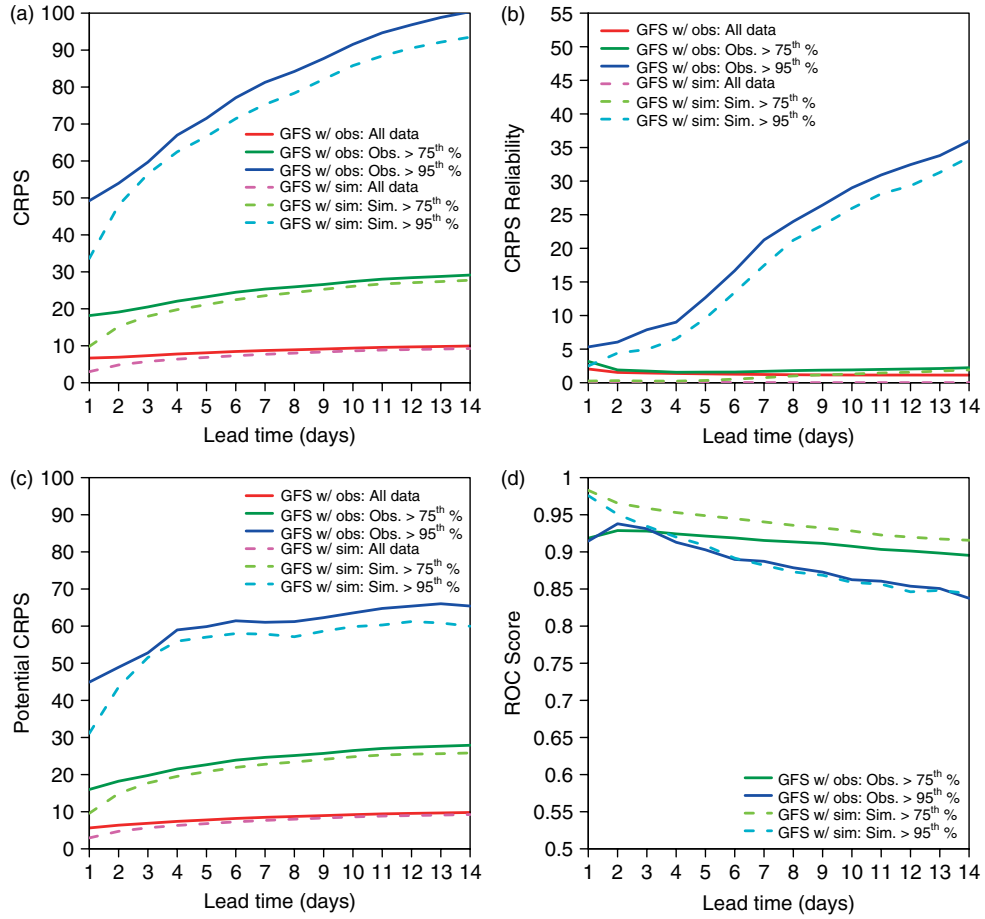

Figure 6. CRPS (a), CRPS Reliability (b), Potential CRPS (c), and ROC Score (d) for the GFS-based flow ensembles verified with observed flows ('GFS w/obs') and simulated flows ('GFS w/sim').

probability threshold. Similarly to the precipitation results shown in Figure 3, for the very high flows, the GFS-based flow ensembles have better CRPS than the climatology-based flow ensembles due to improved reliability (for the >0.95 non-exceedance probability threshold, the relative improvement in reliability varies from 86% at Day 1 to 32% at Day 14).

Regarding the relative contribution of the atmospheric and hydrologic uncertainties, verification statistics are presented in Figure 6 for the GFS-based flow ensembles verified with observed flows (solid lines) and with simulated flows (dashed lines). The forecasts verified with flows simulated from observed hydrometeorological inputs exclude the hydrologic uncertainty (and the observed hydrometeorological input uncertainty), whereas the verification with observed flows includes all sources of uncertainty. Note that the 0.75 and 0.95 non-exceedance probability thresholds correspond to similar flow values for both the observations and the simulations. All four verification statistics indicate that the hydrologic uncertainty is more significant for short lead times and depends on the flow values: for example, for the CRPS (Figure 6(a)), it significantly degrades the score up to Day 7 for all flows and up to Day 2 for very high flows. This indicates that uncertainty in hydrologic initial conditions is a major source of the hydrologic error. However, because of the other hydrologic uncertainty sources (e.g. the model structure and parameters), the hydrologic error tends to degrade the CRPS Reliability, the Potential CRPS, and the ROC Score at all lead times.

5. Conclusions and future work

A strategy for diagnostic verification of hydrologic ensembles is proposed, based on the selection of summary verification metrics and the analysis of the relative contribution of the different sources of error. Such verification could be performed using the EVS software and was illustrated in a case study for experimental ensembles from the HEFS. The results show that the improvement of using the NWP single-valued forecasts in the HEFS ensemble preprocessor (*vs* climatological inputs) for ensemble streamflow prediction is mostly due to improved reliability for very high events. The relative impact of the hydrologic uncertainty is significant for short lead times due to the uncertainty in hydrologic initial conditions. Additional verification studies are underway to include ensembles produced from the HEFS components that account for the hydrologic uncertainty and for other forecast locations to help target future improvements

of the forecasting system and show the value of such improvements to forecasters and users. These verification studies include more detailed verification statistics (including statistics conditioned on the forecast) and more user-oriented verification statistics for operational forecasting. Also planned enhancements to the EVS include the ability to separate the timing (phase) and amplitude errors in hydrologic forecasts. Furthermore, the OHD, the NCEP, and the NWS forecasters are working together and with users to develop meaningful verification products and capabilities to effectively help forecasters and external users in their decision making.

This paper aims to motivate the meteorological and hydrologic research and operations communities for collaborative research and development of verification capabilities and services to generate and communicate verification information for weather, climate, and water forecasts at the catchment scale. One such initiative is the cross-cutting HEPEX verification test-bed, for which the EVS is proposed as one of the verification tools. This verification test-bed aims to address the following challenges in hydrologic ensemble verification: verification of rare events, characterization of the timing error, definition of an optimal set of reference forecasts for skill evaluation, definition of quality measures to be easily integrated in forecasters' and end users' decision process, and development of appropriate methods for multivariate forecasts (e.g. forecasts issued for multiple locations and time steps) and methods to analyze forecast predictability on multiple spatial and temporal scales.

Acknowledgements

The support of this work by the Advanced Hydrologic Prediction Service (AHPS) Program of the National Weather Service (NWS) is gratefully acknowledged. The authors thank Robert Hartmann of the NWS California-Nevada River Forecast Center (CNRFC), Sacramento, California, for providing the hydrometeorological and hydrologic data used in this work.

References

Brown JD, 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. Accepted for publication in *Environmental Modelling and Software*.

Clarke M, Gangopadhyay S, Hay L, Rajagopalan B, Wilby R. 2004. The Schaake Shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology* **5**: 243–262.

Cloke HL, Pappenberger F. 2009. Ensemble flood forecasting: a review. *Journal of Hydrology* **375**: 613–626.

Demargne J, Mullusky M, Werner K, Adams T, Lindsey S, Schwein N, Marosi W, Welles E. 2009. Application of Forecast Verification Science to Operational River Forecasting in the U.S. National Weather Service. *Bulletin of the American Meteorological Society* **90**(6): 779–784.

Demargne J, Wu L, Seo DJ, Schaake J. 2007. Experimental hydrometeorological and hydrologic ensemble forecasts and their verification in the U.S. National Weather Service. *Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management (Proceedings of Symposium HS2004 at IUGG2007, Perugia, July 2007). IAHS Publication* **313** 177–187.

Gupta HV, Beven KJ, Wagener T. 2005. Model calibration and uncertainty estimation. In, *Encyclopedia of Hydrological Sciences*, Anderson M (ed.) John Wiley & Sons, Ltd.: Chichester, UK; 2015–2032.

Hamill TM, Whittaker JS, Mullen SL. 2006. Reforecasts: an important data set for improving weather predictions. *Bulletin of the American Meteorological Society* **87**(1): 33–46.

Hersbach H. 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting* **15**: 559–570.

Jolliffe IT, Stephenson DB. 2003. *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons Ltd: Chichester, UK.

National Research Council of the National Academies (NRC). 2006. Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts. http://www.nap.edu/, accessed 01/21/10.

Pappenberger F, Scipal K, Buizza R. 2008. Hydrological aspects of meteorological verification. *Atmospheric Science Letters* **9**: 43–52.

Schaake J, Demargne J, Hartman R, Mullusky M, Welles E, Wu L, Herr H, Fan X, Seo DJ. 2007. Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth Systems Sciences Discussions* **4**: 655–717.

Welles E, Sorooshian S, Carter G, Olsen B. 2007. Hydrologic verification: A call for action and collaboration. *Bulletin of the American Meteorological Society* **88**: 503–511.

Wilks DS. 2006. *Statistical Methods in Atmospheric Sciences* (2nd ed.). Academic Press: San Diego.