HEFS extends hydrologic ensemble services from 6-hour to year-ahead forecasts and includes additional weather and climate information as well as improved quantification of major uncertainties.

As no forecast is complete without a description of its uncertainty (National Research Council of the National Academies 2006), it is necessary, for both atmospheric and hydrologic predictions, to quantify and propagate uncertainty from various sources in the forecasting system. For informed risk-based decision making, such integrated uncertainty information needs to be communicated to forecasters and users effectively. In an operational environment, ensembles are an effective means of producing uncertainty-quantified forecasts. Ensemble forecasts can be ingested in a user’s downstream application (e.g., reservoir management decision support system) and used to derive probability statements about the likelihood of specific future events (e.g., probability of exceeding a flood threshold). Atmospheric ensemble forecasts have been routinely produced by operational Numerical Weather Prediction (NWP) centers for two decades. Hydrologic ensemble forecasts for long ranges have been initially based on historical observations of precipitation and temperature as plausible future inputs (e.g., Day 1985) in an attempt to account for the uncertainty at the climate time scales. Ensemble forecasts generated in this fashion were considered viable beyond 30 days where the climatic uncertainty would dominate other uncertainty sources. More recently, as the needs for risk-based management of water resources and hazards across weather and climate scales have increased, the research and operational communities have been actively working on integration of the NWP ensembles into hydrologic ensemble prediction systems and quantification of all major sources of uncertainty in such systems. In particular, the Hydrological Ensemble Prediction Experiment (HEPEX; www.hepex.org/), launched in 2004, has facilitated communications and collaborations among the atmospheric community, the hydrologic community, and the forecast users toward improving ensemble forecasts and demonstrating their utility in decision making in water management (Schaake et al. 2007b; Thielen et al. 2008; Schaake et al. 2010).

Ensemble approaches hold great potential for operational hydrologic forecasting. As demonstrated with atmospheric ensemble forecasts, the estimates of predictive uncertainty provide forecasters and users with objective guidance on the level of confidence that they may place in the forecasts. The end users can decide to take action based on their risk tolerance. Furthermore, by modeling uncertainty, hydrologic forecasters can maximize the utility of weather and climate forecasts, which are generally highly uncertain and noisy (Buizza et al. 2005). With the major uncertainties quantified and their relative importance...
analyzed, ensemble forecasting helps identify areas where investments in forecast systems and processes will have the most impact.

Development and implementation of hydrologic ensemble prediction systems is still ongoing and hence only limited operational experience exists. A current status of research using experimental and (pre)operational systems, however, has demonstrated their potential benefits (see, e.g., Cloke and Pappenberger 2009 and Zappa et al. 2010 for references). Recent verification studies of hydrologic ensemble forecasts, however, have demonstrated that forecasts that are retroactively generated using a fixed forecasting system over long time periods include Barthelemy et al. (2009), Jaun and Ahrens (2009), Renner et al. (2009), Hopson and Webster (2010), Demargne et al. (2010), Thiret et al. (2010), Van den Bergh and Roulin (2010), Addor et al. (2011), and Zappa et al. (2012) for short- to medium-range hydrologic forecasts and Kang et al. (2010), Wood et al. et al. (2011), Fundel et al. (2012), Singla et al. (2012), and Yuan et al. (2013) for monthly to seasonal hydrologic ensembles. Objective verification analysis of ensemble forecasts or hindcasts over multiple years should improve not only the science of hydrologic ensemble forecasting but also the utility of hydrologic ensemble forecast products in various downstream applications where decision support systems could be “harnessed” (van Andel et al. 2008).

In the National Oceanic and Atmospheric Administration (NOAA)’s National Weather Service (NWS), an end-to-end Hydrologic Ensemble Forecasting Service (HEFS) is currently being implemented as part of the Advanced Hydrologic Prediction Service (AHPS; McEnery et al. 2005) to address a variety of water information and service needs for flood risk management, water supply management, streamflow regulations, recreation planning, ecosystem management, and others (Raff et al. 2013). Such a wide range of applications requires forcing inputs and hydrologic forecasts at multiple space–time scales and for multiple forecast horizons: from minutes for flash flood predictions in fast responding basins to years for water supply forecasts over larger areas (see examples in McEnery et al. 2005). To account for the forcing input uncertainty, the NWS River Forecast Centers (RFCs) have been using the Ensemble Streamflow Prediction (ESP) component of the National Weather Service River Forecast System (NWSRFS; National Weather Service 2012). ESP produces seasonal probabilistic forecasts of water supply based on the historical observations of precipitation and temperature (the climate being considered stationary and deterministic) and the current hydrologic conditions (Day 1985). HEFS enhances ESP to include short-, medium-, and long-range forcing forecasts, incorporate additional weather and climate information, and better quantify the major uncertainties in hydrologic forecasting. The HEFS provides ensemble forecast and verification products and adds a major new capability to the NWS’s baseline river forecasting system: the Community Hydrologic Prediction System (CHPS).

Overview of the HEFS: Uncertainty in hydrologic predictions comes from many different sources: atmospheric forcing observations and predictions; initial conditions of the hydrologic model, its parameters, and structure; and streamflow regulations among other things. For HEFS, however, uncertainty in atmospheric forcing inputs is typically referred to as input uncertainty and those in all other sources as structural uncertainty (Krzysztofowicz 1999). A hydrologic ensemble prediction system could either model the total uncertainty in the hydrologic output forecasts (e.g., Montanari and Grossi 2008; Coccia and Todini 2011; Weerts et al. 2011; Smith et al. 2012; Regonda et al. 2013) or explicitly account for the major sources of uncertainty, which is the primary approach of HEFS (Seo et al. 2006). As noted by Velázquez et al. (2011), hydrologic ensemble prediction systems presented in the literature often account for the input uncertainty only. Recently, a few systems have included various techniques to address specific hydrologic uncertainties, such as hydrologic data assimilation to reduce and model initial conditions’ uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007). Other uncertainty models include hydrologic uncertainty, Monte Carlo–based techniques to estimate model parameter uncertainty, and postprocessing and multimodel ensemble approaches for hydrologic structural uncertainty modeling (see references in Seo et al. 2007).
forecasting, as well as hindcasting to provide the large sample of events necessary to verify forecast procedures and determine model bias and uncertainty. Comprehensive evaluation of the individual HEFS components as well as the end-to-end system via multiyear hindcasting is underway (Brown 2013; illustrative examples of verification results are presented in this paper.)

In the context of operational hydrologic forecasting in the NWS, the HEFS has been developed to improve upon operational single-valued forecasting and seasonal ESP forecasting while capturing user requirements, which include 1) supporting both real-time ensemble forecasting and hindcasting for large-sample verification and systematic evaluation, 2) maintaining interoperability with the single-valued forecasting system for the short range (given that single-valued forecasting is only a special case of ensemble forecasting), and 3) producing ensemble forecast information that is statistically consistent over a wide range of spatiotemporal scales. The operational hydrologic and water resources models used for both single-valued and probabilistic forecasting are simple conceptual models applied in a lumped fashion, with relatively few parameters estimated by manual calibration (a unique set of parameters being defined for each flow to flooding condition). Expectedly, the hydrologic predictability could be limited in poorly monitored areas, with river gauges malfunctioning (e.g., during flood events) and during rapidly changing hydrometeorological conditions. Also, the estimation of historical information for the RFC forecasters and changes of calibration (a unique set of parameters being defined with relatively few parameters estimated by manual conceptual models applied in a lumped fashion, single-valued and probabilistic forecasting are simple models, as well as enhancements made within the Delft- FEWS community for hydrologic and water resources forecast systems and services.

**METEOREOLOGICAL ENSEMBLE FORECAST PROCESSOR** Reliable and skillful atmospheric ensemble forecasts are necessary for hydrologic ensemble forecasting. Ensemble forecasts from current and past model runs of several atmospheric prediction centers. However, these ensembles are generally biased in the mean, spread, and higher moments (Buizza et al. 2005), both unconditionally on season, storm type, and other attributes. The conditional biases may be particularly large for heavy precipitation events that are crucial in flood forecasting (Hamill et al. 2006; Brown et al. 2012). There are several statistical techniques for estimating the conditional probabil- ity distribution of an (assumed unbiased) observed variable given a potentially biased forecast (see, e.g., references in Brown et al. 2012). These techniques vary in their assumptions about the conditional (or joint) distribution, e.g., marginal model outputs are processed slightly differently since precipitation is intermittent and highly skewed whereas the temper- ature distribution is nearly Gaussian. MEFP uses the normal quantile transform (NQT) to transform observed and forecast precipitation variables into normal variates. The precipitation part of MEFP also includes an explicit treatment of precipitation dependence structure from the ensemble mean (e.g., Hamill et al. 2004; Wilks and Hamill 2007). Therefore, the MEFP uses the single-valued forecasts modified by human forecasters for short-range forecast horizon (up to 7 days) and the mean forecasts from multiple NWP models for mid- to long-range to generate seamless and calibrated hydrometeorological ensembles up to a 1 yr forecast horizon. Precipitation and temperature are processed slightly differently since precipitation is variable and highly skewed whereas the temperature distribution is nearly Gaussian. MEFP uses the normal quantile transform (NQT) to transform observed and forecast precipitation variables into normal variates. The precipitation part of MEFP also includes an explicit treatment of precipitation intermittency using the mixed-type bivariate meta- Gaussian model (Herr and Krzysztofowicz 2005), parametric and nonparametric modeling of the marginal probability distributions, and a parameter optimization under the continuous ranked probability score (CRPS; Hersbach 2000) and other criteria (see Wu et al. 2011 for details). For temperature, the MEFP procedure first generates ensembles of daily marginal models from the conditional distributions (and generates ensembles at subdaily time steps from the daily ensembles through a diurnal variation model.

The above scheme is based on the same interpo- lation procedures used to calculate subdaily historical ensembles (Hamill et al. 2005; Wilks and Hamill 2007; Hamill et al. 2008). Bias correction of precipitation ensemble forecasts is particularly challenging because precipitation amount is intermittent, it depends strongly on space-time scale, and is relatively unpredictable in many cases (e.g., convective events). For hydrologic forecasting with lumped models, the gridded NWP ensembles need to be processed at the basin scale, which requires “downscaling” (described as a change of support in geostatistics) and bias correction. This downscaling includes corrections to match the climatology of the forcings used to calibrate the hydrologic model. The MEFP aims to generate unbiased ensembles that capture the skill of the forecasts from multiple sources for individual basins while preserving the space–time properties of hydrometeorological vari- ables (e.g., precipitation and temperature) across all basins (Schaake et al. 2007a; Wu et al. 2011). For short-range forecasts, human forecasters generally add significant value to single-valued hydrometero- logical forecasts derived from raw NWP forecasts (Charba et al. 2003). Also, postprocessing studies have repeatedly demonstrated that most information from the ensemble comes from the ensemble mean (e.g., Hamill et al. 2004; Wilks and Hamill 2007). Therefore, the MEFP uses the single-valued forecasts modified by human forecasters for short-range forecast horizon (up to 7 days) and the mean forecasts from multiple NWP models for mid- to long-range to generate seamless and calibrated hydrometeorological ensembles up to a 1 yr forecast horizon. Precipitation and temperature are processed slightly differently since precipitation is variable and highly skewed whereas the temperature distribution is nearly Gaussian. MEFP uses the normal quantile transform (NQT) to transform observed and forecast precipitation variables into normal variates. The precipitation part of MEFP also includes an explicit treatment of precipitation intermittency using the mixed-type bivariate meta- Gaussian model (Herr and Krzysztofowicz 2005), parametric and nonparametric modeling of the marginal probability distributions, and a parameter optimization under the continuous ranked probability score (CRPS; Hersbach 2000) and other criteria (see Wu et al. 2011 for details). For temperature, the MEFP procedure first generates ensembles of daily marginal models from the conditional distributions (and generates ensembles at subdaily time steps from the daily ensembles through a diurnal variation model.

The above scheme is based on the same interpo- lation procedures used to calculate subdaily historical ensembles (Hamill et al. 2005; Wilks and Hamill 2007; Hamill et al. 2008). Bias correction of precipitation ensemble forecasts is particularly challenging because precipitation amount is intermittent, it depends strongly on space-time scale, and is relatively unpredictable in many cases (e.g., convective events). For hydrologic forecasting with lumped models, the gridded NWP ensembles need to be processed at the basin scale, which requires “downscaling” (described as a change of support in geostatistics) and bias correction. This downscaling includes corrections to match the climatology of the forcings used to calibrate the hydrologic model. The MEFP aims to generate unbiased ensembles that capture the skill of the forecasts from multiple sources for individual basins while preserving the space–time properties of hydrometeorological vari- ables (e.g., precipitation and temperature) across all basins (Schaake et al. 2007a; Wu et al. 2011). For short-range forecasts, human forecasters generally add significant value to single-valued hydrometeor- logical forecasts derived from raw NWP forecasts (Charba et al. 2003). Also, postprocessing studies have repeatedly demonstrated that most information from the ensemble comes from the ensemble mean (e.g., Hamill et al. 2004; Wilks and Hamill 2007). Therefore, the MEFP uses the single-valued forecasts modified by human forecasters for short-range forecast horizon (up to 7 days) and the mean forecasts from multiple NWP models for mid- to long-range to generate seamless and calibrated hydrometeorological ensembles up to a 1 yr forecast horizon. Precipitation and temperature are processed slightly differently since precipitation is variable and highly skewed whereas the temperature distribution is nearly Gaussian. MEFP uses the normal quantile transform (NQT) to transform observed and forecast precipitation variables into normal variates. The precipitation part of MEFP also includes an explicit treatment of precipitation intermittency using the mixed-type bivariate meta- Gaussian model (Herr and Krzysztofowicz 2005), parametric and nonparametric modeling of the marginal probability distributions, and a parameter optimization under the continuous ranked probability score (CRPS; Hersbach 2000) and other criteria (see Wu et al. 2011 for details). For temperature, the MEFP procedure first generates ensembles of daily marginal models from the conditional distributions (and generates ensembles at subdaily time steps from the daily ensembles through a diurnal variation model.)
by the output dimensionality. We therefore suggest examining in the future alternative approaches, which are more robust and less sensitive to the output dimensionality. For example, methods 2 and 3 model explicitly the precipitation intermittency (Schaake et al. 2007a); methods 2 and 3 model explicitly the precipitation intermittency, with methods 3 and 3 adding parameter optimization based on the CRPS. Since, for single-valued forecasts, the CRPS collapses to the mean absolute error, the CRPS values are compared to the mean absolute error of the conditioning single-valued forecast. Figure 2 indicates that the quality of MEFP-generated precipitation ensembles has improved significantly with the explicit intermittency modeling and that the technique captures the skill in the conditioning single-valued forecast very well. In Fig. 3 (from Wu et al. 2011), the reliability diagram and the relative operating characteristic (ROC) curve for method 3 indicate that the MEFP precipitation ensembles are reliable (left plot) and capture very well the discriminatory skill (right plot) in the single-valued precipitation forecasts. Wu et al. (2011) also show that independent validation results (based on leave-one-year-out cross validation) are similar to dependent validation (i.e., parameter estimation) when using a historical archive of about 6 years. This suggests that the MEFP calibration with a similar or longer record length allows realizing in real-time applications the level of performance obtained in dependent validation, even if some degradation may be expected for rare events.

For medium range (up to 14 forecast lead days), the single-valued forecasts are obtained as the ensemble means from the frozen version (circa 1998) of the Global Forecast System (GFS; Hamill et al. 2006) of the National Weather Service’s National Centers for Environmental Prediction (NCEP). A 30-yr reforecast archive at T62 resolution and stored on a 2.5° grid is available for the MEFP calibration. Verification results of GFS-based forcing ensembles are described in Schaake et al. (2007a) and Demargne et al. (2010) for the NFDC1 basin.

Figure 4 shows the verification results from dependent validation for the 14-day GFS-based precipitation ensembles compared to the climatology-based precipitation ensembles. All ensembles were produced at a 6-h time step from 2000 to 2005, with 45 ensemble members, using method 3, and were verified with the EVS as daily totals. The mean error is reported for the ensemble mean (commonly used as a single-valued representation of the ensemble in operational forecasting), along with the continuous ranked probability skill score (CRPSS), which describes the overall quality of the probabilistic forecast in reference to the climatology-based ensembles. At short lead times, the precipitation ensembles are relatively unbiased in the unconditional sense as evidenced by the mean error for the precipitation intermittency threshold; however, they underforecast for larger observed events. These Type-II conditional biases are common in ensemble forecasting systems since model calibration typically favors average conditions and such conditional biases are more difficult to remove with postprocessing (Brown et al. 2012). When comparing to the climatology-based ensembles for all 14 lead days, the MEFP-generated ensembles expectedly showed reduced conditional biases. The quality of GFS-based precipitation decreases rapidly with increasing forecast lead time as evidenced by the increased mean

![Image](https://example.com/image.png)
error and the reduction in CRPSS. However, because of the relatively large predictability of orographic precipitation in the Sierra Nevada during the cold season in particular, GFS-based ensembles show useful skill in terms of CRPSS until 10 days.

As part of the ongoing comprehensive evaluation of HESS ensembles, Browne (2013) analyzed verification results of GFS-based precipitation and temperature ensemble hindcasts for a 14-day forecast horizon for four pairs of headwater–downstream test basins located in California, Colorado, Kansas–Oklahoma, and Pennsylvania–New York. The GFS-based precipitation ensembles generally show skill against climatology-based ensembles for the first week but little or no skill in the second week. However, results vary significantly with basin locations (e.g., reduced precipitation predictability in the southern plains), seasons (e.g., less skill during the dry season), and magnitudes (e.g., underestimation of the probability of precipitation and, more problematically, large precipitation amounts), which underlines the need for a systematic and comprehensive evaluation of MEEP ensembles across the different RFCs.

MEEP has recently been enhanced to ingest forecast from the NCEP’s latest Global Ensemble Forecast System (GEFS), which was implemented in February 2012. The new version of the GEFS uses the latest GFS model version v9.0 with an increased horizontal resolution of T254 (~55 km) for 8 days and an improved vertical resolution for all 16 days; it also includes uncertainty modeling enhancements (see Wei et al. 2008 and Hou et al. 2013, manuscript submitted to Tellus, for details). A new 25-yr ensemble reforecast dataset has been completed by using the configuration of the current operational GFS and is available for public access (Hamill et al. 2013). For the longer range, the MEEP ingests single- valued forecasts from the NCEP’s Climate Forecast System (GFS version 2; Saha et al. 2013), which has been operational since February 2011 and has shown skill against climatology for hydrological ensemble forecasting (e.g., Yuan et al. 2013). MEEP constructs lagged ensemble forecasts from the single- valued GFS forecasts to estimate the ensemble mean (used as single- valued forecast to drive the MEEP statistical model) for a forecast horizon up to 9 months. MEEP requires long hindcast datasets of weather and climate forecasts from a fixed model to correct biases in the single-valued forecasts, particularly for rare events. Several studies have demonstrated that utilizing the reforecast dataset from the frozen version of a NWP model significantly improves the skill of temperature and precipitation forecasts (in particular for heavy precipitation events), as well as precipitation forecasting (Werner et al. 2005; Hamill et al. 2006; Wilks and Hamill 2007; Hamill et al. 2008).

Validation of short- to long-term ensembles for various RFC basins is underway to evaluate the expected performance of MEEP for producing seamless and skillful ensembles.

In the future, MEEP should include forecasts from other NWP models (e.g., the Short-Range Ensemble Forecast System (SREF) system produced by the NCEP (Du et al. 2009), techniques to estimate precipitation from the combination of different NWP model output variables (e.g., total column precipitable water), and additional and/or alternative postprocessing techniques, for example, to incorporate information from the ensemble spread and higher moments (Brown and Seo 2010). In the experimental Meteorological Model- Based Ensemble Forecast System (Philpott et al. 2012), three Eastern Region RFCs and a Southern Region RFC are also investigating the use of SREF and GEFS ensembles, as well as North American Ensemble Forecast System (NAEFS) ensembles, all produced and bias corrected (at the grid scale) by the NCEP (Cui et al. 2012) (experimental products available at www.arh.noaa.gov/nmesfs/). Grand-ensemble datasets such as The Observing System Research and Predictability Experiment (THORPEX) Interactive Global Ensemble (TIGGE; Park et al. 2008) have significant potential to capture uncertainties in the initial conditions, the model parameterizations, and the data assimilation technique, and the model structure through the use of atmospheric ensembles from different NWP models (e.g., Pappenberger et al. 2008; He et al. 2009, 2010). However, the use of any NWP model ensembles in hydrologic modeling requires a long reforecast dataset in order to calibrate the meteorological ensemble forecast processor as well as the hydrologic and water resources models for rare events.

**HYDROLOGIC ENSEMBLE POSTPROCESSING.** Sources of hydrologic bias and uncertainty may be unknown or poorly specified in hydrologic ensemble prediction systems. Therefore, a range of statistical postprocessing techniques have been developed to account for the collective hydrologic uncertainty (Krzyztofowicz 1999; See et al. 2006; Coccia and Todini 2011; Brown and Seo 2013; He et al. 2009, 2010). The use of any NWP model ensemble forecasts in hydrologic modeling requires a long reforecast dataset in order to calibrate the meteorological ensemble forecast processor as well as the hydrologic and water resources models for rare events.

In the HEFS, the EnsPost (See et al. 2006) hydrologic uncertainty model estimates the uncertainty in a lumped form. Since MEEP generates bias-corrected hydrometeorological ensembles that reflect the input uncertainty, EnsPost is calibrated with simulated streamflow (i.e., generated from perfect future meteorological forcing) without any manual modifications of model states and parameters. The hydrologic uncertainty is, therefore, modeled independently of forecast lead time. The postprocessed streamflow ensemble results from integration of the input and hydrologic uncertainties and hence reflect the total uncertainty. The current version of the EnsPost employs a parsimonious statistical model that combines probability matching and time series modeling. Parsimony is important to reduce data requirements and, therefore, the sample uncertainty of the estimated parameter values. The procedure adjusts each ensemble trace via recursive linear regression in the normal space (see Seo et al. 2006 for details). The regression is a first-order autoregressive model with an exogenous variable, or ARX(1,1), and uses normal-quantile-transformed historical simulation and verifying observations. The regression parameter is optimized for different seasons, and different flow conditions are taken into account that the correlation depends greatly on flow magnitude and season. Recently, this model...
has been modified to better simulate temporal variability in the postprocessed streamflow ensembles by accounting for dependence in the normal space between the residual error of the model fit and the observed streamflow, as well as the serial correlation in the residual error.

EnsPost is currently applied to daily observed and forecast streamflows; after statistical postprocessing, the adjusted ensemble values are disaggregated to subdaily flows. In See et al. (2006) and subsequent studies for other locations, the EnsPost shows satisfactory results for short forecast horizons and for all ranges of flow. However, independent validation shows slightly degraded results in comparison to dependent validation when EnsPost parameters were estimated from a 20-yr record, mainly owing to uncertainties in the empirical cumulative distribution functions of observed and simulated flows. See et al. (2006) underlined that in real-time applications, when the postprocessor parameters may be regularly (e.g., annually) updated using more than 20 years of data, the performance of EnsPost would be similar or better than the obtained independent validation results. Examples of cross validation results are shown in Fig. 5 for postprocessed flow ensemble hindcasts produced with perfectly known future forcing. The (left) Reliability diagrams and (right) ROC curves relative to the 0.95 nonexceedance probability threshold show good reliability (left plot) and discriminatory skill similar to the single-valued model predictions (right plot) for the first and fifth lead days. However, the current version of EnsPost is of limited utility for complex flow regulations and does not explicitly account for timing errors in the streamflow simulations (see Lin et al. 2011). Regarding the quality of HEFS flow ensembles, examples of dependent verification results are given for raw and postprocessed flow ensemble hindcasts produced by the Hydrologic Processor and EnsPost using the GFS-based precipitation and temperature ensembles generated by MEFP. For flow hindcasting, the Hydrologic Processor is first run in simulation mode with the observed precipitation and temperature time series to generate the historical initial conditions for all hindcast dates. Based on the historical initial conditions, flow ensemble hindcasts are produced by the Hydrologic Processor for each hindcast date using the MEFP precipitation and temperature ensemble hindcasts, and then postprocessed by EnsPost. To evaluate the performance gain using MEFP and EnsPost, flow ensembles produced by the Hydrologic Processor (using the same retrospective initial conditions) from climatological forcing ensembles are used as reference forecasts. The example verification results are given for the NFDC1 basin, for which 6-h ensemble hindcasts were produced from 1979 to 2005 and verified with EVS as daily average flows. The comparisons of dependent and independent validation results for MEFP and EnsPost in the previous studies (e.g., Wu et al. 2011; See et al. 2006) have shown their robustness. Thus, the following dependent validation results for HEFS-generated flow ensembles give a reasonable indication of the expected performance of HEFS in real-time applications, when both MEFP and EnsPost are calibrated with more than 25 years of data, even if some degradation is expected for rare events. As illustrated in Figs. 2–4 for the NFDC1 basin, MEFP precipitation ensembles perform well, particularly when compared with climatological ensembles. The marginal value of EnsPost depends largely on the magnitude of the systematic bias in the model-simulated streamflow. For the NFDC1 basin, the model simulation is of very high quality with a volume bias of only about 1%. As such, one may expect the contribution from the EnsPost to be modest, coming mostly from improved reliability by adding spread to the streamflow ensembles. Figure 6 shows the mean error for the ensemble means and the CRPSS for the postprocessed flow ensembles and raw flow ensembles in reference to the climatology-based flow ensembles. The GFS-based flow ensembles exhibit a conditional bias consistent with the conditional bias of the precipitation ensembles: overforecasting of small events and underforecasting of large events. However, owing to hydrologic persistence or “basin memory,” the quality of the flow ensembles declines more slowly than that of the precipitation ensembles. Regarding CRPSS results, the sharp increase in skill between the first and second forecast days is due to the fact that, for the first lead day, the climatology-based flow ensembles too have good skill owing to persistence, which results in reduced skill score for the GFS-based flow ensembles. The comparison of the CRPSS values for the raw flow ensembles and the postprocessed flow ensembles shows that most of the flow forecast skill comes from the MEFP component, with limited impact of EnsPost. The additional improvement by EnsPost is marginal because of small hydrologic biases and uncertainties in this basin. It decreases very fast within the first few days as a reflection of the fast-decaying memory in the initial conditions, noting that the prior observation is a predictor in the EnsPost. However, as pointed out by Brown (2013), the overall skill of GFS-based postprocessed flow ensembles in reference to climatology-based flows, as well as the relative contributions of the MEFP (with GFS forecasts) and EnsPost components, depend on the basin location (as illustrated in Fig. 7 with basins located in four different RFCs), flow amount,
and season. In Fig. 8, examples of ensemble traces for two different basins illustrate how MEEP and EnsPost may help to predict a large flow event 5 days in advance but with a peak timing error (top plots), and may produce ensembles with a much reduced spread compared to climatology-based flows but with a low bias tendency (bottom plot).

The performance of EnsPost depends largely on the availability of long-term observed and model-simulated flows and the assumption that the streamflow climatology is stationary over a multidecadal period. If additional stratification of the observed-simulated flow dataset is necessary for parameter estimation to improve the model fit for specific conditions (e.g., snowmelt), the EnsPost will require an even larger dataset for its calibration. In the future, for areas where observed and simulated flow data are available at subdaily scales (6 hourly or hourly), direct modeling of the subdaily flow will be necessary for improved performance. The use of multiple temporal scales of aggregation to improve bias correction at longer ranges is under investigation. Evaluation of other bias-correction techniques (including those used for atmospheric forcings) is also ongoing (e.g., van Andel et al. 2012) to find the best approaches for different forecasting situations and forecast attributes.

Moreover, EnsPost needs to be currently applied without any manual modifications of model states and parameters to maintain the consistency between the real-time ensemble flows and the simulated flows used for its calibration, as well as the EnsPost-generated streamflow hindcasts and verification results. Therefore, for real-time ensemble prediction, the set of model states used in HEFS is generated with a simulation time window long enough to minimize the impact of any modifications previously applied in single-valued forecasting. Obviously, EnsPost needs to evolve along with the data assimilator component to utilize automated DA procedures. Meanwhile, given that the current manual modifications address significant limitations in the operational models and datasets, we recommend analyzing the potential impact of these modifications on the performance of HEFS flow ensembles. Such comprehensive evaluation could offer guidance on best operational practices for applying manual modifications and cost-effective transitioning of experimental automated DA capabilities into operational ensemble forecasting.

**ENSEMBLE VERIFICATION.** To evaluate the performance of HEFS for both research and operational forecasting purposes, ensemble verification is required. Key attributes of forecast quality include the degree of bias of the forecast probabilities, whether unconditionally or conditionally upon the forecasts (reliability or Type-I conditional bias) or observations (Type-II conditional bias), the ability to discriminate between different observed events (i.e., to issue distinct probability statements), and skill relative to a baseline forecasting system (Jollife and Stephenson 2003; Wilks 2006). Ensemble forecasting systems, such as HEFS, are intended for a wide range of practical applications, such as flood forecasting, river navigation, and water supply forecasting. Therefore, forecast quality needs to be evaluated for a range of observed and forecast conditions in terms of forecast horizon, space–time scale, seasonality, and magnitude of event. The EVS, built on the Ensemble Verification System (Brown et al. 2010; freely available from www.mcs.anl.gov/evs.html), was designed to support conditional verification of forcing and hydrologic ensembles, generated by HEFS, as well as operational forecasts. The EVS is a flexible, modular, and open-source software tool programmed in Java to allow cost-effective collaborative research and development with academic and private institutions and rapid research-to-operations transition of scientific advances.

Key features of EVS include the following (see Brown et al. 2010 for details):

- the ability to evaluate forecast quality for any continuous numerical variable (e.g., precipitation, temperature, streamflow, river stage) at specific forecast locations (points or areas) and for any temporal scale or forecast lead time;
- the ability to evaluate the quality of an ensemble forecasting system conditional upon many factors, such as forecast lead time, seasonality, temporal aggregation, magnitude of event (defined in various ways, such as exceedance of a real-valued threshold or climatological probability), and values of auxiliary variables (e.g., quality of flow ensembles conditional upon the amount of observed precipitation);
- the ability to evaluate key attributes of forecast quality, such as reliability, discrimination, and
skil, at varying levels of detail, ranging from highly summarized (e.g., skill scores such as CRPS or NSE) to highly detailed (e.g., box plots of con-
ditional errors); the ability to aggregate the forecasts in time (e.g., hourly to daily) and to evaluate aggregate perfor-
ance over a range of forecast locations, either by pooling pairs or computing a weighted average of the verification metrics from several locations; generating graphical and numerical outputs in a range of file formats (R scripts are also provided for further analysis and generation of custom graphs); the ability to implement a verification study via the graphical user interface (GUI) or to batch process a large number of forecast locations on the com-
mand line, using a project file in an XML format, the EVS can also be run within CHPS—e.g., to produce diagnostic verification results for one or multiple hindcast scenarios); and the ability to estimate the sampling uncertainty in the verification metrics using the station block bootstrap—synthetic realizations of the original paired data are repetitively generated and the verification metrics are computed for each sample to estimate a bootstrap distribution of the verifica-
tion metrics, from which the percentile confidence intervals are then derived.

EVS is regularly enhanced to address needs from modelers and forecasters as HEFS is being imple-
mented and expanded across all RFCs and since the Ensemble Verification System is being used in other projects such as HEPEX.

GRAPHICS GENERATOR. Communicating uncertainty information to a wide range of end users represents a challenge. As hydrologic ensemble fore-
casting is relatively new, much research is needed to define the most effective methods of presenting such information. A key support system that maximizes their utility (Cloke and Pappenberger 2009). Challenges in communicating hydrologic ensembles include how to understand the ensemble forecast information (e.g., value of the ensemble mean, relation between spread and skill), how to use such information (e.g., in coordination with deterministic forecasts), and how to communicate it (e.g., spaghetti plots versus plume charts), even to nonexperts (Demenit et al. 2010). A variety of prac-
tices for communicating ensemble information have been presented by Bruen et al. (2010) for seven European ensemble forecasting platforms and by Ramos et al. (2007) and

Demerit et al. (2013) for the European Flood Alert System. Pappenberger et al. (2013) formulated rec-
ommendations for effective visualization for both com-
munication of probabilistic flood forecasts among experts, acknowledging that there is no overarching agreement and one-size-fits-all solution.

In HEFS, the Graphics Generator (GraphGen), a generic software tool for CHPS, enables forecasters to generate and visualize information for internal decision support during operations as well as disseminate the final products to end users. This tool is expected to be accessed externally through a web service interface, which will allow the uncertainty-quantified forecast and verification information to be tailored to the needs of specific external users. GraphGen includes the functionality of the NWSRF5 Ensemble Streamflow Prediction Analysis and Display Program (National Weather Service 2012), such as generation of spaghetti plots, expected value chart to describe the ensemble distribution (minimum, maximum, mean, and standard deviation), exceedance probability bar graph for a few probability categories and for a given product (e.g., monthly volume), and exceedance probability distribution plot using current initial conditions as reference to historical simulations (see examples in McEnery et al. 2005). HEFS also needs the ability to aggregate the forecasts in time (e.g., 90% chance to exceed flood impact thresholds) as well as short-term values (e.g., expected value charts for all forecast lead times). New products to visualize the ensemble distribution include such as box-and-whisker plots with quantiles from the ensemble distribution, ensemble consistency tables, and visualization of peak timing uncertainty and magnitude uncertainty. In addition, information is needed to help forecasters and modelers understand the forecast and verify the ensemble forecasts in the context of the estimated uncertainty. Several RFCs make prototype ensemble products and information available to their customers (see http://water.weather.gov/ahps/). These interfaces allow user-specific views to help enhance the operational national web interface for AHPS (http://

water.weather.gov/ahps/). Furthermore, verification information needs to be provided along with forecast information to sup-
port decision making (Demargne et al. 2009). Simi-
lar approaches have been reported by Bartholmes et al. (2009), Renner et al. (2009), and van Andel et al. (2010). These interfaces allow the browsern results to maximize the utility of hydrologic and water resources forecast products and services.

The end-to-end HEFS provides, among other things, long range streamflow and water resources forecast and verification products that are generated by 1) the MEFP, which ingests weather and climate forecasts from multiple Numerical Weather Prediction models to produce seasonal and forecast precipitation and temperature ensembles at the hydrologic basins scales; 2) the Hydrologic Processor, which inputs the forcing ensembles into a suite of hydrologic, hydraulic, and reservoir models; 3) the EnPost, which models the collective hydrologic uncertainty and corrects

for biases in the streamflow ensemble; 4) the EVS, which verifies the forcing and streamflow ensembles to help identify the main sources of skill and bias in the forecasts; and 5) the Graphics Generator, which enables forecasters to derive and visualize products and information from the ensembles. Evaluation of the HEFS through multyear hindcasting and large-

scale verification is currently underway and results obtained so far show positive skill and reduced bias in the short to medium term when compared to climate

ology-based ensembles and single-valued fore-
casts. However, the performance varies significantly with, for example, forecast horizons, basin locations, seasons, and magnitudes, which underlines the need to provide uncertainty-aware forecasts by generating a comprehensive use and needs assessment of the water management community, stressing in particular the need of more detailed information on product skill and uncertainty, guidance for synthesizing the large amount of available hydrologic information to support decision making on probabilistic forecasting principles and risk-based decision making (Raff et al. 2013). Increased collaborations between forecasters, scien-
tists (including social and behavioral scientists), and decision makers should help to link and integrate decision support processes with uncertainty-based forecasts, develop innovative training and education activities to pro-

mote a common understanding, and, ultimately, increase the effective use of probabilistic forecasts for decision making (Ramos et al. 2013; Demerit et al. 2013; Pappenberger et al. 2013). To this end, the NWS Hydrology Program and the RFCs are involved in a number of outreach and training activities, as well as ongoing collabora-
tions with the New York City Department of Environ-
mental Protection. Finally, as CHPS and HEFS are based on the Delft-FEWS platform, complementary visualization techniques and decision support systems are expected to be shared within the Delft-FEWS com-
munity, which will reduce the initial condition uncertainty. Although not implemented in the first version of the HEFS, a number of TA have been developed and are cur-
cently integrated into the AHPS (e.g., maps of the snow accumulation and ablation model, and hydrologic routing models to simulate real-time conditions and adjust model states within the assimilation window (Seo et al. 2003, 2009; Lee et al. 2013). Further activities are needed to improve the data assimilation component, enhancements are also planned to account for the parametric and structural uncertainties in the hydrologic models. As shown by Georgakakos et al. (2004) and Velazquez et al. (2008), some aspects of the current models outperform individual model predictions as long as model-specific biases can be corrected.
Obviously, the different uncertainty modeling approaches available in the HEFS and in other research and operational systems will need to be rigorously compared via ensemble verification to define optimized systems for operational hydrologic ensemble predictions. Close collaborations between scientists, forecasters, and end users from the atmospheric and hydrologic communities, through projects such as the HEPEX, help support such intercomparison, as well as address the following ensemble challenges:

- seamlessly combine probabilistic forecasts from short to long ranges and from multiple models while maintaining consistent spatial and temporal relationships across different scales and variables;
- include forecaster guidance on forcing input forecasts and hydrologic model operations, especially in the short term;
- improve accuracy of both meteorological and hydrologic models and reduce the cone of uncertainty for effective decision support;
- improve the uncertainty modeling of rare events (e.g., record flooding or drought) when availability of analogous historical events is very limited;
- integrate and leverage conditional uncertainty associated with NWP and human adjusted forecasts for research and operational purposes;
- improve computing power, database, and data storage, with forecasts becoming available at higher resolution and from an increasing number of models, to produce long hindcast datasets for all forcing inputs and hydrologic outputs for research and operation purposes;
- improve the understanding of how uncertainty and verification information is interpreted and used in practice by different groups (including forecasters and end users) to provide this information in a form and context that is easily understandable and useful to customers; and
- develop innovative training and education activities for both scientists and practice the ensemble paradigm in hydrologic and water resources services and increase the effectiveness of probabilistic forecasts in risk-based decision making.

ACKNOWLEDGMENTS. This work has been supported by the National Oceanic and Atmospheric Administration (NOAA) through the Advanced Hydrologic Prediction Service (AHPS) Program and the Climate Predictions Program for the Americas (CLIPPER Program Office). The research and development of HEFS over the last decade has involved multiple scientists and forecasters from the Office of Hydrologic Development and the RFCs. The authors would also like to thank Dr. Yuqiong Liu for her valuable contribution and three anonymous reviewers.

REFERENCES


Abstract

NOAA’s National Weather Service (NWS) is implementing a short- to long-range Hydrologic Ensemble Forecast Service (HEFS). The HEFS addresses the need to quantify uncertainty in hydrologic forecasts for flood risk management, water supply management, streamflow regulation, recreation planning, and ecosystem management, among other applications. The HEFS extends the existing hydrologic ensemble services to include short-range forecasts, incorporate additional weather and climate information, and better quantify the major uncertainties in hydrologic forecasting. It provides, at forecast horizons ranging from 6 h to about a year, ensemble forecasts and verification products that can be tailored to users’ needs.

Based on separate modeling of the input and hydrologic uncertainties, the HEFS includes 1) the Meteorological Ensemble Forecast Processor, which ingests weather and climate forecasts from multiple numerical weather prediction models to produce bias-corrected forcing ensembles at the hydrologic basin scales; 2) the Hydrologic Processor, which inputs the forcing ensembles into hydrologic, hydraulic, and reservoir models to generate streamflow ensembles; 3) the hydrologic Ensemble Postprocessor, which aims to account for the total hydrologic uncertainty and correct for systematic biases in streamflow; 4) the Ensemble Verification Service, which verifies the forcing and streamflow ensembles to help identify the main sources of skill and error in the forecasts; and 5) the Graphics Generator, which enables forecasters to create a large array of ensemble and related products. Examples of verification results from multiyear hindcasting illustrate the expected performance and limitations of HEFS. Finally, future scientific and operational challenges to fully embrace and practice the ensemble paradigm in hydrology and water resources services are discussed.