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 greatest benefit.

Development and implementation of hydrologic ensemble prediction systems is still ongoing and hence only limited operational experience exists. A number of case studies using experimental and (pre) operational systems, however, have demonstrated their potential benefits (see, e.g., Cloke and Pappenberger 2009 and Zappa et al. 2010 for references). Recent verification studies of hydrologic ensemble forecasting have been retroactively generated using a fixed forecasting system over long time periods include Barthelements et al. (2009), Jaun and Ahrens (2009), Renner et al. (2009), Hopson and Webster (2010), Demargne et al. (2010), Thiriet et al. (2010), Van den Bergh and Roulun (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. (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 “trained” (van Andel and Alder 2008).

In the National Oceanic and Atmospheric Administration (NOAA)’s National Weather Service (NWS), an end-to-end Hydrologic Ensemble Forecast Service (HEFS) is currently being implemented as part of the Advanced Hydrologic Prediction Service (AHPS; McInery 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 McInery 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 already considered model and initial condition uncertainties) 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 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). The next section presents an overview of the HEFS and its various components. In the subsequent four sections, the individual components are described in more detail and selected illustrative verification results are presented to demonstrate HEFS’s benefits. Finally, future scientific and operational challenges for improving hydrologic ensemble forecasting services are discussed.

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 others. DA techniques are implemented to reduce and model the initial condition uncertainties, such as hydrologic data assimilation (Zappa et al. 2011; Smith et al. 2012; Velázquez et al. 2011), hydraulic models to produce streamflow ensembles. The data assimilation (DA) process currently consists of manual data 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 already considered model and initial condition uncertainties) 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 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). The next section presents an overview of the HEFS and its various components. In the subsequent four sections, the individual components are described in more detail and selected illustrative verification results are presented to demonstrate HEFS’s benefits. Finally, future scientific and operational challenges for improving hydrologic ensemble forecasting services are discussed.

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 others. DA techniques are implemented to reduce and model the initial condition uncertainties, such as hydrologic data assimilation (Zappa et al. 2011; Smith et al. 2012; Velázquez et al. 2011), hydraulic models to produce streamflow ensembles. The data assimilation (DA) process currently consists of manual data 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 already considered model and initial condition uncertainties) 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 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). The next section presents an overview of the HEFS and its various components. In the subsequent four sections, the individual components are described in more detail and selected illustrative verification results are presented to demonstrate HEFS’s benefits. Finally, future scientific and operational challenges for improving hydrologic ensemble forecasting services are discussed.

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 others. DA techniques are implemented to reduce and model the initial condition uncertainties, such as hydrologic data assimilation (Zappa et al. 2011; Smith et al. 2012; Velázquez et al. 2011), hydraulic models to produce streamflow ensembles. The data assimilation (DA) process currently consists of manual data 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 already considered model and initial condition uncertainties) 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 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). The next section presents an overview of the HEFS and its various components. In the subsequent four sections, the individual components are described in more detail and selected illustrative verification results are presented to demonstrate HEFS’s benefits. Finally, future scientific and operational challenges for improving hydrologic ensemble forecasting services are discussed.

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 others. DA techniques are implemented to reduce and model the initial condition uncertainties, such as hydrologic data assimilation (Zappa et al. 2011; Smith et al. 2012; Velázquez et al. 2011), hydraulic models to produce streamflow ensembles. The data assimilation (DA) process currently consists of manual data 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 already considered model and initial condition uncertainties) 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 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). The next section presents an overview of the HEFS and its various components. In the subsequent four sections, the individual components are described in more detail and selected illustrative verification results are presented to demonstrate HEFS’s benefits. Finally, future scientific and operational challenges for improving hydrologic ensemble forecasting services are discussed.
forecasting, as well as hindcasting to provide the large sample of events necessary to verify forecast procedures and develop new event sampling 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-scale 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 river basin, or “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. For example, operational reservoir regulations and diversions is challenging because of the lack of reliable information for the RFC forecasters and changes of reservoir operations to adjust to the current and forecast flow situation. Also, the estimation of historical past reservoir operation and hindcasting may not be consistent with real-time meteorological model inputs, owing to changes in tools (e.g., gauges versus radar for precipitation estimation) and models, as well as changes in inputs. To address these data and model challenges, the RFCs have longstanding practices to apply in a subjective way manual modifications of model states and parameters for single-valued forecasting (see Raff et al. 2013 for details on RFC practices)—modifications that are not currently included in HEFS. The initial HEFS prototype system, referred to as the Experimental Ensemble Forecast System (XEFS; www.nws.noaa.gov/oh/XEFS/), began testing at selecting prototype HEFS The implementation is based on three software development releases to five test RFCs, from spring 2012 to fall 2013. The development phase is targeted to be completed by the end of 2013 with HEFS implementation to all 13 RFCs and associated model systems (e.g., precipitation and model systems). The project has been in agreement with the New York City Department of Environmental Protection, which needs these new probabilistic forecast services to more efficiently and effectively manage the water supply system for New York City (Pyke and Porter 2012). Similarly to NWS operational single-valued hydrologic forecasting, HEFS uses CHPs, an open service-oriented architecture built on the Delft-FEWS framework (Werner et al. 2004). It facilitates incorporation of new models and tools, establishes interoperability with partners, and accelerates research to operations. CHPs is critical in supporting the NOAA Integrated Water Resources Science and Services in partnership with federal agencies (e.g., U.S. Army Corps of Engineers (USACE) and U.S. Geological Survey) that have complementary operational missions in water science, observation, prediction, and management. Also, Delft-FEWS provides an open interface to various data sources and multiple hydrologic and hydraulic forecasting models (see examples of ensemble hydrologic prediction systems based on Delft-FEWS in Werner et al. 2009, Renner et al. 2009, and Schellekens et al. 2011). The system for hydrologic coupling with different hydrologic and hydraulic models, as well as enhancements made within the Delft-FEWS community for hydrologic and water resources forecast systems and services.

**METEOROLOGICAL ENSEMBLE FORECAST PROCESSOR.** Reliable and skillful atmospheric ensemble forecasts are necessary for hydrologic ensemble forecasting. Ensemble forecasts from NWP models can be modified for example an atmospheric prediction centers. However, these models are generally biased in the mean, spread, and higher moments (Buizza et al. 2002), even for lowskill forecasts. The precipitation part of MEFP is intermittent, it depends strongly on the observed record. Therefore, the MEFP uses the single-valued 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 ensemble mean forecasts from multiple NWP models for mid- to long-range generation of ensembles. Both approaches are very attractive computationally, requiring only the conditional Gaussian assumption. Also, additional value to single-valued hydrometeorological variables (e.g., precipitation and temperature). Therefore, the MEFP uses the single-valued forecasts modified by human forecasters for short-range forecast horizon (up to 7 days) and the ensemble mean forecasts from multiple NWP models for mid- to long-range generation of ensembles. The above scheme is based on the same interpolylation procedures used to calculate subdaily historical RFCs in time series and account for the diurnal cycle assumed in the operational calibration process (Anderson 1973).
by the output dimensionality. We therefore suggest examining in the future alternative approaches, which are low dimensional (e.g., analogs) for improved space–time rank structure.

In general, the forecast uncertainty and skill are time–space dependent. Even though the forecast skill at the individual time steps may be limited, especially for long lead times, the skill of forecasts aggregated over multiple time steps is likely to be useful and needs to be exploited for hydrologic and water resources applications. Therefore, the MEFP calibration and ensemble forecasting procedures are also applied to a set of precipitation accumulations and temperature averages defined by the user across different forecast periods from the individual time steps (e.g., n-day events and x-month events up to the maximum available forecast horizon for each forecast source). The final ensemble members at the individual time steps are sequentially produced by the Schaake shuffle for the original and aggregated temporal scales according to increasing forecast skill at the individual scales and for the different forecast sources, with the highest skill having the greatest influence on the final values (see Schaake et al. 2007a for details).

MEFP has been experimentally implemented and evaluated at several RFCs using single-valued forecasts from various sources for a number of different forecast horizons. For the short-range forecast horizon, MEFP uses RFC operational single-valued forecasts as modified by the human forecasters. Depending on forecast locations, these forecasts are available from 1 to 5 forecast lead days for precipitation and up to 7 forecast lead days for temperature. Validation results were reported by 1) Schaake et al. (2007a) for precipitation and temperature for one basin in California, 2) Demargne et al. (2007) for precipitation ensembles (and corresponding streamflow ensembles) for five basins in Missouri and Oklahoma, and 3) Wu et al. (2011) for three basins in Pennsylvania, Arkansas and Missouri, and California. Figure 2 (from Wu et al. 2011) shows the CRPS values of MEFP-generated 6-h precipitation ensembles for the first lead day for the North Fork of the American River basin (NFDC1; 875 km²) near Sacramento, California, using three different methods. Method 1 uses an implicit treatment of precipitation intermittency (Schaake et al. 2007a); methods 2 and 3 model explicitly the precipitation intermittency, with method 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 from various sources for a number of different forecast horizons, with the highest skill having the greatest influence on the final values (see Schaake et al. 2007a for details).

MEFP has been experimentally implemented and evaluated at several RFCs using single-valued forecasts from various sources for a number of different forecast horizons. For the short-range forecast horizon, MEFP uses RFC operational single-valued forecasts as modified by the human forecasters. Depending on forecast locations, these forecasts are available from 1 to 5 forecast lead days for precipitation and up to 7 forecast lead days for temperature. Validation results were reported by 1) Schaake et al. (2007a) for precipitation and temperature for one basin in California, 2) Demargne et al. (2007) for precipitation ensembles (and corresponding streamflow ensembles) for five basins in Missouri and Oklahoma, and 3) Wu et al. (2011) for three basins in Pennsylvania, Arkansas and Missouri, and California. Figure 2 (from Wu et al. 2011) shows the CRPS values of MEFP-generated 6-h precipitation ensembles for the first lead day for the North Fork of the American River basin (NFDC1; 875 km²) near Sacramento, California, using three different methods. Method 1 uses an implicit treatment of precipitation intermittency (Schaake et al. 2007a); methods 2 and 3 model explicitly the precipitation intermittency, with method 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 from various sources for a number of different forecast horizons, with the highest skill having the greatest influence on the final values (see Schaake et al. 2007a for details).

MEFP has been experimentally implemented and evaluated at several RFCs using single-valued forecasts from various sources for a number of different forecast horizons. For the short-range forecast horizon, MEFP uses RFC operational single-valued forecasts as modified by the human forecasters. Depending on forecast locations, these forecasts are available from 1 to 5 forecast lead days for precipitation and up to 7 forecast lead days for temperature. Validation results were reported by 1) Schaake et al. (2007a) for precipitation and temperature for one basin in California, 2) Demargne et al. (2007) for precipitation ensembles (and corresponding streamflow ensembles) for five basins in Missouri and Oklahoma, and 3) Wu et al. (2011) for three basins in Pennsylvania, Arkansas and Missouri, and California. Figure 2 (from Wu et al. 2011) shows the CRPS values of MEFP-generated 6-h precipitation ensembles for the first lead day for the North Fork of the American River basin (NFDC1; 875 km²) near Sacramento, California, using three different methods. Method 1 uses an implicit treatment of precipitation intermittency (Schaake et al. 2007a); methods 2 and 3 model explicitly the precipitation intermittency, with method 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 from various sources for a number of different forecast horizons, with the highest skill having the greatest influence on the final values (see Schaake et al. 2007a for details).

MEFP has been experimentally implemented and evaluated at several RFCs using single-valued forecasts from various sources for a number of different forecast horizons. For the short-range forecast horizon, MEFP uses RFC operational single-valued forecasts as modified by the human forecasters. Depending on forecast locations, these forecasts are available from 1 to 5 forecast lead days for precipitation and up to 7 forecast lead days for temperature. Validation results were reported by 1) Schaake et al. (2007a) for precipitation and temperature for one basin in California, 2) Demargne et al. (2007) for precipitation ensembles (and corresponding streamflow ensembles) for five basins in Missouri and Oklahoma, and 3) Wu et al. (2011) for three basins in Pennsylvania, Arkansas and Missouri, and California. Figure 2 (from Wu et al. 2011) shows the CRPS values of MEFP-generated 6-h precipitation ensembles for the first lead day for the North Fork of the American River basin (NFDC1; 875 km²) near Sacramento, California, using three different methods. Method 1 uses an implicit treatment of precipitation intermittency (Schaake et al. 2007a); methods 2 and 3 model explicitly the precipitation intermittency, with method 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 from various sources for a number of different forecast horizons, with the highest skill having the greatest influence on the final values (see Schaake et al. 2007a for details).
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 HRES ensembles, Browning (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 (GFSF), 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 for various other applications (e.g., hydrologic forecasting, parameterization techniques, for example, to incorporate information from additional and/or alternative postprocessing techniques, for instance, 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 GFS 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 http://erh.noaa.gov/mmefs/). 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 initialization, the model parameterizations, 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; Seo et al. 2006; Coccia and Todini 2011; Brown and Seo 2013; and references therein). They aim to produce reliable (i.e., conditionally unbiased) hydrologic ensemble forecasts from single-valued forecasts or “raw” ensemble forecasts, sometimes with the aid of covariates, accounting only for the hydrologic uncertainty in the forecasts. The regression model is described by a complete density function (e.g., Krzyztofowicz 1999; Seo et al. 2006; Montanari and Grossi 2008; Todini 2008; Bogner and Pappenberger 2011) or several thresholds of the distribution (e.g., Solomatine and Ineson 2009; Brown 2012). Examples of postprocessing techniques for hydrologic ensemble prediction systems include error correction based on the last known forecast error (Velázquez et al. 2009), an autoregressive error correction with the most recent modeled error (Renner et al. 2009; Hopson and Webster 2010; in the latter, postprocessing is also applied to multimodel ensembles), bias correction similar to the MEEP temperature methodology for long-term ESP streamflow ensembles (Wood and Schaake 2008), a Bayesian postprocessor for ensemble streamflow forecasts (Reggiani et al. 2009), error correction for multiple temporal scales based on wavelet transformation (Bogner and Kalas 2008; Bogner and Pappenberger 2011), and a generalized linear regression model using multiple temporal scales (Zhao et al. 2011). To help establish the reliability of different statistical postprocessors and predictors under varied forecasting conditions, the HEPEX project includes an initiative to intercompare postprocessing techniques in order to develop recommendations for their operational use in hydrologic ensemble prediction systems (van Andel et al. 2012).

In the HEFS, the EnsPost (Seo et al. 2010) algorithm for the collection of hydrologic uncertainty is 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 ensembles result 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, reduce the sampling uncertainty of the estimated parameter values. The algorithm adjusts each ensemble trace via recursive linear regression in the normal space (see Seo et al. 2006 for details). The regression is an autoregressive model with an exogenous variable, or ARX(1,1), and uses normal quantile-transformed historical simulation and verifying observation. The regression parameter is optimized for different seasons, and the seasonal factor is 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 currently applied to daily observed and forecast streamflows; after statistical postprocessing, the adjusted ensemble values are disaggregated to subdaily flows. In Seo 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. Seo 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 daily flow ensemble hindcasts were generated for the NFDC1 basin using 38 years of observed–simulated flow records. In Fig. 5, the reliability diagram and the ROC curve relative to a threshold of 95th-percentile flow indicate 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 ensembles 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 were 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; Seo 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 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,
To evaluate for Atlantic RFC (MARFC) (see JANUARY 2014 | ENSEMBLE VERIFICATION. The transitioning of experimental automated DA capabilities into operational ensemble forecasting offers objective guidance on best operational practices for applying manual modifications and cost-effective algorithms on the performance of HEFS flow ensembles. Such comprehensive evaluation could 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 are 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 objective 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 (Jolliffe 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.rws.noaa.gov/oh/evs.html), was designed to support conditional verification of forcing and hydrologic ensembles, generated by HEFS, as well as hydrologic ensemble forecasting systems. 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
skill, at varying levels of detail, ranging from highly summarized (e.g., skill scores such as CPEB for a boxplot or detailed (e.g., box plots of conditional errors); the ability to aggregate the forecasts in time (e.g., hourly to daily) and to evaluate aggregate performance 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 (e.g., PDF, HTML, or XML); 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 command line, using a project file in an XML format; the ability to estimate the sampling uncertainty in the verification metrics using the stationary 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 verification 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 implemented in other areas and associated with other Verification Systems such as HEEPEX.

**GRAPHICS GENERATOR.** Communicating uncertainty information to a wide range of end users represents a challenge. As hydrologic ensemble forecasting is relatively new, much research is needed to define the most effective methods of presenting such information. Verification support systems need to maximize 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 decision support forecasts), and how to communicate it (e.g., spaghetti plots versus plume charts), even to nonexperts (Demeritt et al. 2010). A variety of practices have been presented by Brun et al. (2010) for seven European ensemble forecasting platforms and by Ramos et al. (2007) and Demeritt et al. (2013) for the European Flood Alert System. Pappenberger et al. (2013) formulated recommendations for effective visualization of both ensemble and deterministic forecasts in the context of uncertainty in both upstream (e.g., spaghetti plots versus plume charts), even to (deterministic forecasts), and how to communicate the resulting information to end users in a clear and effective manner. One such approach is the Web interface for the CBRFC water supply forecasts available from www.cbrfc.noaa.gov. It enables users to generate a wide range of long-term, uncertainty-quantified probabilistic forecasts and verification statistics. Such visualization tools provide insights into the strengths and weaknesses of the forecasts and can help users assess potential forecast errors in the real-time forecasts. Along with the probabilistic forecast information, the visualized CBRFC products will include context information (e.g., historical lowest and highest, specific years of interest—El Niño, La Niña, or neutral), recent forecasts and corresponding observations, forecasts from alternative scenarios, as well as spatial spreads for real-time forecasts. The selection of analogs (i.e., past forecasts that are analogous to the current forecast) and the display of diagnostic verification statistics from similar conditions provide important contextual information tailored to the specific real-time forecasting situation.

Through customer evaluations of the AHPS website and the NWS Hydrology Program, the NWS has recognized the need to better communicate hydrologic forecast uncertainty information for the end users to understand better and use such information more effectively in their decision making. The NWS, USACE, and Bureau of Reclamation conducted a comprehensive use and needs assessment of the water management community, stressing in particular the need for more detailed information on product skill and uncertainty, guidance for synthesizing the large amount of hydrometeorological information, and training on probabilistic forecasting principles and risk-based decision making (Raff et al. 2013). Increased collaborations between forecasters, scientists (including social and behavioral scientists), and decision makers should help to develop better understanding of the forecast processes with uncertainty-based forecasts, develop innovative training and education activities to promote a common understanding, and, ultimately, improve the effectiveness of probabilistic forecasts (e.g., for the Sacramento River Basin; Ramos et al. 2013; Demeritt 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 collaboration with the New York City Department of Environmental Protection. Finally, as CHIPS and HEFS are based on the Delft-FEWS platform, comprehensive visualization techniques and decision support systems are expected to be shared within the Delft-FEWS community. The verification results to maximize the utility of hydrologic and water resources forecast products and services.

**CONCLUSIONS AND FUTURE CHALLENGES.** The end-to-end HEFS provides, on the one hand, long range uncertainty-quantified forecast and verification products that are generated by 1) the MEEF, which ingest weather and climate forecasts from multiple Numerical Weather Prediction models to produce seasonal and longer range precipitation and temperature ensembles at the hydrologic basinscale; 2) the Hydrologic Processor, which inputs the forcing ensembles into a suite of hydrologic, hydraulic, and reservoir models; and 3) the EmPost, 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 sample verification is currently underway and results obtained so far show positive skill and reduced bias in the short to medium term when compared to climatology-based ensembles and single-valued forecasts. However, the performance varies significantly with, for example, forecast horizons, basin locations, seasons, and magnitudes, which underlines the need to provide a more comprehensive verification 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 NWSRFC 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 compared to historical simulations (see examples in McEnery et al. 2005). HEFS also needs to be adapted to the hydrologic forecast and verification metrics from several locations (e.g., 90% chance to exceed flood impact thresholds) as well as individual locations (e.g., expected value charts for all forecast lead times). New products to visualize the forcing and streamflow ensembles, including the verification metrics, 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 about how forecasters should help to understand uncertainty in future forecasts in the context of the estimated uncertainty.

Several RFMCs make prototype ensemble products and information available to their customers (see www.water.weather.gov/evs for the Delft-FEWS-based probabilistic forecast system). The Hydrologic Processor, which inputs the forcing and streamflow ensembles that are aggregated into the operational national web interface for AHPS (http://water.weather.gov/ahps/). Furthermore, verification information needs to be provided along with forecast information to support decision making (Demargne et al. 2010). Similar approaches have been reported by Bartholmes et al. (2007), Renner et al. (2009), and van Andel et al. (2010). GraphGen has been presented by Brun et al. (2010) for seven European ensemble forecasting platforms and by Ramos et al. (2007) and Demeritt et al. (2013) for the European Flood Alert System. Pappenberger et al. (2013) formulated recommendations for effective visualization of both ensemble and deterministic forecasts in the context of uncertainty in both upstream (e.g., spaghetti plots versus plume charts), even to (deterministic forecasts), and how to communicate the resulting information to end users in a clear and effective manner. One such approach is the Web interface for
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 optimal models 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 interchange, 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 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 useful and clear information in a form and context that is easily understandable and useful to customers; and
- develop innovative training and education activities that will familiarize and practice the ensemble paradigm in hydrology 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 under the Climate Program Office. The major milestone 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


JANUARY 2014 | AMS | 14

JANUARY 2014 | AMS | 17

AMERICAN METEOROLOGICAL SOCIETY

14

17

BYF

BYF
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.