Verification of Precipitation Forecasts from NCEP’s Short-Range Ensemble Forecast (SREF) System with Reference to Ensemble Streamflow Prediction Using Lumped Hydrologic Models

JAMES D. BROWN
NOAA/National Weather Service, Office of Hydrologic Development, Silver Spring, Maryland, and University Corporation for Atmospheric Research, Boulder, Colorado

DONG-JUN SEO
Department of Civil Engineering, The University of Texas at Arlington, Arlington, Texas

JUN DU
NOAA/NWS/NCEP/Environmental Modeling Center, Camp Springs, Maryland

(Manuscript received 7 April 2011, in final form 13 December 2011)

ABSTRACT
Precipitation forecasts from the Short-Range Ensemble Forecast (SREF) system of the National Centers for Environmental Prediction (NCEP) are verified for the period April 2006–August 2010. Verification is conducted for 10–20 hydrologic basins in each of the following: the middle Atlantic, the southern plains, the windward slopes of the Sierra Nevada, and the foothills of the Cascade Range in the Pacific Northwest. Mean areal precipitation is verified conditionally upon forecast lead time, amount of precipitation, season, forecast valid time, and accumulation period. The stationary block bootstrap is used to quantify the sampling uncertainties of the verification metrics. In general, the forecasts are more skillful for moderate precipitation amounts than either light or heavy precipitation. This originates from a threshold-dependent conditional bias in the ensemble mean forecast. Specifically, the forecasts overestimate low observed precipitation and underestimate high precipitation (a type-II conditional bias). Also, the forecast probabilities are generally overconfident (a type-I conditional bias), except for basins in the southern plains, where forecasts of moderate to high precipitation are reliable. Depending on location, different types of bias correction may be needed. Overall, the northwest basins show the greatest potential for statistical postprocessing, particularly during the cool season, when the type-I conditional bias and correlations are both high. The basins of the middle Atlantic and southern plains show less potential for statistical postprocessing, as the type-II conditional bias is larger and the correlations are weaker. In the Sierra Nevada, the greatest benefits of statistical postprocessing should be expected for light precipitation, specifically during the warm season, when the type-I conditional bias is large and the correlations are strong.

1. Introduction
Reliable forecasts of precipitation and temperature are essential for operational streamflow forecasting. They are needed at space–time scales ranging from minutes and kilometers (e.g., for flash flood guidance) to multiple years and entire regions (e.g., for water supply outlooks). To produce reliable streamflow forecasts at multiple space–time scales, the River Forecast Centers (RFCs) of the U.S. National Weather Service (NWS) are evaluating temperature and precipitation forecasts from a range of Numerical Weather Prediction (NWP) models. These include the Short-Range Ensemble Forecast system (SREF; Du et al. 2009), the Global Ensemble Forecast System (GEFS; Toth et al. 1997), and the Climate Forecast System (CFS; Saha et al. 2006) of the National Centers for Environmental Prediction (NCEP). Collectively, the SREF, GEFS, and CFS provide atmospheric forecasts from a few hours to several months into the future and cover hydrologic basins of a few hundred square kilometers to several million square kilometers.

Corresponding author address: James D. Brown, NOAA/National Weather Service, Office of Hydrologic Development, 1325 East–West Highway, Silver Spring, MD 20910.
E-mail: james.d.brown@noaa.gov

DOI: 10.1175/JHM-D-11-036.1

© 2012 American Meteorological Society
However, all of these models are subject to error and uncertainty, including systematic modeling errors that are conditional upon complex atmospheric states and forcing mechanisms (Stensrud et al. 2000; de Elia and Laprise 2003; Eckel and Mass 2005; Jones et al. 2007; Clark et al. 2009; Schwartz et al. 2010; Schumacher and Davis 2010).

Uncertainties in atmospheric forecasts combine with uncertainties in hydrologic modeling and lead to uncertain streamflow predictions (Brown and Heuvelink 2005). Depending upon varied hydrologic states and basin characteristics, the atmospheric uncertainties may contribute significantly to the overall uncertainties in hydrologic forecasting (Kobold and Suselj 2005; Buurger et al. 2009; Pappenberger and Buizza 2009; Zappa et al. 2010; Mascaloro et al. 2010). Thus, understanding the atmospheric uncertainties is essential for generating reliable and skillful hydrologic forecasts. In operational hydrologic forecasting, ensemble techniques are increasingly used to quantify and propagate uncertainty (Epstein 1969; Brown and Heuvelink 2005). For example, the NWS RFCs produce ensemble forecasts of streamflow at a variety of lead times (Seo et al. 2006; Schaake et al. 2007). In one experimental operation, ensemble traces of precipitation and temperature are generated with an ensemble preprocessor (EPP; Schaake et al. 2007; Wu et al. 2011). The EPP estimates the conditional probability distribution of the future observation given a single- valued forecast, of which several forecasts are currently used in EPP (Wu et al. 2011). Elsewhere, the mid-Atlantic (MA), Ohio (OH), and Northeast (NE) RFCs are developing a meteorological model-based ensemble forecast system (MMEFS) that uses “raw” ensemble forecasts from the operational GEMFs and SREF to produce experimental hydrologic forecasts for the short to medium range. The forcing inputs, whether from EPP or MMEFS, are input into the Ensemble Streamflow Prediction (ESP) subsystem of the Hydrologic Ensemble Forecasting Service (HEFS), from which ensemble traces of streamflow are output. Verification of the forcing and flow ensembles is necessary to establish the key sources of error and uncertainty in the HEFS (Demargne et al. 2010) and to identify the benefits of improved NWP and hydrologic modeling.

Forecasts from operational NWP models have improved significantly in recent years. Deterministic models have benefited from more resolved spatial grids and time steps together with new physical parameterizations and data assimilation schemes (Warner 2011). Alongside these improvements, ensemble prediction systems have benefited from developments in stochastic modeling. These include new algorithms for initializing ensemble forecasts (Stensrud et al. 2000; Wei et al. 2008; Schwartz et al. 2010) and for assimilating uncertain weather observations with ensemble and variational techniques (Whitaker et al. 2008; Houtekamer et al. 2009), as well as incorporating additional sources of uncertainty, such as model structural uncertainties (Rabier 2006; Yuan et al. 2009; Raynaud et al. 2012). Indeed, when deterministic forecasts are unskillful, probabilistic forecasts may nevertheless contain skill, particularly for large events and forecasts of “noisy” variables, such as precipitation and temperature (de Elia and Laprise 2003). Alongside developments in NWP models, statistical postprocessors can further improve the reliability and resolution of forcing ensembles, providing they are adequately calibrated and their statistical assumptions are met (Wilks 2006; Yuan et al. 2007; Hamill et al. 2008; Unger et al. 2009).

Nevertheless, questions remain about the relative benefits of new deterministic versus stochastic modeling techniques (Eckel and Mass 2005; Clark et al. 2009; Weigel et al. 2008, 2009), particularly for quantitative precipitation forecasts (McCollor and Stull 2008), and how they might contribute to hydrologic forecasting. For example, errors in the parameterization of finescale precipitation mechanisms, such as orographic forcing and thermal convection, combine with uncertainties in the initial conditions and model structure, and quickly saturate (Yuan et al. 2005, 2009). To account for these errors, regional-scale NWP models increasingly employ convection-allowing resolutions (CAR) of ~4 km or better (Clark et al. 2009, 2010, 2011; Schwartz et al. 2010). Elsewhere, multimodel and multiphysics schemes (Du et al. 2009) and stochastic physics (Buizza et al. 1999) have lengthened the “effective lead times” of precipitation forecasts (e.g., Ruiz et al. 2012). Yet any improvements in operational models must be weighed against the computing resources required to implement them and, more importantly, to evaluate them with appropriately large verification samples. In practice, upgrades to operational models are rarely accompanied by long-term hindcasting and verification experiments (Hamill et al. 2006). This can impede the calibration of statistical postprocessors for use in operational ESP and obscure the hydrologic benefits of improved atmospheric modeling.

The SREF is a multicore, multiphysics, ensemble prediction system that provides a compromise between “reasonable resolution” of the atmospheric models and consideration of multiple sources of uncertainty, although the nature of this compromise is, clearly, application dependent. The SREF was implemented at the NCEP in May 2001 and initially comprised 10 ensemble members from theEta and Regional Spectral Model (RSM) models (Du and Tracton 2001). Subsequent
updates have increased the physics diversity by adding members from the Weather Research and Forecasting (WRF) models [Nonhydrostatic Mesoscale Model (NMM) and Advanced Research WRF (ARW) cores], which has increased membership from 10 to 21 ensemble members. The SREF has been verified for specific variables, years, model cores, and seasons by Hamill and Colucci (1998), Yuan et al. (2005, 2007), Du et al. (2009), and Charles and Colle (2009), among others. For example, Yuan et al. (2005) verify precipitation forecasts from the RSM model for the 2002/03 cool season in the Southwest United States. The RSM forecasts were most skillful along the California coastline and on the windward slopes of the Sierra Nevada, with significantly worse skill in the Great Basin and the Colorado basin (except over mountain peaks). However, Yuan et al. (2005) also note that uncertainties in the quantitative precipitation estimates (QPEs) used to verify the RSM ensembles were a significant factor controlling the apparent lack of skill in the RSM for the Great Basin and other areas with limited QPE coverage. Charles and Colle (2009) verified the SREF ensembles from the 2004–07 cool seasons, with an emphasis on the strengths and positions of extratropical cyclones. In general, they found a lack of spread across the central and eastern United States, particularly at longer lead times, and too much spread in the eastern Pacific, eastern Canada, and western Atlantic.

Grid-based verification of the SREF (e.g., Yuan et al. 2005, 2007; Charles and Colle 2009) confers the advantage of being spatially exhaustive, particularly in areas of complex terrain where forecast skill can vary over short distances (Yuan et al. 2005). However, for hydrologic applications, there is a need to verify the SREF at different spatial scales, over multiple years, in different climate zones, for both warm and cool seasons, and for a range of accumulation periods. For example, lumped hydrologic modeling is an integral part of the NWS HEFS. This requires mean areal precipitation (MAP) over hydrologic basins. In general, atmospheric forecasts are less meaningful at their nominal grid resolution than over aggregated areas, particularly for near-surface temperatures and precipitation, which comprise finescale orographic and surface flux components (Harris et al. 2001; Brussolo et al. 2008). Similarly, while hydrologic forecasts are usually based on 6-hourly MAPs, daily accumulations are, in general, more meaningful for operational hydrologic forecasting (i.e., for the range of basin scales used in operational forecasting; see Pappenberger and Buizza 2009 also). Finally, hydrologic models are sensitive to biases in atmospheric forcing, and streamflow postprocessors often assume unbiased forcing (e.g., Zhao et al. 2011). Thus, appropriate bias corrections must be developed for hydrologic applications. This requires verification over longer periods of time (and, ideally, large areas) in order to establish the performance of statistical postprocessors that are calibrated on pooled data and applied to operational models that undergo frequent changes.

This paper evaluates the quality of the SREF precipitation forecasts from April 2006 to August 2010 with an emphasis on their use in ESP. The forecasts are verified for selected basins in four climate regions. Verification is performed conditionally upon the amount of precipitation, forecast lead time, season, and time of day, among other things. However, the analysis focuses on moderate to heavy precipitation amounts, as these are critical for the RFCs at short forecast lead times (e.g., for flood prediction). The paper is organized in three parts: 1) a description of the data and verification methodology; 2) the verification results and analysis, which are ordered by conditioning variable (amount of precipitation, season, etc.); and 3) the conclusions and opportunities for future work.

2. Materials and methods

a. Study area

Verification was performed for 10–20 contiguous basins in each of four RFCs (Fig. 1). Contiguous basins were chosen to increase the verification sample size via spatial pooling. In particular, this increased the sample of moderate and heavy precipitation amounts, as storms in the ~100–200-km mesoscale range, which can be detected by SREF, can miss individual basins. Each study basin comprises one or more NWS operational forecast points for which the HEFS is calibrated. Thus, in follow-up work, the SREF will be input to the HEFS and tested for its ability to generate reliable and skillful streamflow ensembles at the hydrologic basin scale. While spatial pooling was limited to 10–20 basins in each RFC (Fig. 1), the basin groups were selected to represent multiple climate regions. These include the middle Atlantic (MARFC), the southern plains [Arkansas–Red Basin (AB) RFC], the windward slopes of the Sierra Nevada [California Nevada (CN) RFC], and the coastal mountains of the Pacific Northwest [Northwest (NW) RFC]. Subsequently, each basin group is referred to by its identifying RFC (Fig. 1). However, the operating areas of the RFCs are much larger than the basin groups considered in this study (Fig. 1).

b. Datasets

Operational forecasts of total precipitation were provided by NCEP for a ~4.5-yr period from April 2006 to August 2010 (contact the first author for availability).
The forecasts comprise 4-times-daily model runs at 0300, 0900, 1500, and 2100 UTC from November 2007 to August 2010 and 2-times-daily model runs at 0900 and 2100 UTC from April 2006 to November 2007 (the 0300 and 1500 UTC runs were not archived for the earlier period). Each forecast cycle comprises 3-hourly forecasts for lead times of 3–87 h, with accumulations valid for the preceding 3 h. The current operational SREF employs perturbations of the initial conditions, which are based on a mixture of the Global Forecast System (GFS) ensemble transform perturbations (Wei et al. 2008) and regional-bred vectors as well as perturbations of the lateral boundary conditions from the GEFS. Major upgrades to the SREF system occurred in December 2005.

FIG. 1. Study basins with mean elevations in meters above mean sea level (MSL). The MARFC basins comprise RTDP1, PORP1, MPLP1, HUNP1, WIBP1, SLYP1, SPKP1, LWSP1, NPTP1, and SAXP1. The ABRFC basins comprise WSCO2, QUAO2, KNSO2, WTTO2, ELMA4, BSGM7, SVYA4, LSGM7, TALO2, SLSA4, INCM7, TIFM7, ELD02, SPAO2, JOPM7, and PEN02. The CNRFC basins comprise FMDC1, RRGC1, UNVC1, HLLC1, EDOC1, MRYC1, CFWC1, CBAC1, MFAC1, NFDC1, FOLC1, and HLEC1. The NWRFC basins comprise GARW1, LNDW1, ISSW1, RNTW1, MORW1, SQUW1, TANW1, CRNW1, AUBW1, HHDW1, and MMRW1.
and November 2009 (Du et al. 2009). The last upgrade comprised a change in member composition, with the substitution of fourEta members for WRF members, and an increase in horizontal resolution from ~40 to ~32 km, among other things (Du et al. 2009). While these changes are expected to improve forecast skill and conditional bias (e.g., Du et al. 2009), they were not seen to substantially change the behavior of the SREF ensembles. Consequently, verification results are shown for the extended period of April 2006–August 2010.

Two sources of observed precipitation were available for this study—namely 1) a gridded Climatology-Calibrated Precipitation Analysis (CCPA) from NCEP (D. Hou 2011, personal communication) and 2) basin-averaged observed precipitation from the NWS RFCs (RFC QPE). The CCPA QPE data was available for the period January 2002–August 2010 on the ~4-km grid of the Hydrologic Rainfall Analysis Project (HRAP). The precipitation estimates are based on the RFC stage-IV multisensor precipitation analysis (Seo 1998). However, unlike the raw stage IV, which is quality controlled by each RFC, the CCPA QPE is further bias corrected in a consistent way. Specifically, the stage-IV QPEs are climatologically adjusted using the NOAA Climate Prediction Center’s (CPC) Unified Global Daily Gauge Analysis. Calibration is performed at the temporal and spatial resolution of the CPC analysis (daily, $1/8^\circ$) by aggregating the stage-IV QPEs and estimating the parameters of a linear regression between the aggregated QPEs and the CPC analysis. The adjusted QPEs are then partitioned into six-hour accumulations on the HRAP grid using the fractional precipitation amounts from the corresponding times and grid cells in the raw stage-IV QPEs. MAP was obtained for each study basin by averaging over the $g$ HRAP grid cells that fell within the basin:

$$\text{MAP} = \frac{1}{g} \sum_{i=1}^{g} f_i p_i, \quad f_i \in [0, 1]; \quad p_i \geq 0,$$

where $f_i$ is the fraction of the $i$th grid cell that falls within the basin boundary and $p_i$ is the amount of precipitation in the $i$th grid cell. Figure 2 shows the precipitation climatologies of each basin group for accumulation periods of 6, 12, and 24 h. The climatologies were derived by pooling the CCPA QPEs across the 10–20 study basins in each RFC.

While verification was performed with the CCPA QPEs, the RFC QPEs were obtained for comparison. The RFC QPEs are less uniform than the CCPA QPEs, comprising a range of dates, accumulation periods, basin averaging techniques, and data sources, including gauge-based, radar-based, and gauge-adjusted radar estimates. The quality control also varies between RFCs, with some RFCs (notably NWRFC) employing custom station weights when deriving the MAPs for hydrologic simulations (see below). Nevertheless, the RFC and CCPA observations are highly correlated, with correlation coefficients of 0.9, 0.91, 0.9, and 0.88 for MA-, AB-, CN-, and NWRFCs, respectively. Cross correlations with the ensemble mean forecast are also high, both conditionally and conditionally upon precipitation threshold (not shown). Also, as indicated in Fig. 3, the unconditional quantiles of the CCPA and RFC QPEs are reasonably consistent for MA-, AB-, and CNRFCs. However, while NWRFC exhibits good correlations for the 24-h period, there are substantial differences in the unconditional quantiles (Fig. 3), with ~35% more precipitation in the CCPA estimates at all precipitation thresholds (but varying between basins). Correspondence with NWRFC suggests that manual calibration of the RFC QPEs may be responsible. In particular, the station weights used to derive the MAPs are manually adjusted to reduce bias in the flow simulations (B. Gillies, NWRFC, 2011, personal communication), and this can transfer biases from the hydrologic model to the precipitation estimates.

Verification pairs were derived from the SREF ensemble forecasts and corresponding QPEs for accumulation periods of 6, 12, and 24 h. The forecasts were initialized (and hence valid) at odd hours of 0300, 0900, 1500, and 2100 UTC. In contrast, the observations were valid at even hours of 0000, 0600, 1200, and 1800 UTC. The first forecast was, therefore, dropped from each cycle, and 6-h accumulations were derived from the remaining forecasts. When accumulating by forecast lead time, the resulting 6-h accumulations are $[3–9, \ldots, 81–87]$, the 12-h accumulations are $[3–15, \ldots, 75–87]$, and the 24-h accumulations are $[3–27, \ldots, 63–87]$.

**c. Verification strategy**

The aim of the verification is to 1) examine the conditional skill and biases in the SREF precipitation forecasts for selected hydrologic basins and, thereby, 2) guide the development of bias-correction techniques for hydrologic applications. For each basin group, verification results were computed conditionally upon forecast lead time, amount of precipitation, season, forecast valid time, and accumulation period. Limited combinations of these attributes were also considered (e.g., season and amount of precipitation), but were often constrained by the sampling uncertainties of the verification metrics. Indeed, in verifying atmospheric models, there is always a trade-off between the sampling uncertainty and stationarity (or applicability) of the verification results, whether pooling samples in space, time, or across varied observed or forecast amounts. There are few general guidelines for pooling samples, and verification results should always
be viewed as limited and contingent. For example, aggregation across multiple years and seasons, varied terrain, and different atmospheric states and storm types (among other variables) is unavoidable, and can lead to a false impression of conditional bias and skill (e.g., Hamill and Juras 2006). To avoid excessive aggregation, long-term hindcasts from a frozen NWP model are preferred over a limited sample of operational forecasts, but were not available in this study (or, in general, for major upgrades of operational models).

Verification was performed with the NWS Ensemble Verification System (EVS; Brown et al. 2010) and the sampling uncertainties were quantified with the stationary block bootstrap (Politis and Romano 1994). Here, blocks of adjacent pairs are sampled randomly, with replacement, from the \( n \) available pairs for each basin and forecast lead time. The central index of each block has a discrete uniform distribution on \( \{1, \ldots, n\} \) and its length, \( b \), has a geometric distribution with probability of success, \( p = 1/b \), in order to avoid nonstationarity in the bootstrap sample (see Politis and Romano 1994 and Lahiri 2003 for details). Experimental correlograms were computed from the sample data for a range of precipitation thresholds. A block length of \( b = 7 \) days was found to capture most of the temporal dependence in precipitation across all RFCs and thresholds. The resampling was repeated 10 000 times, and the verification metrics computed for each sample. Confidence

![Graphs showing climatological probability distributions for observed accumulations](image-url)

**Fig. 2.** (a)–(d) Climatological probability distributions for observed accumulations. The thick lines denote the climatological distributions pooled across all basins. The shaded areas correspond to the minimum and maximum values from the individual basins. The dashed lines mark the boundaries of the shaded areas. The dashed lines parallel to the axes highlight a climatological exceedance probability of 0.01.
intervals were derived from the bootstrap sample with a nominal coverage probability of 0.9—that is, [0.05, 0.95]. When pooling pairs from multiple basins, perfect spatial dependence was assumed (i.e., each randomly sampled block was used for all basins). In general, bootstrap confidence intervals do not provide unbiased estimates of coverage probabilities (see Lahiri 2003), and should only be regarded as indicative for large events. Also, observational uncertainties were not considered.

Key attributes of forecast quality are obtained by examining the joint probability distribution of the observed variable, \( Y \), and the forecast variable, \( X \), \( f_{XY}(x, y) \). The joint distribution can be factored into \( f_{Y|X}(y|x) \times f_X(x) \), which is known as the “calibration-refinement” (CR) factorization and \( f_{X|Y}(x|y) \times f_Y(y) \), which is known as the “likelihood–base rate” (LBR) factorization (Murphy and Winkler 1987). Differences between \( f_X(x) \) and \( f_Y(y) \) describe the unconditional biases in the forecast probabilities. The conditional probability distribution function (pdf), \( f_{Y|X}(y|x) \), describes the type-I conditional bias or “reliability” of the forecast probabilities when compared to \( f_X(x) \) and resolution when only its sensitivity to \( X \) is considered. In operational forecasting, the reliability of a forecast may be improved through statistical postprocessing, which aims to estimate \( f_{Y|X}(y|x) \) given a raw ensemble forecast. For a given level of reliability, forecasts with smaller spread (i.e., sharp forecasts) are sometimes preferred over more diffuse ones, as they contribute less uncertainty to decision making (Gneiting et al. 2007). The conditional pdf, \( f_{X|Y}(x|y) \), describes...
the type-II conditional bias of the forecasts when compared to $f_Y(y)$ and discrimination when only its sensitivity to $Y$ is considered. For any given attribute of forecast quality, there are several possible metrics or measures of quality. Some of these measures, such as the Brier score (BS; Brier 1950), can be decomposed algebraically into more detailed measures on the CR and LBR distributions (Hersbach 2000; Bradley et al. 2004). The appendix summarizes the key metrics and measures used in this paper.

When verifying forecasts of continuous random variables, such as precipitation and streamflow, verification is often performed for discrete events (Jolliffe and Stephenson 2012; Wilks 2006). To compare the verification results between basins and seasons, for different forecast lead times and valid times, and for different accumulation periods, common events were identified for each of MA-, AB-, CN-, and NWRFCs. Specifically, for each RFC and accumulation period, $a$, the CCPA QPEs were pooled across all study basins and used to compute an empirical, climatological, distribution function, $\hat{F}_{n,a}^{-1}(x)$. Real-valued thresholds were then determined for $k = 50$ climatological exceedance probabilities, $c_p$, $\hat{F}_{n,a}^{-1}(c_p), c_p \in [0, 1]; p = 1, \ldots, k$, including the probability of precipitation (PoP) threshold, which was identified separately for each basin and accumulation period. Measures that depend continuously on the data, such as the mean error, were derived from the conditional sample in which the observed value exceeded the threshold. Measures defined for discrete events, such as the BS, were computed from the observed and forecast probabilities of exceeding the threshold.

3. Results and analysis

a. Forecast lead time

Figure 4 shows the correlation of the observed variable and the ensemble mean forecast for the 6-h accumulations at lead times of 9–87 h. The correlations are shown for all data and for precipitation amounts exceeding the real-valued thresholds with climatological exceedance probabilities of $\{0.1, 0.01, 0.001\}$. At light to moderate precipitation thresholds, $\{all\ data, 0.1\}$, the correlation declines systematically with increasing forecast lead time in all RFCs. At high precipitation thresholds, $\{0.01, 0.001\}$, the correlations are more sensitive to model initialization time than forecast lead time, particularly in the mountainous terrain of NWRFC, although the sampling uncertainties are large. This is evidenced by the 6-hourly cycle in correlation with increasing forecast lead time. Before 2008, when the 0300 and 1500 UTC forecasts are missing (see above), forecasts initialized at 0300 and 1500 UTC with valid times of 0000 and 1200 UTC do not contribute to the verification results at lead times $\{9, 21, 33, 45, 57, 69, 81\}$. Similarly, forecasts initialized at 0300 and 1500 UTC with valid times of 0600 and 1800 UTC do not contribute to the verification results at lead times $\{15, 27, 39, 51, 63, 75\}$. Thus, slight differences in forecast quality between the 0300/1500 and 0900/2100 UTC cycles are reflected in the forecast lead times. Figure 5 shows the continuous rank probability skill score (CRPSS; Hersbach 2000), relative to sample climatology, for lead times of 9–87 h. Again, the forecast skill declines smoothly with increasing forecast lead time and shows some sensitivity to model initialization time, particularly in NWRFC. However, unlike the correlation coefficient, which measures the quality of the ensemble mean forecast, the CRPSS is also distinguishable for high precipitation amounts. Overall, CNRFC shows the best forecast skill, with an equivalent or better CRPSS at 2 days ahead than the other RFCs show at 9 h ahead (see Yuan et al. 2005 also). While the CRPSS declines consistently across all RFCs with increasing forecast lead time, the dependence on precipitation amount varies between RFCs.

b. Precipitation amount

Figures 6 and 7 show the Brier skill score (BSS) for 24-h precipitation totals at three forecast lead times—namely 4–27, 28–51, and 52–75 h. Figure 6 shows the CR factorization of the BSS into relative reliability (or type-I conditional bias) and relative resolution (see appendix). Figure 7 shows the LBR factorization into relative type-II conditional bias, discrimination, and sharpness. The scores are plotted against climatological exceedance probability, $c_p$ (note: the thresholds are spaced on a logit scale—i.e., $\log_{10}[c_p/(1 - c_p)]$—but labeled with actual $c_p$). Figure 8 shows the correlation between the observed variable and the forecast ensemble mean together with the relative mean error (RME) of the ensemble mean and the CRPSS. The RME comprises the average error as a fraction of the average observed value. Again, the scores are plotted by increasing precipitation threshold for each 24-h accumulation period. Figure 9 shows the reliability, or type-I conditional bias, of the forecast probabilities for selected precipitation thresholds in each RFC (Hsu and Murphy 1986).

The skill with which the forecasts predict the exceedance of a fixed precipitation threshold depends strongly on the threshold value, with substantially better BSS for light to moderate precipitation than either PoP or heavy precipitation (Fig. 6). The reduced performance for PoP and heavy precipitation reflects the conditional biases in the forecasts. In particular, the forecasts overestimate...
PoP and light precipitation when no precipitation is observed and underestimate heavy precipitation when heavy precipitation occurs, both in terms of forecast probabilities and amounts. This is apparent in the type-II conditional bias of the BSS (Fig. 7), the RME of the ensemble mean forecast (Fig. 8), and in the quantiles of the raw ensemble members (Fig. 10). Figure 10 compares the unconditional quantiles of the observed and forecast distributions for each ensemble member as well as the forecast ensemble mean. The forecast quantiles comprise 24-h precipitation totals with lead times of 27–51 h. As the SREF membership changed in late 2009 (Du et al. 2009), results are shown for the period April 2006–November 2009. In contrast to the quantiles of the individual members in Fig. 10, the ensemble mean preserves the time indexing across the individual members. Thus, in quantile space, the relative trajectories of the ensemble mean and the ensemble members will depend upon the time-dependent conditional biases in the ensemble mean. In MA- and ABRFCs, these trajectories are substantially different for moderate to heavy precipitation amounts, with much larger negative biases in the ensemble mean forecast than the bulk of the individual members (see Fig. 8 also).

In general, the type-I conditional biases are larger for PoP and heavy precipitation than moderate precipitation, as indicated in Fig. 6. However, unlike the type-II conditional biases, the reliability diagrams (Fig. 9) suggest

FIG. 4. (a)–(d) Correlation of the 6-h observed and ensemble mean accumulations by lead time. Results are shown for all data and for conditional subsets in which the observed value is greater than a given climatological exceedance probability. The shaded areas represent the 5th–95th confidence intervals for the score value and the dashed lines represent the boundaries of these shaded areas (used in subsequent plots also).
In other words, the forecast probabilities are overconfident conditionally upon the forecast amount/probability, but underestimate heavy observed precipitation conditionally upon the observed amount. While ABRFC shows the largest (relative) type-II conditional biases of any RFC (Fig. 7), it also shows the most reliable forecasts of heavy precipitation (Fig. 6 and Fig. 9).

Figure 11 shows the relative operating characteristic (ROC) for each RFC. The ROC shows the ability of the forecasts to discriminate between selected events and nonevents across a range of “decision thresholds”; that is, forecast probabilities at which a decision is taken (Green and Swets 1966). The ROC curves were fitted under the assumption of bivariate normality between the probability of detection (POD) and the probability of false detection (POFD) (see appendix) and are shown together with the empirical pairs of POD and POFD for three event thresholds—namely \( c_p = \text{PoP}, c_p = 0.05, c_p = 0.01 \).

The most skillful forecasts of heavy precipitation are for the NWRFC area (see Yuan et al. 2005 also), both in terms of the BSS (Fig. 7) and CRPSS (e.g., Figure 8). This is explained by the relatively high correlations (Fig. 8) and relatively low unconditional and conditional biases in the ensemble mean forecast (Figs. 8 and 10). In addition, the NWRFC forecasts are highly resolved for heavy precipitation events (Fig. 6) and also show good relative discrimination (Figs. 7 and 11). However, most of the heavy precipitation events in NWRFC originate from the cool season (October–March), with substantially different, and larger, biases present in the warm season. Time-of-day effects were also found to be significant in NW- and CNRFCs (see below).
The most skillful forecasts of light to moderate precipitation are for the CNRFC area, despite the lack of reliability for PoP and very light precipitation (Figs. 9 and 6). In terms of the CR factorization, this is reflected in the strong resolution component of the BSS, which offsets the type-I conditional biases. Also, while the type-I conditional biases are high for PoP and very light precipitation, a threshold exists at which the ensemble mean is conditionally unbiased (Fig. 8). In terms of the LBR factorization, the CNRFC forecasts show good relative discrimination across a wide range of event thresholds (Fig. 7). Indeed, they are significantly more discriminatory than the climatological probability forecast for all decision thresholds in the ROC diagram (Fig. 11). Interestingly, while the forecasts from CNRFC are unreliable for PoP and very light precipitation, they are
relatively sharp (i.e., contain smaller spread than the sample climatology) when compared to the other RFCs (Fig. 7). Thus, the lack of reliability may originate from overconfident or underspread forecast probability distributions as well as a conditional bias in the ensemble mean forecast (Fig. 8). While most RFCs would benefit from statistical postprocessing to reduce the type-I conditional biases, the type-II conditional biases are more substantial, particularly for heavy precipitation in MA- and ABRFCs (see below also). However, the type-II conditional biases are more difficult to remove through statistical postprocessing, as bias correction is generally concerned with the CR factorization of the joint distribution (i.e., the type-I conditional biases).

c. Season

Seasonal verification was performed for the “warm” and “cool” seasons in each RFC. The seasons were
defined, respectively, as April–September and October–March, inclusive. Verification was performed for increasing amounts of observed precipitation, with thresholds defined at fixed climatological probabilities (see above). To compare forecast quality between seasons, the climatological probabilities were derived from the overall observed sample, ensuring fixed absolute amounts of precipitation between seasons (noting that hydrologic models respond to amounts of precipitation, rather than climatological probabilities). Figure 12 shows the correlation of the observed variable with the forecast ensemble mean for daily accumulations of 4–27, 28–51, and 52–75 h. Results are shown for the full period (top row), cool season (middle row), and warm season (bottom row). Figure 13 shows the corresponding results for the RME of the ensemble mean forecast. Figures 14 and 15 show the CR and LBR factorizations of the BSS, respectively. However, unlike Figs. 6 and 7,
Figs. 14 and 15 comprise the BSS results for the warm season only.

The correlation between the forecasts and observations depends strongly on season (Fig. 12). In general, the differences between RFCs are greatest during the cool season, when the forecast skill is better (Fig. 12). For example, the correlations are consistently weaker in CNRFC than in AB- and NWRFCs for moderate to heavy precipitation (by ~0.2 units) and consistently weaker than MARFC for PoP and light precipitation (by ~0.15 units). However, during the warm season, the correlations for PoP and light precipitation are significantly better in CNRFC, particularly in the second and third accumulation periods when spinup errors have dissipated (see Yuan et al. 2005 also). For cool season precipitation in AB-, CN-, and NWRFCs, the correlations decline reasonably smoothly with increasing precipitation threshold (Fig. 12), but decline rapidly with increasing precipitation amount in MARFC, particularly for the first 24-h accumulation.

For a given amount of precipitation, the correlations are similar across all RFCs during the warm season (with slightly higher correlations in CNRFC for light precipitation, as indicated above). However, the BSS varies more with location (Figs. 14 and 15). This is largely due to the conditional biases in the forecasts, which vary with location during both the warm and cool seasons. For example, in CN- and NWRFCs, the forecasts show larger conditional biases for light and heavy precipitation thresholds, respectively (Fig. 14). In NWRFC, the
type-I conditional biases are greatest for heavy precipitation amounts during the first two accumulation periods. As a result, the forecasts are too sharp, given the initial condition uncertainties (Fig. 15), as well as unreliable (Fig. 14). In contrast to the cool season results, the large conditional biases in the warm season for NWRFC are compounded by relatively weak resolution and discrimination, specifically at high precipitation thresholds (Fig. 15). While the forecasts from CNRFC are relatively unreliable for PoP and light precipitation amounts during the warm season, the type-II conditional biases are smaller for moderate and heavy precipitation amounts than in other RFCs (notwithstanding the sampling uncertainties; Fig. 15). Also, the forecasts are substantially sharper and more discriminatory.

The forecasts from ABRFC are relatively skillful in predicting moderate to large precipitation events during both the warm and cool seasons. This is explained by the small type-I conditional biases (see the overall results in Fig. 9 also). Indeed, in ABRFC, the conditional biases are primarily type II in origin and are similar in magnitude to those in MARFC. Conversely, the type-I conditional biases are consistently smaller in ABRFC than MARFC for moderate to heavy precipitation. These differences are more pronounced during the cool season (not shown), when the forecasts from ABRFC are significantly more reliable, and much more resolved, than...
in MARFC. Nevertheless, there is a strong type-II conditional bias in ABRFC during both the cool and warm seasons (Fig. 15), which originates from a conditional bias in the ensemble mean forecast (Fig. 13).

d. Time of day

Joint conditioning on season and time of day was restricted by the large sampling uncertainties of the verification metrics. Instead, verification was performed separately for each of the warm and cool seasons (see above) and for the “morning” and “afternoon” periods, which were defined as 0600–1800 and 1800–0600 UTC, respectively. To control for the different model initialization times—and hence forecast lead times—that occupied these morning and afternoon periods, only the 0900 and 2100 UTC initializations were verified. The 0900 and 2100 UTC forecasts produced 12-h accumulations at lead times of \{33, 45, 57, 69, 81\} hours for both the morning and afternoon periods. Results are shown for the morning, afternoon, and combined periods at a forecast lead time of 33 h. The climatological exceedance probabilities were derived from the 12-h accumulations for the combined period. Unfortunately, there was insufficient data to isolate the effects of model initialization time on forecast quality.

Figure 16 shows the correlation of the observed variable with the ensemble mean forecast, together with the ROC score and BSS (see appendix). Selected components of the CR and LBR factorizations of the BSS are shown in Fig. 17. In general, the effects of forecast valid
time are to increase correlation and BSS during the afternoon periods in AB-, NW-, and CNRFCs and to reduce correlation and BSS during the afternoon periods in MARFC, with the greatest differences for moderate and heavy precipitation amounts. The BSS factorizations show improved resolution, sharpness, and type-II conditional bias during the afternoon periods in NW- and CNRFCs (Fig. 17), with only small changes in event discrimination (Fig. 16). However, while the NWRFC shows positive skill across all precipitation thresholds, CNRFC shows negative BSS during the morning periods for moderate and heavy precipitation amounts, and little or no skill during the afternoon periods (Fig. 16). In MARFC, the loss of skill in the afternoon periods originates from an increase in reliability and a decline in resolution (not shown), or from a reduction in sharpness (Fig. 17). Overall, the effects of both time of day and season are to exaggerate the differences in forecast quality between RFCs, particularly for high precipitation thresholds.

Fig. 12. Correlation coefficient for daily precipitation totals by climatological exceedance probability for each season (row) and lead time (column).
e. Accumulation period

The forecasts were verified at three levels of temporal aggregation—namely 6, 12, and 24 h. The aggregated forecasts were derived by accumulating precipitation in each ensemble trace. For example, three 24-h precipitation totals were derived by summing the contributions from lead times of 4–27 h (day 1), 28–51 h (day 2), and 52–75 h (day 3). The corresponding 12-hourly accumulations were then 4–15 and 16–27 h (day 1), 28–39 and 40–51 h (day 2), and 52–63 and 64–75 h (day 3). To compare the different accumulations at corresponding forecast lead times and valid times, the verification results from the 6- and 12-h accumulations were averaged to the 24-h accumulation periods. Simple averaging is appropriate for additive measures—that is, measures whose outputs depend linearly on the inputs. For the correlation coefficient, averaging was applied to the coefficient of determination (i.e., the square of the correlation coefficient), which is an additive measure. Results are also shown for BSS. While the Brier score is additive, the BSS is not, in general (Jolliffe and
Stephenson 2012). However, for a given accumulation period, the climatological variance used in the BSS, \( c_p^2(1 - c_p^2) \), does not vary with forecast lead time. Thus, the aggregated BSS was computed from

\[
1 - \frac{\sum_{i=1}^{m} BS_i^a}{m \times c_p^2(1 - c_p^2)}.
\]

where \( BS_i^a \) is the \( i \)th of \( m \) BS values from which the accumulation is formed. Verification was performed for multiple thresholds in each RFC. The thresholds were obtained from an observed (climatological) sample with the same accumulation period as the forecasts.

Figure 18 shows the correlation of the observed variable with the forecast ensemble mean for 6-, 12- and 24-h accumulations at lead times of 28–51 h. Figure 19 shows the corresponding results for the BSS. For PoP and light

![Figure 18](image-url)
precipitation amounts, all RFCs show an increase in skill, both in correlation and BSS, with increasing accumulation period. In addition, NW- and ABRFCs show significant increases in BSS for moderate and heavy precipitation amounts (Fig. 19). For example, in ABRFC, the nominal value of the BSS increases from 0.01 for a 1-in-1000 (>0.001) accumulation at 6 h to 0.09 at 24 h. However, the sampling uncertainty also increases from 6 to 24 h, which is in keeping with the trade-off between scale-dependent modeling skill and sample size. In contrast, the correlations between the ensemble mean forecast and observed variable decline for moderate and heavy precipitation in CNRFC (Fig. 18), despite the significant gains in BSS (Fig. 19).

Notwithstanding the overall gains in BSS with increasing accumulation period in AB-, CN-, and NWRFCs, there are substantial and complex differences in the various components of the BSS (not shown). In terms
of the CR decomposition, the effects of aggregation are to reduce the (relative) reliability for PoP and light precipitation across all RFCs, but to significantly increase reliability for moderate to heavy precipitation in AB- and CNRFCs. In contrast, AB- and NWRFCs show little change in reliability for moderate to heavy precipitation, but much improved resolution for the 12- and 24-h accumulations. In terms of the LBR decomposition, the type-II conditional biases of the BSS are significantly reduced in NWRFC for moderate to heavy precipitation amounts (e.g., from 0.55 at 6 h to 0.4 at 24 h for $c_p = 0.01$), with smaller reductions in MA- and CNRFCs. The aggregated forecasts are also much sharper for moderate to heavy precipitation in MA-, CN-, and NWRFCs (e.g., from 0.5 at 6 h to 0.76 at 24 h for $c_p = 0.01$ in CNRFC). However, sharpness declines in all RFCs for PoP and light precipitation (e.g., 0.87 at 6 h to 0.62 at 24 h for PoP in ABRFC). Discrimination is

![FIG. 16. Selected verification scores (rows) for daily precipitation totals by climatological exceedance probability and time of day (columns).](image-url)
slightly improved in NWRFC for all precipitation thresholds, with smaller changes in other RFCs. These patterns allude to a complex relationship between forecast skill and temporal scale, which must be addressed when developing bias corrections for the SREF forecasts.

Finally, as forecast quality can depend on spatial as well as temporal scale (Pappenberger and Buizza 2009), the verification results were computed for each basin separately and ordered by increasing basin size. In practice, this led to significantly increased sampling uncertainties, particularly for the moderate to heavy precipitation amounts. Thus, while the effects of basin scale are likely to depend on complex storm characteristics and spatial attributes other than simply area (such as elevation, shape, and orientation relative to the prevailing winds, etc.), conditioning was restricted to basin size, with forecast valid time and season as secondary controls. Specifically, the verification results were
conditioned separately for the warm and cool seasons and for the morning and afternoon periods in each basin. Notwithstanding these limitations, forecast quality did not vary consistently with basin area across any RFC, precipitation threshold, season, or forecast valid time. For example, in MA- and ABRFCs, the individual basins produced very similar CRPSS (e.g., 0.38–0.42 for the ABRFC basins at 24 h). In CN- and NWRFC, these fluctuations were more pronounced, but were not consistently related to basin area.

4. Summary and conclusions

Reliable and skillful hydrometeorological forecasts are essential for operational hydrologic forecasting. To produce reliable streamflow forecasts at multiple space–time scales, the River Forecast Centers (RFCs) of the U.S. National Weather Service (NWS) are evaluating precipitation and temperature forecasts from a range of Numerical Weather Prediction (NWP) models. This paper examines the skill and biases of precipitation forecasts from NCEP’s Short-Range Ensemble Forecast system (SREF; Du et al. 2009) for 10–20 basins in each of four RFCs. Contiguous basins are chosen to increase the verification sample size via spatial pooling. The basin groups are selected to represent different climate regions—namely the middle Atlantic (MARFC), the southern plains [Arkansas–Red Basin (AB) RFC], the windward slopes of the Sierra Nevada [California Nevada (CN) RFC], and the coastal mountains of the Pacific Northwest [North West (NW) NWRFC]. For each RFC, verification results are computed conditionally upon...
forecast lead time, amount of precipitation, season, forecast valid time, and accumulation period. Limited interactions are also considered (e.g., between season and amount of precipitation), but are often constrained by the sampling uncertainties of the verification metrics. The sampling uncertainties are quantified with the stationary block bootstrap (Politis and Romano 1994).

In general, the forecast quality declines smoothly with increasing forecast lead time in all RFCs. However, in NWRFC, the conditional biases are greater for the first two lead times (6 and 12 h), particularly for heavy precipitation amounts. The forecast skill is also lower in CNRFC for the first 6-h accumulation, particularly for light to moderate precipitation amounts during the warm season. In the future, this will be addressed by improved initialization of the SREF. For example, a hybrid data assimilation system, comprising elements of the ensemble Kalman filter and three-dimensional variational assimilation, is currently being implemented at NCEP. However, the forecast skill and biases were generally more sensitive to precipitation amount than forecast lead time. For example, the forecast skill is better for moderate precipitation amounts than either PoP/light precipitation or heavy precipitation. This reflects a conditional bias in the ensemble forecasts with increasing precipitation amount. During the cool season, the forecasts from MA-, CN-, and NWRFCs overestimate PoP and light precipitation when no precipitation is observed and underestimate heavy precipitation when heavy precipitation occurs (i.e., a type-II conditional bias). Consequently, there is a precipitation threshold in each of MA-, CN-, and NWRFCs at which the forecasts are unconditionally unbiased in the ensemble mean. For this reason (among others), pooling of verification results across several or
all precipitation thresholds could be highly misleading (see Hamill and Juras 2006 also). During the warm season, light precipitation is again underforecast in MA-, CN-, and NWRFCs. However, in ABRFC, there is a large negative type-II conditional bias across all precipitation thresholds in both seasons. Here, the ensemble mean forecast is consistently too low, given the observed precipitation amount. In contrast, for moderate and heavy precipitation amounts, the type-I conditional biases are smallest in ABRFC, where the forecasts are conditionally reliable, given the sampling uncertainties. While most RFCs would benefit from statistical postprocessing to reduce the type-I conditional biases, there are large type-II conditional biases across all RFCs, but particularly for heavy precipitation in MA- and ABRFCs.

Overall, the differences between RFCs are greatest during the cool season, when the forecast skill is higher. For example, the correlations are consistently weaker in CNRFC than in AB- and NWRFCs for moderate to heavy precipitation amounts and consistently weaker than MARFC for PoP and light precipitation. While the correlations decline gradually with increasing precipitation amount in AB-, CN-, and NWRFCs, they declined rapidly with increasing precipitation in MARFC. This is driven by poor performance during the cool season in MARFC. The most skillful forecasts of heavy precipitation occur during the cool season in NWRFC, where significant conditional biases are offset by strong correlations between the ensemble mean forecast and the observed variable across all precipitation thresholds. However, during the warm season, these conditional biases are no longer offset (in terms of skill) by strong correlations. Rather they are compounded by weakened correlations, which contribute to reduced resolution and discrimination. The forecasts from ABRFC are relatively skillful in predicting moderate to large precipitation events in both seasons. However, in ABRFC, the forecast skill is greatly enhanced by the small type-I conditional biases. Thus, bias correction should be approached differently in ABRFC, where the dominant bias is type II (specifically originating from the ensemble mean) than NWRFC, where there is a large type-I conditional bias and strong correlations during the cool season. Overall, NWRFC is a good candidate for statistical postprocessing, particularly during the cool season (see Hamill and Whitaker 2006 also). While the type-I conditional biases are greatest in MARFC, particularly for heavy precipitation during the cool season, the correlations are also relatively low, which reduces the potential for statistical postprocessing. In CNRFC, the greatest benefits of postprocessing should be expected for PoP and light precipitation, specifically during the warm season (and for the second and third 24-h accumulation periods), when the type-I conditional biases are high and the correlations are strong.

For PoP and light precipitation amounts, all RFCs show some increase in skill with increasing accumulation period (from 6 to 24 h). In NW- and ABRFCs, the aggregated forecasts are also more skillful for moderate and heavy precipitation. However, the overall increases in BSS with increasing accumulation period originate from complex differences between the various components of the BSS. Consequently, the selection of an “optimal” accumulation period for statistical postprocessing (among other things) may be less straightforward than implied by the correlations alone.

In future work, several statistical postprocessors (e.g., Slaughter et al. 2007; Wilks 2009; Brown and Seo 2010) will be evaluated for their performance in removing the type-I and type-II conditional biases from the SREF precipitation forecasts. Using the NWS Hydrologic Ensemble Forecast Service (HEFS; Demargne et al. 2010), streamflow predictions will be generated with the bias-corrected precipitation forecasts and verified with the EVS. Indeed, understanding the quality of the precipitation forecasts is a necessary but only preliminary step toward understanding the hydrologic potential of the SREF, as the latter depends on the former via its complex (multiscale, multivariable, and multimodel) joint probability distribution.

Acknowledgments. This work was supported by the National Oceanic and Atmospheric Administration (NOAA) through the Advanced Hydrologic Prediction Service (AHPS) and the Climate Prediction Program for the Americas (CPPA). We thank Dingchen Hou of the National Centers for Environmental Prediction (NCEP) for providing the Climatology-Calibrated Precipitation Analysis (CCPA) dataset.

APPENDIX

Verification Scores

a. Brier score and Brier skill score

The Brier score (BS; Brier 1950) measures the mean square error of \( n \) predicted probabilities and corresponding observed outcomes that \( Q \) exceeds \( q \):

\[
BS = \frac{1}{n} \sum_{i=1}^{n} [F_X(q_i) - F_Y(q_i)]^2, \quad \text{where}
\]

\[
F_X(q_i) = \Pr(X_i > q_i) \quad \text{and}
\]

\[
F_Y(q_i) = \begin{cases} 
1, & Y_i > q_i; \\
0, & \text{otherwise}. 
\end{cases}
\]
By conditioning on the predicted probability, and partitioning over J discrete categories, the BS is decomposed into the calibration-refinement (CR) measures of type-I conditional bias or “reliability” (REL/T1), resolution (RES), and uncertainty (UNC) (see also Bradley et al. 2004):

$$\text{BS} = \frac{1}{n} \sum_{j=1}^{J} N_j [F_{Y_j}(q) - F_Y(q)]^2$$

$$- \frac{1}{n} \sum_{j=1}^{J} N_j [F_Y(q) - F_Y(q)]^2 + \sigma_Y^2(q).$$

(A2)

Here, $F_Y(q)$ represents the average relative frequency (ARF) with which the observation exceeds the threshold, $q$. The term $F_{Y_j}(q)$ represents the conditional observed ARF, given that the predicted probability falls within the $j$th of $J$ probability category, which happens $N_j$ times. A sequence of $n$ discrete probability forecasts is reliable or “calibrated” in terms of the BS if the forecast probability matches the observed ARF for those cases when the forecast is issued with probability, $F_{X_j}(q)$. The sequence is resolved in terms of the BS when the forecasts readily distinguish between different observed outcomes; that is, when the observed ARF in each forecast category is substantially different from the climatological mean. Forecast errors are penalized quadratically by the BS [Eq. (A2)], and event probabilities converge to 0 or 1 in the tails of the climatological probability distribution. Thus, other factors being equal, an extreme threshold will produce a smaller BS, and the BS will be dominated by errors from the complementary event. In terms of the latter, it is useful to interpret the trajectory of the BS over increasing values of $q$. In terms of the former, the BS may be normalized by the climatological variance, $\sigma_Y^2(q)$—that is, the uncertainty term in Eq. (A3), which leads to the Brier skill score (BSS):

$$\text{BSS} = 1 - \frac{\text{BS}}{\sigma_Y^2(q)} = \frac{\text{RES}}{\sigma_Y^2(q)} - \frac{\text{REL}}{\sigma_Y^2(q)}.$$  

(A3)

The CR factorization of the BSS implies that a sequence of $n$ discrete probability forecasts will be skillful, relative to the climatological variance, when the resolution is greater than the conditional bias—that is, when the (conditional) square bias is smaller than the (conditional) average square difference between the observed ARF and the climatological ARF.

By conditioning on the $K = 2$ two possible observed outcomes, $\{0,1\}$, the BS is decomposed into the likelihood-base rate (LBR) measures of type-II conditional bias (T2), discrimination (DIS), and sharpness (SHA):

$$\text{BS} = \left(1 - \frac{\text{SHA}}{\sigma_Y^2(q)} + \frac{\text{DIS}}{\sigma_Y^2(q)} - \frac{T2}{\sigma_Y^2(q)}\right).$$  

(A4)

Here, $F_{X_j}(q)$ represents the average probability with which $X$ is predicted to exceed $q$, given that $Y$ exceeds $q (k = 1)$ or does not exceed $q (k = 2)$, where $N_k$ is the conditional sample size for each case. The BSS is then given by

$$\text{BSS} = 1 - \frac{\text{SHA}}{\sigma_Y^2(q)} + \frac{\text{DIS}}{\sigma_Y^2(q)} - \frac{T2}{\sigma_Y^2(q)}.$$  

(A5)

The LBR factorization of the BSS implies that a sequence of $n$ discrete probability forecasts will be skillful when the forecasts are sharper than climatology (i.e., when the forecast variance is smaller than the climatological variance), and when the forecasts are more discriminatory, relative to climatology, than conditionally biased (see Bradley et al. 2004 also).

b. Continuous ranked probability score and skill score

The continuous ranked probability score (CRPS) measures the integral square difference between the cumulative distribution functions of the observed and predicted variables:

$$\text{CRPS} = \int [F_X(q) - F_Y(q)]^2 \, dq.$$  

(A6)

In practice, the CRPS is averaged across $n$ of pairs of forecasts and observations. The continuous ranked probability skill score (CRPSS) comprises a ratio of the mean CRPS for the main prediction system, CRPS, and a reference system, $\text{CRPS}_{\text{REF}}$:

$$\text{CRPSS} = \frac{\text{CRPS} - \text{CRPS}_{\text{REF}}}{\text{CRPS}_{\text{REF}}}.$$  

(A7)

c. Relative operating characteristic score

The relative operating characteristic (ROC; Green and Swets 1966) measures the trade-off between correctly
forecasting that a discrete event will occur (probability of detection or POD) and incorrectly forecasting that it will occur (probability of false detection or POFD). This trade-off is expressed as a decision threshold, \( d \), at which the forecast probability triggers some action. The ROC plots the POD versus the POFD for all possible values of \( d \) in \([0, 1]\). For a particular threshold, the empirical POD is

\[
POD = \frac{\sum_{i=1}^{n} I_Y(F_X(q) > d | Y > q)}{\sum_{i=1}^{n} I_Y(Y > q)},
\]

where \( I \) denotes the indicator function. The empirical POFD is

\[
POFD = \frac{\sum_{i=1}^{n} I_Y(F_X(q) > d | Y \leq q)}{\sum_{i=1}^{n} I_Y(Y \leq q)}.
\]

Here, the relationship between the POD and POFD is assumed bivariate normal (Hanley 1988; Metz and Pan 1999):

\[
POD = \Phi[a + b\Phi^{-1}(POFD)], \quad a = \frac{\mu_{POD} - \mu_{POFD}}{\sigma_{POD}} \quad \text{and} \quad b = \frac{\sigma_{POFD}}{\sigma_{POD}},
\]

and \( \Phi \) is the cumulative distribution function of the standard normal distribution. The means of the POD and POFD are \( \mu_{POD} \) and \( \mu_{POFD} \), respectively, and their corresponding standard deviations are \( \sigma_{POD} \) and \( \sigma_{POFD} \). The ROC score measures the area under the ROC curve (AUC) after adjusting for the climatological base rate (i.e., ROC score = 2AUC - 1). The AUC is an analytical function of the binormal model parameters

\[
AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right).
\]

Calculation of the fitted ROC score amounts to estimating the parameters, \( a \) and \( b \), of the linear relationship between the POD and the POFD in normal space, for which ordinary least squares regression was used.

REFERENCES


