Use of Medium-Range Numerical Weather Prediction Model Output to Produce Forecasts of Streamflow

MARTYN P. CLARK

Center for Science and Technology Policy Research, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

LAUREN E. HAY

U.S. Geological Survey, Denver, Colorado

(Manuscript received 20 March 2003, in final form 25 August 2003)

ABSTRACT

This paper examines an archive containing over 40 years of 8-day atmospheric forecasts over the contiguous United States from the NCEP reanalysis project to assess the possibilities for using medium-range numerical weather prediction model output for predictions of streamflow. This analysis shows the biases in the NCEP forecasts to be quite extreme. In many regions, systematic precipitation biases exceed 100% of the mean, with temperature biases exceeding 3°C. In some locations, biases are even higher. The accuracy of NCEP precipitation and 2-m maximum temperature forecasts is computed by interpolating the NCEP model output for each forecast day to the location of each station in the NWS cooperative network and computing the correlation with station observations. Results show that the accuracy of the NCEP forecasts is rather low in many areas of the country. Most apparent is the generally low skill in precipitation forecasts (particularly in July) and low skill in temperature forecasts in the western United States, the eastern seaboard, and the southern tier of states. These results outline a clear need for additional processing of the NCEP Medium-Range Forecast Model (MRF) output before it is used for hydrologic predictions.

Techniques of model output statistics (MOS) are used in this paper to downscale the NCEP forecasts to station locations. Forecasted atmospheric variables (e.g., total column precipitable water, 2-m air temperature) are used as predictors in a forward screening multiple linear regression model to improve forecasts of precipitation and temperature for stations in the National Weather Service cooperative network. This procedure effectively removes all systematic biases in the raw NCEP precipitation and temperature forecasts. MOS guidance also results in substantial improvements in the accuracy of maximum and minimum temperature forecasts throughout the country. For precipitation, forecast improvements were less impressive. MOS guidance increases the accuracy of precipitation forecasts over the northeastern United States, but overall, the accuracy of MOS-based precipitation forecasts is slightly lower than the raw NCEP forecasts.

Four basins in the United States were chosen as case studies to evaluate the value of MRF output for predictions of streamflow. Streamflow forecasts using MRF output were generated for one rainfall-dominated basin (Alapaha River at Statenville, Georgia) and three snowmelt-dominated basins (Animas River at Durango, Colorado; East Fork of the Carson River near Gardnerville, Nevada; and Cle Elum River near Roslyn, Washington). Hydrologic model output forced with measured-station data were used as ''truth'' to focus attention on the hydrologic effects of errors in the MRF forecasts. Eight-day streamflow forecasts produced using the MOS-corrected MRF output as input (MOS) were compared with those produced using the climatic Ensemble Streamflow Prediction (ESP) technique. MOS-based streamflow forecasts showed increased skill in the snowmelt-dominated river basins, where daily variations in streamflow are strongly forced by temperature. In contrast, the skill of MOS forecasts in the rainfall-dominated basin (the Alapaha River) were equivalent to the skill of the ESP forecasts. Further improvements in streamflow forecasts require more accurate local-scale forecasts of precipitation and temperature, more accurate specification of basin initial conditions, and more accurate model simulations of streamflow.

1. Introduction

Rapid population growth and economic development, along with changing social demands on freshwater resources, have imposed new challenges on water management in many regions in the United States. Managers must balance the need to retain as much water as possible in reservoirs to meet the needs of irrigation, hydropower generation, and domestic consumption, along with needs such as ensuring an adequate supply of water for recreational uses, as well as meeting stringent water quality standards, regulations for maintenance of aquatic

Corresponding author address: Dr. Martyn P. Clark, Center for Science and Technology Policy Research, University of Colorado, 1333 Grandview Ave., UCB 488, Boulder, CO 80309-0488.
E-mail: clark@vorticity.colorado.edu

ecosystems, and the special needs for the protection of threatened or endangered species. Reservoir space also must be maintained to protect downstream homes, farms, and businesses from flooding.

Accurate streamflow forecasts can play a key role in optimizing the use of water. Traditionally, hydrologic forecasts in the United States have been made using the climatic Ensemble Streamflow Prediction (ESP) procedure (Day 1985). In this approach, a hydrologic model is driven with observed precipitation and temperature data up to the beginning of the forecast to estimate basin initial conditions. Then precipitation and temperature data for the same date from every other year in the historical record are used to produce ensemble forecasts of streamflow. For example, an 8-day forecast initialized on 1 January 2004 could use station observations from 2 to 9 January 1950 as inputs for ensemble 1, station observation from 2 to 9 January 1951 as inputs for ensemble 2 . . . and station observations from 2–9 January 1999 as inputs for ensemble 50. When these ensembles are run through a hydrologic model, the method provides an ensemble of possible streamflow given the antecedent conditions (e.g., soil moisture, water equivalent of the accumulated snowpack) at the start of the forecast. Forecast accuracy is therefore dependent on accurate specification of conditions over the basin at the start of the forecast and the influence of those conditions on the basin hydrologic response. Accuracy also is dependent on the similarity between future weather conditions and the ensembles of historic data from previous years. This approach works well in river systems where substantial lag times are introduced because of storage of water in snowpack or subsurface and ground-water reservoirs. However, because the methodology of ESP weights equally the history for each year in the historical record, the approach often yields a wide range of possible outcomes and low probabilistic forecast skill.

A number of studies have suggested that it is possible to improve the accuracy of probabilistic streamflow forecasts by including in the ESP approach information from meteorological forecasts and climate outlooks [e.g., see the plans for an Advanced Hydrologic Prediction System (AHPS) by the National Weather Service in the United States (Connelly et al. 1999)]. As a first step in this direction, Hamlet and Lettenmaier (1999) modified the ESP approach by restricting ensemble members to years that are similar in terms of the phase of the El Niño–Southern Oscillation (ENSO) and the phase of the Pacific decadal oscillation (PDO). In most cases this provides a set of ensembles that are more tightly clustered than the full ensemble. On shorter time scales, further reductions in ensemble spread may be realized by replacing the ensemble of data from previous years with output from atmospheric forecast models.

This paper explores the utility of atmospheric forecasts for hydrologic predictions on time scales of up to 8 days. This study first evaluates the systematic biases and the accuracy in Medium-Range Forecast Model (MRF) predictions of precipitation and temperature over the contiguous United States and introduces procedures to improve raw MRF output through downscaling. Results are based on the 40+ yr archive of 8-day atmospheric forecasts from the National Centers for Environmental Prediction (NCEP) reanalysis project (described later). As an example application of this approach, this study assesses the hydrologic forecast accuracy obtained when forcing a distributed-hydrologic model with the MRF output for four basins across the contiguous United States.

2. The NCEP forecast archive

a. Project overview

The NCEP reanalysis project (Kalnay et al. 1996; Kistler et al. 2001) produced a retroactive 40+ year record of global atmospheric fields and surface fluxes derived from a numerical weather prediction and data assimilation system kept unchanged over the analysis period. Use of a fixed model eliminates pseudoclimate jumps in archived time series associated with frequent upgrades in the operational modeling system used at NCEP and allows an assessment of the accuracy of a Numerical Weather Prediction (NWP) model over a long time period. However, temporal inconsistencies can still be present because of changes through time in the amount, type, and quality of the available assimilation data. The model used for the reanalysis is identical to the Medium-Range Forecast Model implemented operationally at NCEP in January 1995, except that the horizontal resolution is twice as coarse in the reanalysis version. Every 5 days, a single realization of an 8-day atmospheric forecast was run. For the period 1958–98, this provides more than 2500 8-day forecasts that can be compared with observations.

b. NCEP Medium-Range Forecast Model description

The NCEP reanalysis is performed with a T62 model (approximately 1.9° horizontal resolution) with 28 vertical sigma levels and the spectral statistical interpolation (SSI) for assimilation (Kalnay et al. 1996). Assimilation data are formatted into a common standard World Meteorological Organization (WMO) binary universal format representation (BUFR) and then evaluated by quality control procedures (Dey and Morone 1985; Woollen et al. 1994; DiMego 1988; Kalnay et al. 1996). Data sources include rawinsonde profiles, surface marine reports from the Comprehensive Ocean–Atmosphere Data Set (COADS), aircraft observations of wind and temperature, synoptic reports of surface pressure over land, vertical temperature profiles from the Television Infrared Operation Satellite (TIROS) Operational Vertical Sounder (TOVS) over the ocean, TOVS temperature sounding over land above 100 hPa, surface wind speeds from the Special Sensor Microwave Imager (SSM/I) and satellite cloud drift winds.

Two types of precipitation are computed: convective and grid scale (dynamic). Convection is based on a simplified Arakawa–Schubert scheme (Pan and Wu 1994) that was found to result in improved prediction of precipitation over the continental United States and the tropics as compared to the previous Kuo parameterization (Kalnay et al. 1996). Dynamic precipitation is parameterized by starting at the top layer and checking for supersaturation. If supersaturated, latent heat is released to adjust the specific humidity and temperature to saturation, with the excess water falling to the next lower layer. If this next layer is supersaturated then adjustment to saturation occurs again, and the amount of precipitation is added to that from the higher layer. However, if the layer is unsaturated, some or all of the precipitation is evaporated. The process continues downward with all precipitation that penetrates to the bottom layer allowed to fall to the surface.

3. Station data

This study uses daily precipitation and maximum and minimum temperature data from a network of over 11 000 National Weather Service (NWS) manual cooperative climate observing stations across the contiguous United States. These data were extracted from the National Climatic Data Center (NCDC) Summary of the Day Dataset by J. Eischeid, National Oceanic and Atmospheric Administration (NOAA) Climate Diagnostics Center, Boulder, Colorado (Eischeid et al. 2000). Quality control performed by NCDC includes the procedures described by Reek et al. (1992) that flag questionable data based on checks for (a) outliers, based on extreme values defined for each state; (b) internal consistency among variables (e.g., maximum temperature less than minimum temperature); (c) constant temperature (e.g., 5 or more days with the same temperature are suspect); (d) excessive diurnal temperature range; (e) invalid relations between precipitation, snowfall, and snow depth; and (f) unusual spikes in temperature time series. Records at most of these stations start in 1948 and continue through 1998. We restrict use to only the ''best'' stations in the Eischeid archive. These are defined as those with less than 10% missing or questionable data during the period 1958–99.

Observation times for stations in the co-op network are mixed. Some co-op observers take measurements in the morning, some observers take measurements in the afternoon, and some observers take measurements in the evening. Specific observation times sometimes vary through time, and are not known for all stations. To address these inconsistencies, the forecast model output was averaged for the three 12-h periods surrounding the day of the observation.

4. Accuracy of the NCEP model

a. Systematic model biases

As a first step in evaluating the accuracy of the NCEP model, the systematic biases in NCEP temperature and precipitation forecasts are examined. Mean biases are evaluated using monthly climatologies of precipitation and temperature derived from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) system (Daly et al. 1994). PRISM uses multiple linear regression techniques to distribute monthly climatologies of precipitation and temperature from a dense network of stations to a 2-km digital elevation model (DEM) over the contiguous United States. The PRISM climatologies are available commercially from Oregon State University's Spatial Climate Analysis Service. In this analysis, the PRISM climatologies were regridded to the NCEP Gaussian grid over the contiguous United States using an average of values from all PRISM grids within each NCEP grid box. The elevation of the resampled-PRISM DEM matches the elevation of the NCEP grid almost perfectly (not shown) and avoids introducing artificial biases associated with differences between the elevation of model grid points and the elevation of individual stations in the NCDC archive (e.g., see Briggs and Cogley 1996). Corresponding climatologies (1961–90) for the NCEP model were computed by averaging the 12-hourly data for each day in the forecast cycle. The 12-h MRF output represents an average precipitation rate and average temperature for the 12 h prior to the forecast time.

Systematic biases in precipitation and 2-m air temperature are summarized for the months of January and July in Fig. 1. The figure shows biases for day+0 (top row) and then for each subsequent forecast lead time. Precipitation biases are expressed as a percentage of the observed (PRISM) mean. On day+0 when the NCEP model atmosphere is strongly constrained by observed data, significant biases are evident in both precipitation and 2-m temperature. Most apparent are the negative temperature biases in the western United States in January and the previously documented positive precipitation biases in the southeastern United States in July (e.g., see Janowiak et al. 1998; Trenberth and Guillemot 1998). The bias characteristics change with forecast lead time (Fig. 1). In the most general sense this reflects the NCEP model ''drifting'' away from the observed climate toward the model's climate. In January, note the evolution of positive precipitation biases over the western Great Plains, the reduction of the negative day+0 temperature biases over the western United States, and the emergence of positive temperature biases over the northern Great Plains. Of note for July is the disappearance of the positive precipitation biases in the southeastern United States and the strengthening of negative temperature biases in the same region.

These biases are presented as an example of substantial biases in state-of-the-art NWP models. Exam-

FIG. 1. Systematic biases in NCEP forecasts of precipitation and 2-m air temperature, showing biases for day+0 (top row) and then biases for each subsequent forecast lead time. Precipitation biases are expressed as a percentage of the PRISM climatology, and temperature biases are expressed as a departure from the PRISM climatology (°C).

ples of biases in other NWP models and other regions are provided for day+0 output by Chelliah and Ropelewski (2000), Serreze and Hurst (2000), Reid et al. (2001), and Hagemann and Gates (2001). Chelliah and Ropelewski (2000) compared tropospheric temperature from the Microwave Sounding Unit Channel 2 (MSU Ch2) with estimates from the NCEP reanalysis, the European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis, and the National Aeronautics and Space Administration Data Assimilation Office (NASA DAO) reanalysis. NCEP and ECMWF temperatures averaged over a near-global domain (80°N–80°S) were ap-

proximately 2°C higher than the MSU Ch2 values, with those from the NASA DAO reanalysis exhibiting positive biases of 3°C. Serreze and Hurst (2000) diagnosed problems with the NCEP and ECMWF reanalysis in simulating monthly Arctic precipitation. They found that both models underestimate precipitation over the Atlantic side of the Arctic. The most significant problem is a large overprediction of summertime convective precipitation over Arctic land areas in the NCEP reanalysis. Hagemann and Gates (2001) used NCEP and ECMWF reanalysis output to drive a hydrologic model for several large river basins throughout the globe. Of particular note was a wintertime cold bias in the ECMWF 2-m temperatures over high latitudes, resulting in a delay in spring streamflow. Excessive summer precipitation over Northern Hemisphere land areas in the NCEP reanalysis resulted in positive streamflow biases. Such biases need to be removed before NWP model output can be used in hydrologic applications.

b. Forecast accuracy

If biases are systematic, the NCEP model may still have considerable skill in forecasting day-to-day variations in precipitation and temperature. This is indicated by the results of Kalnay et al. (1998), who used the same forecast archive that is used in this study and showed that the NCEP model has appreciable skill in forecasting daily variability in 500-hPa height up to forecast lead times of 8 days. The accuracy of NCEP precipitation and 2-m maximum temperature forecasts is computed by interpolating the NCEP model output for each forecast day to the location of each station in the NWS cooperative network (see section 2b) and computing an appropriate skill score. The skill of precipitation forecasts is measured by Spearman rank correlations, and the skill of the 2-m maximum temperatures is measured by the explained variance using squared Pearson correlations (i.e., the r^2 value). Spearman rank correlations are more appropriate than Pearson correlations when normal distributions cannot be assumed (as is the case for daily precipitation). We use Kupier's skill score (Wilks 1995) to evaluate the accuracy of precipitation occurrence predictions, that is, how well forecasted wet (dry) days match observed wet (dry) days. To avoid the possibility of spuriously high correlations that can result from matching zero-precipitation days, Spearman rank correlations for precipitation are only computed for days when both the station and NCEP model report precipitation. This reduces the number of days for analysis, particularly in dry regions and at longer forecast lead times when precipitation occurrence in the NCEP model is poorly matched with precipitation occurrence in observed records. Skill scores are only computed if there are more than 50 valid days available for analysis. More details on the skill scores are provided in the appendix.

The intent of comparisons between the NCEP model output and station observations is not to assess the true model skill (which would be done by interpolating the station observations to the NCEP grid, with topographic corrections) but to assess the potential utility of the NCEP MRF output at the local scales important for water resource applications. The goal of this study is to determine if global-scale forecast models contain useful local-scale information. Note however that the Pearson and Spearman correlation statistics are not sensitive to differences in the mean (appendix), so the effects of differences between grid-box and station elevations are reduced.

The accuracy of NCEP precipitation and 2-m maximum temperature forecasts is presented in Fig. 2 for the months of January and July. The skill at each individual station is represented by a colored dot. The NCEP model is shown to capture important aspects of day-to-day variations in precipitation and 2-m maximum temperature. In particular, note the modest skill in January precipitation forecasts over California and the upper Midwest states at the beginning of the forecast cycle and the high skill in January 2-m maximum temperature forecasts (through to day+4) over the eastern half of the United States. In July, the skill of precipitation forecasts is rather low across the entire country, but the 2-m maximum temperature forecasts exhibit high skill over the Pacific Northwest and the Midwest states where there is a high frequency of summertime clear days. Low forecast skill is evident in January for precipitation throughout the Rocky Mountains. The 2-m maximum temperature forecasts in January exhibit low skill over the Rocky Mountains and Appalachians, and in July the 2-m maximum temperature forecasts have lower skill east of the Mississippi River. In all cases precipitation forecasts have much lower skill than the 2-m maximum temperature forecasts.

The generally poor precipitation forecasts limit the use of global-scale NWP model output in river basins where the surface hydrology is dominated by rainfall. In river basins dominated by snowmelt (where the surface hydrology is controlled by variations in temperature), difficulties in providing accurate precipitation forecasts are less important. The case studies presented later in this paper demonstrate that while precipitation forecasts from the NCEP global-scale NWP model are not accurate enough to provide credible predictions of streamflow in river basins where the surface hydrology is dominated by rainfall, the NCEP temperature forecasts do provide useful predictions of streamflow in river basins dominated by snowmelt.

5. Improvement of raw NCEP NWP output

a. Background

Given the large systematic biases in the NCEP model and the poor skill in precipitation and 2-m air temperature forecasts in some regions, it is necessary to use

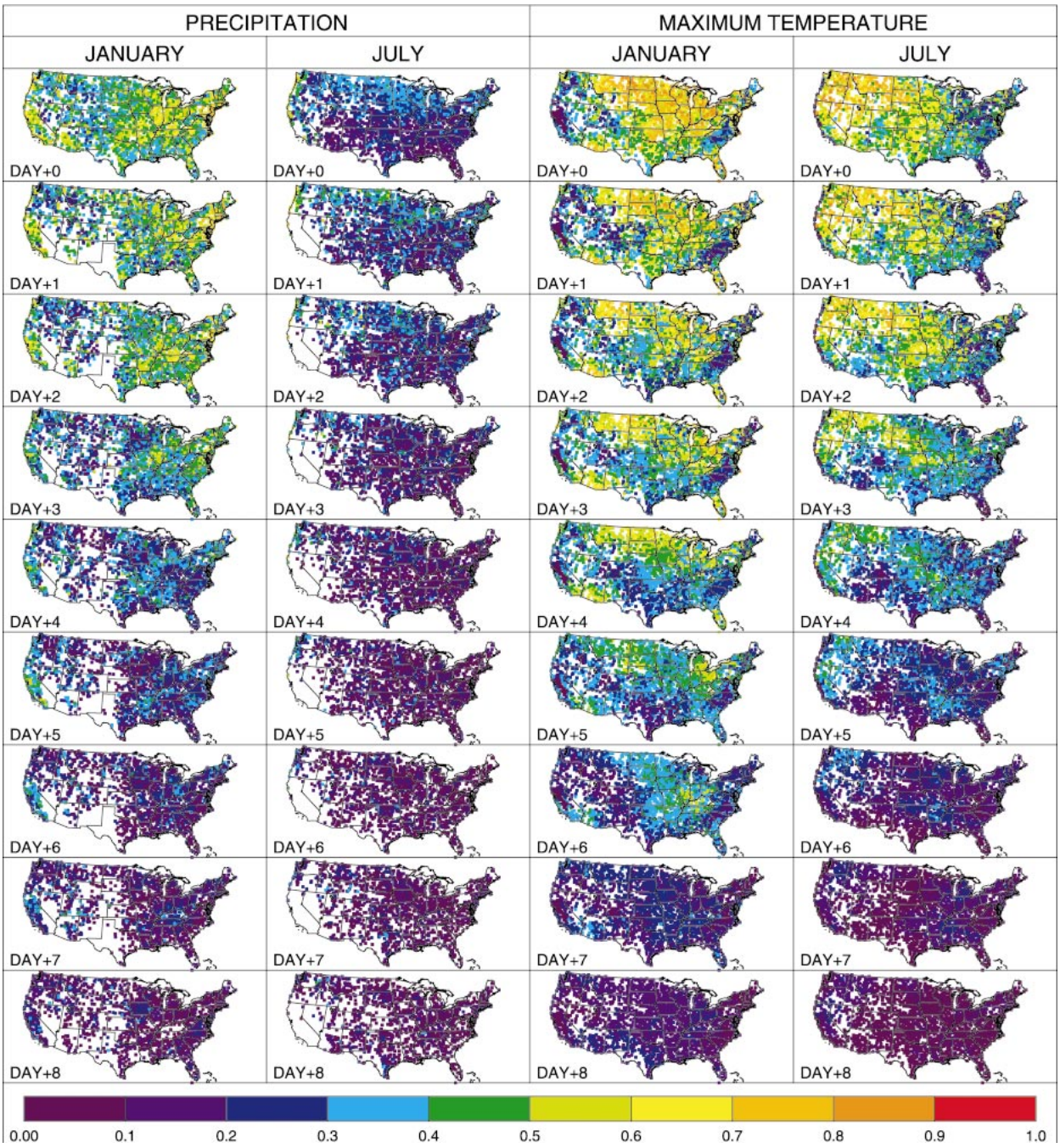

FIG. 2. Accuracy of the raw NCEP precipitation and 2-m max temperature forecasts, showing forecast skill for day+0 (top row) and then skill for each subsequent forecast lead time. Forecast skill for precipitation forecasts is assessed using Spearman rank correlations, and forecast skill for 2-m max temperature forecasts is assessed using squared Pearson correlations (r^2).

methods that may improve upon the raw forecasts. The technique of model output statistics (MOS; e.g., Glahn and Lowry 1972; Antolik 2000) may be useful for this purpose. MOS downscaling approaches develop empirical relations between gridpoint values of NWP model output (e.g., vertical velocity, total column precipitable water, static stability) and observed data. An advanced MOS system was entered in the 1996–97 National Collegiate Weather Forecasting Contest and finished better than approximately 97% of the human forecasters who entered the contest (Vislocky and Fritsch 1997). The disadvantage of MOS is that the MOS equations must be developed using an archive of forecasts from the same model that is used in the operational setting. The

practice at NCEP and other modeling centers is to frequently implement a new improved version of the operational model, meaning that the length of the forecast archive from the operational model may be too short to develop reliable MOS equations.

b. MOS technique

In the MOS technique used in this study, variables included in the NCEP forecast archive were used as predictors in a multiple linear regression approach to forecast precipitation occurrence, precipitation amounts, maximum temperature, and minimum temperature for stations in the National Weather Service cooperative network (Fig. 2). The MOS technique used in this study includes three main steps: preprocessing of the station data, development of the regression equations, and application of the regression equations—including stochastic modeling of the regression residuals to generate ensemble forecasts.

For the first step, the station time series of precipitation are preprocessed. For precipitation occurrence, the daily precipitation data at a given station are converted to a binary time series of 1's (wet days) and 0's (dry days); the regression equation thus predicts the probability of precipitation (see also Antolik 2000). For precipitation amount, the station precipitation data (only wet days) is transformed to a normal distribution using a nonparametric probability transform (e.g., Panofsky and Brier 1963). To do this, we compute the cumulative probability of observed precipitation (based on the ranked time series) and the cumulative probability of a standard normal distribution (mean of zero and standard deviation of one). The cumulative probability of each daily precipitation total in the observed record is matched with the cumulative probability in the standard normal distribution, and the precipitation value is replaced with the corresponding z score. For example, a precipitation value of 16.25 mm may have a cumulative probability of 0.84 and correspond to a z score of 1.0. The ranked daily precipitation data for the dependent sample is saved for a later retransform of the downscaled precipitation predictions. In the retransform, a linear interpolation is generally necessary because the cumulative probability of the downscaled z score lies between the cumulative probability of two of the ranked observed values. In rare cases when the cumulative probability of the downscaled z score is smaller (larger) than the lowest (highest) cumulative probability in the ranked observed time series, it is ascribed the lowest (highest) observed value in the dependent time period.

Multiple linear regression with forward selection is used to develop the MOS equations (Antolik 2000). The forward selection procedure first identifies the predictor variable (e.g., total column precipitable water), which explains the most variance of the predictand (e.g., precipitation at a station location). It then searches through the remaining variables and selects the variable that reduces the largest portion of the remaining unexplained variance in combination with the variable already chosen. If the improvement in explained variance exceeds a given threshold (taken here as 1%), the variable is included in the multiple linear regression equation. The remaining variables are examined in the same way until no further improvement is obtained based on the correlation threshold. The MOS equations are developed over the period 1958–76 and validated over the period 1977–98, which represent two different climate regimes over the North Pacific Ocean and North America (Mantua et al. 1997). A separate regression equation is developed for each station, each forecast lead time, and each month.

To provide a fairly complete description of forecasted atmospheric conditions, a large pool of potential predictor variables is tested in the multiple linear regression model (Table 1). Predictor variables from the NCEP archive include geopotential height, temperature, wind, and humidity at five pressure levels (300, 500, 700, 850, and 1000 hPa), various surface flux variables (e.g., downwelling shortwave radiation flux, 24-h accumulated precipitation), and derived variables such as vorticity advection, zonal and meridional moisture fluxes, and stability indices. All predictor variables are taken from grid boxes within a 500-km search radius of the station being modeled and interpolated to the station location using Cressman (inverse distance) interpolation. Grid-binary predictors are also used (Jensenius 1992; Antolik 2000). Gridded fields of the downwelling shortwave radiation flux, the precipitation rate, 850-hPa relative humidity, and the modified-K stability index (Charba 1977; Peppler and Lamb 1989) were compared against threshold values (in this case, tercile values for each quantity, computed separately for each month). For each grid point, each day with a data value above a given threshold was assigned a ''1,'' and each day below that threshold was assigned a ''0.'' These gridded binary fields were then interpolated to the station locations in an identical manner to the standard and computed predictors, providing a continuous predictor time series bounded by zero and one. All predictor variables are lagged to account for possible temporal phase errors in the atmospheric forecasts. For example, regression models developed for forecast day+3 include variables at forecast lead times of 48, 60, 72, 84, and 96 h. In Colorado, mountain standard time, hour+72 corresponds to 1700 local time 3 days from the start of the forecast for variables reporting a snapshot of atmospheric conditions (e.g., 500-hPa height), and hour+72 corresponds to the period 0500 to 1700 local time 3 days from the start of the forecast for variables reporting 12-h averages (e.g., downwelling shortwave radiation flux).

Cross-validation procedures are used to avoid overspecification of the regression equation or chance selection of a set of insignificant variables (Michaelsen 1987; Allen 1971). The dependent sample (1958–76) is randomly broken into two periods. For a given com-

TABLE 1. Percentage of equations that a given predictor variable is selected for use in the MOS system, and (in parentheses) the percentage of equations that a given variable is the first variable for MOS-based predictions of max temperature (TMAX), min temperature (TMIN), precipitation occurrence (POCC), and precipitation amounts (PRCP). Data are aggregated over all stations, all forecast lead times, and for the five lagged time periods. See text for more details. Downwelling shortwave radiation flux is only used for daytime forecast periods (0000–1200 UTC). High-pass-filtered 500- and 1000-hPa height is computed using all time periods (i.e., midpoint minus the mean of all five periods).

	Jan				Jul			
	TMAX	TMIN	POCC	PRCP	TMAX	TMIN	POCC	PRCP
Air temperature 300 hPa	1.74 (0.01)	1.76 (0.03)	3.15 (0.29)	5.23 (1.39)	1.66 (0.74)	4.27 (2.24)	2.93 (0.37)	4.57 (1.22)
Air temperature 500 hPa	2.07 (0.06)	1.79 (0.03)	2.50 (0.14)	2.97 (0.62)	3.54 (2.35)	4.35 (2.70)	3.57 (0.64)	3.56 (0.98)
Air temperature 700 hPa	4.51 (1.98)	2.82 (0.62)	2.93 (0.25)	2.84 (0.61)	5.12 (4.06)	3.79 (2.37)	2.46 (0.26)	3.26 (0.80)
Air temperature 850 hPa	12.87 (9.14)	8.35 (5.48)	2.76 (0.16)	3.11 (0.67)	13.32 (11.75)	9.12 (7.23)	2.79 (0.30)	2.72 (0.67)
Air temperature 1000 hPa	42.06 (34.62)	38.37 (31.75)	2.59 (0.26)	2.95 (0.68)	20.98 (19.07)	10.05 (8.21)	3.37 (0.55)	3.32 (0.80)
Relative humidity 300 hPa	1.85 (0.01)	1.77 (0.01)	2.97 (0.23)	4.54 (1.10)	0.77 (0.12)	1.12 (0.06)	3.70 (0.88)	5.58 (1.84)
Relative humidity 500 hPa	2.41 (0.01)	1.85 (0.01)	3.88 (0.44)	4.78 (1.00)	1.38 (0.20)	1.45 (0.12)	9.96 (2.27)	6.26 (2.17)
Relative humidity 700 hPa	3.62 (0.00)	2.11 (0.00)	6.99 (1.23)	4.57 (1.34)	1.72 (0.66)	1.12 (0.08)	9.43 (4.01)	4.73 (1.72)
Relative humidity 850 hPa	4.72 (0.01)	2.98 (0.02)	3.79 (0.68)	3.56 (0.57)	2.66 (1.41)	1.63 (0.38)	4.55 (1.90)	4.41 (1.59)
+ Relative humidity 850 hPa (grid binary 1)	2.93 (0.02)	2.50 (0.01)	3.15 (0.37)	3.75 (0.48)	1.23 (0.37)	1.15 (0.14)	2.62 (0.16)	5.23 (1.47)
+ Relative humidity 850 hPa (grid binary 2)	3.58 (0.01)	2.92 (0.01)	4.29 (1.16)	3.74 (0.65)	1.88 (0.80)	1.38 (0.27)	3.41 (0.33)	5.45 (1.88)
+ Relative humidity 850 hPa (grid binary 3)	3.95 (0.01)	2.54 (0.00)	6.28 (2.23)	4.12 (0.85)	2.16 (0.85)	1.47 (0.30)	3.39 (0.66)	5.23 (2.00)
Relative humidity 1000 hPa	4.99 (0.02)	3.23 (0.04)	3.76 (0.93)	4.36 (1.13)	2.74 (1.53)	1.99 (0.40)	7.52 (4.41)	4.43 (1.59)
Geopotential height 300 hPa	2.36 (0.37)	1.95 (0.02)	3.16 (0.40)	2.86 (0.56)	4.89 (3.78)	6.62 (4.95)	3.86 (0.80)	2.69 (0.63)
Geopotential height 500 hPa	3.30 (0.90)	1.79 (0.03)	3.87 (0.78)	2.68 (0.54)	3.32 (2.35)	2.36 (1.09)	3.99 (1.15)	2.80 (0.67)
Geopotential height 700 hPa	2.55 (0.37)	2.01 (0.00)	5.08 (1.51)	3.48 (0.88)	1.24 (0.48)	1.13 (0.15)	4.16 (1.28)	3.21 (0.77)
Geopotential height 850 hPa	2.08 (0.05)	1.74 (0.00)	5.18 (1.69)	3.93 (1.30)	1.05 (0.31)	1.23 (0.12)	3.83 (1.02)	3.74 (1.18)
Geopotential height 1000 hPa	2.30 (0.01)	2.05 (0.09)	4.43 (1.52)	4.50 (1.77)	1.48 (0.74)	1.77 (0.53)	2.94 (0.57)	4.04 (1.33)
Zonal wind 300 hPa	2.55 (0.11)	2.94 (0.02)	4.08 (0.30)	4.47 (1.11)	1.83 (0.90)	2.11 (0.75)	4.76 (1.29)	4.47 (1.65)
Zonal wind 500 hPa	2.62 (0.06)	3.32 (0.01)	3.42 (0.15)	4.39 (1.19)	1.45 (0.53)	1.93 (0.48)	4.65 (1.11)	3.83 (1.09)
Zonal wind 700 hPa	2.14 (0.04)	3.74 (0.03)	3.19 (0.35)	4.53 (1.33)	1.25 (0.30)	1.31 (0.10)	3.61 (0.69)	3.58 (0.90)
Zonal wind 850 hPa	2.67 (0.06)	2.77 (0.08)	3.20 (0.57)	4.37 (1.06)	0.95 (0.14)	1.06 (0.06)	3.93 (0.63)	3.52 (0.91)
Zonal wind 1000 hPa	6.72 (0.27)	3.58 (0.15)	4.60 (0.82)	4.21 (0.82)	1.59 (0.56)	1.32 (0.14)	3.33 (0.43)	4.54 (1.20)
Meridional wind 300 hPa	1.89 (0.01)	2.01 (0.05)	4.53 (1.00)	5.72 (2.46)	0.97 (0.18)	1.65 (0.29)	3.96 (0.72)	4.23 (1.41)
Meridional wind 500 hPa	1.94 (0.03)	2.09 (0.07)	4.52 (1.05)	4.84 (1.94)	0.70 (0.11)	1.36 (0.12)	3.45 (0.57)	3.86 (1.16)
Meridional wind 700 hPa	2.07 (0.06)	2.36 (0.06)	4.68 (1.34)	4.61 (1.88)	0.89 (0.10)	1.36 (0.15)	2.69 (0.24)	4.30 (1.34)
Meridional wind 850 hPa	3.25 (0.02)	3.03 (0.01)	3.48 (0.57)	4.08 (1.28)	1.05 (0.15)	1.24 (0.15)	2.81 (0.23)	3.85 (1.33)
Meridional wind 1000 hPa	4.54 (0.01)	4.97 (0.01)	3.33 (0.23)	3.95 (0.84)	1.23 (0.17)	1.48 (0.18)	2.98 (0.34)	4.30 (1.37)
Relative vorticity 300 hPa	1.71 (0.00)	1.93 (0.01)	2.84 (0.15)	3.82 (0.84)	0.96 (0.22)	1.33 (0.20)	3.17 (0.28)	5.15 (1.40)
Relative vorticity 500 hPa	2.16 (0.00)	1.97 (0.00)	3.31 (0.38)	3.71 (0.71)	1.31 (0.37)	1.24 (0.13)	2.82 (0.15)	4.25 (1.26)
Relative vorticity 700 hPa	2.11 (0.00)	1.75 (0.01)	4.00 (0.77)	4.32 (1.13)	1.15 (0.27)	1.23 (0.13)	3.68 (0.44)	4.82 (1.59)
Relative vorticity 850 hPa	2.86 (0.01)	2.26 (0.01)	3.59 (0.44)	4.02 (0.89)	1.12 (0.15)	1.07 (0.07)	5.44 (0.92)	5.03 (1.76)
Relative vorticity 1000 hPa	3.56 (0.02)	3.87 (0.01)	3.00 (0.23)	3.97 (0.78)	0.86 (0.14)	1.19 (0.07)	4.89 (1.14)	5.75 (1.78)
Modified K (Mod-K) stability index	2.21 (0.08)	2.51 (0.18)	2.65 (0.07)	3.24 (0.37)	1.32 (0.40)	1.97 (0.44)	2.58 (0.38)	4.25 (1.23)
+ Mod-K stability index (grid binary 1)	2.19 (0.02)	2.63 (0.10)	2.47 (0.15)	3.63 (0.55)	0.91 (0.16)	1.32 (0.27)	3.49 (0.55)	4.59 (1.38)
+ Mod-K stability index (grid binary 2)	2.27 (0.12)	3.27 (0.47)	3.28 (0.57)	3.25 (0.75)	0.94 (0.21)	1.16 (0.13)	6.48 (1.96)	4.54 (1.56)
+ Mod-K stability index (grid binary 3)	2.22 (0.06)	2.53 (0.19)	7.05 (2.50)	4.52 (1.40)	1.01 (0.21)	1.13 (0.12)	7.99 (2.75)	6.31 (2.48)
2-m Air temperature	36.08 (29.61)	41.32 (35.16)	3.34 (0.15)	2.77 (0.78)	14.44 (12.71)	20.62 (18.63)	3.73 (0.84)	3.37 (0.87)
Downwelling longwave radiation flux	4.42 (0.39)	5.42 (2.16)	6.19 (1.96)	5.58 (2.54)	1.67 (0.61)	12.38 (9.76)	5.12 (2.27)	4.66 (1.82)
Downwelling shortwave radiation flux	3.09 (0.01)	1.20 (0.01)	3.21 (0.63)	2.92 (0.83)	0.99 (0.23)	0.98 (0.18)	7.91 (4.45)	3.20 (1.36)
+ Shortwave radiation flux (grid binary 1)	2.41 (0.00)	1.27 (0.00)	6.57 (1.85)	3.07 (1.00)	0.92 (0.24)	0.81 (0.15)	8.63 (4.93)	3.76 (1.55)
+ Shortwave radiation flux (grid binary 2)	2.32 (0.01)	1.09 (0.00)	2.73 (0.62)	2.28 (0.57)	0.87 (0.20)	0.93 (0.12)	7.48 (4.60)	2.79 (0.81)
+ Shortwave radiation flux (grid binary 3)	1.60 (0.01)	1.27 (0.00)	1.65 (0.06)	2.62 (0.51)	0.75 (0.19)	1.25 (0.26)	2.94 (0.73)	3.04 (0.87)

TABLE 1. (*Continued*)

	Jan				Jul			
	TMAX	TMIN	POCC	PRCP	TMAX	TMIN	POCC	PRCP
Precipitation rate	2.23 (0.01)	2.21 (0.06)	11.85 (4.65)	26.55 (15.66)	1.35 (0.36)	1.23 (0.13)	12.65 (7.11)	9.21 (5.12)
+ Precipitation rate (grid binary 1)	4.79 (0.02)	2.76 (0.05)	24.24 (17.59)	5.72 (2.58)	1.36 (0.41)	1.26 (0.19)	14.53 (9.79)	4.73 (1.79)
+ Precipitation rate (grid binary 2)	4.62 (0.06)	3.13 (0.06)	23.51 (17.01)	7.46 (4.21)	1.37 (0.37)	1.28 (0.14)	11.30 (6.62)	5.35 (2.35)
+ Precipitation rate (grid binary 3)	3.22 (0.01)	2.24 (0.01)	21.66 (13.60)	13.55 (8.40)	1.27 (0.32)	1.01 (0.09)	9.26 (3.65)	7.41 (3.46)
Specific humidity 300 hPa	2.07 (0.01)	1.93 (0.01)	2.31 (0.16)	3.55 (0.78)	0.77 (0.10)	1.02 (0.10)	3.50 (0.61)	5.05 (1.70)
Specific humidity 500 hPa	2.13 (0.01)	2.06 (0.02)	2.68 (0.23)	4.44 (1.06)	1.08 (0.15)	1.45 (0.19)	6.24 (1.59)	5.76 (2.19)
Specific humidity 700 hPa	2.87 (0.03)	2.04 (0.07)	4.22 (0.83)	5.07 (2.11)	1.39 (0.25)	1.62 (0.44)	5.18 (1.89)	5.47 (1.83)
Specific humidity 850 hPa	4.14 (0.62)	4.93 (1.90)	2.75 (0.25)	2.93 (0.74)	1.48 (0.39)	1.81 (0.63)	2.96 (0.49)	4.70 (1.90)
Specific humidity 1000 hPa	5.84 (2.39)	14.72 (9.38)	2.59 (0.08)	2.98 (0.71)	1.44 (0.58)	9.68 (7.27)	3.43 (0.86)	4.39 (1.53)
Specific humidity × zonal wind 300 hPa	2.74 (0.01)	2.21 (0.01)	3.33 (0.65)	4.06 (1.08)	1.55 (0.58)	1.34 (0.27)	5.93 (2.03)	5.15 (1.86)
Specific humidity × zonal wind 500 hPa	2.79 (0.02)	2.78 (0.02)	3.23 (0.50)	4.31 (1.00)	1.74 (0.45)	1.68 (0.21)	6.73 (1.27)	4.92 (1.60)
Specific humidity × zonal wind 700 hPa	2.62 (0.03)	3.14 (0.20)	3.80 (0.57)	4.32 (1.03)	1.40 (0.32)	1.32 (0.09)	7.04 (1.45)	4.31 (1.32)
Specific humidity × zonal wind 850 hPa	2.90 (0.06)	3.03 (0.27)	2.95 (0.29)	4.15 (0.93)	1.02 (0.16)	1.06 (0.08)	3.58 (0.68)	4.12 (1.14)
Specific humidity × zonal wind 1000 hPa	5.32 (0.12)	2.67 (0.08)	4.02 (0.57)	4.29 (1.01)	1.49 (0.54)	1.30 (0.18)	3.33 (0.61)	4.70 (1.44)
Specific humidity × meridional wind 300 hPa	1.94 (0.01)	1.69 (0.02)	4.28 (1.18)	5.28 (2.17)	0.74 (0.13)	1.44 (0.14)	3.77 (0.59)	5.17 (1.69)
Specific humidity × meridional wind 500 hPa	1.94 (0.03)	1.94 (0.02)	5.86 (2.08)	5.81 (2.45)	0.91 (0.15)	1.10 (0.05)	3.89 (0.47)	4.37 (1.31)
Specific humidity × meridional wind 700 hPa	2.09 (0.02)	1.99 (0.03)	5.67 (1.86)	5.77 (2.27)	0.96 (0.16)	1.03 (0.09)	3.11 (0.34)	4.49 (1.60)
Specific humidity × meridional wind 850 hPa	2.33 (0.03)	2.72 (0.01)	2.94 (0.34)	4.52 (1.14)	0.96 (0.16)	1.12 (0.12)	2.61 (0.29)	4.00 (1.34)
Specific humidity × meridional wind 1000 hPa	3.56 (0.01)	3.55 (0.01)	2.76 (0.14)	3.98 (0.98)	1.23 (0.18)	1.19 (0.13)	3.15 (0.33)	4.81 (1.43)
Atmospheric thickness (500–1000 hPa)	7.97 (5.06)	5.28 (2.68)	2.54 (0.20)	2.88 (0.57)	12.21 (10.82)	15.61 (13.77)	2.32 (0.21)	2.55 (0.68)
Atmospheric thickness (700–1000 hPa)	16.77 (12.70)	11.26 (8.04)	2.58 (0.16)	2.75 (0.64)	13.06 (11.68)	12.79 (10.88)	2.72 (0.24)	2.95 (0.65)
Modified K stability index (squared)	1.34 (0.02)	1.62 (0.03)	2.08 (0.05)	4.86 (1.13)	1.27 (0.38)	1.39 (0.15)	5.45 (1.80)	5.13 (1.77)
Log (precipitation rate × 86 400 + .01)	4.30 (0.12)	2.36 (0.06)	10.09 (5.81)	6.34 (3.38)	1.06 (0.27)	1.33 (0.18)	6.38 (2.70)	5.32 (2.13)
500 hPa height (high-pass filtered)	0.40 (0.00)	0.44 (0.00)	0.59 (0.05)	0.76 (0.19)	0.23 (0.06)	0.31 (0.03)	0.74 (0.08)	1.43 (0.38)
1000 hPa height (high-pass filtered)	0.47 (0.00)	0.79 (0.00)	0.50 (0.02)	0.66 (0.12)	0.16 (0.03)	0.25 (0.03)	0.59 (0.06)	1.07 (0.33)

bination of variables, the equation is trained on the first period (true-dependent sample) and validated on the second period (pseudoindependent sample). This process is repeated five times for the same combination of variables. The selection of the set of predictor variables is based on the average explained variance from the five pseudoindependent samples. Once the variable set is selected, coefficients in the regression equation are estimated using the entire dependent sample. Typically, between three and eight variables are used in the MOS equations.

A final step in the downscaling procedure is stochastic modeling of the residuals in the multiple linear regression equations to provide an assessment of model uncertainty and permit the generation of probabilistic forecasts. For maximum and minimum temperature, this is achieved by extracting a random number from a normal Gaussian distribution (mean of zero and standard deviation of one), multiplying the random number by the standard deviation of the regression residuals, and adding this product to the forecast of temperature. For precipitation, we first determine precipitation occurrence. A random number is drawn from a uniform distribution ranging from zero to one. If the random number is lower than the forecasted probability of precipitation occurrence, the day is classified as a precipitation day. Precipitation amounts are only forecasted for precipitation days. After forecasting precipitation amounts, residuals are modeled stochastically using methods identical to those used for maximum and minimum temperature, and then the forecasted (normally distributed) precipitation amounts are transformed back to the original gamma-type distribution of observed precipitation using the nonparametric probability transform techniques described earlier. The stochastic modeling of the regression residuals inflates the variance of precipitation and temperature forecasts, reducing problems of variance underestimation that are typical of regression-based models.

c. Forecast improvements

Output from the MOS system does not contain the large biases evident in the raw NCEP predictions. Figure 3 illustrates the observed and modeled long-term (1977–98) mean for maximum and minimum temperature, precipitation occurrence, and precipitation amounts for the four midseason months of January, April, July, and October. Each point represents the observed and modeled 1977–98 mean for an individual station. The median absolute bias, computed over all stations for individual months, is summarized in Table 2. For all variables, the MOS system reproduces spatial variations in the long-term climatologies (Fig. 3). The median absolute bias in maximum and minimum temperature at individual stations is less than 0.5°C in all months apart from December, and the median absolute bias for precipitation at individual stations is less than 15% of the mean in all months except for July and August (Table 2). The lower scatter between observed and modeled long-term mean temperatures (Fig. 3) occurs because the temporal variations in temperature at a given station are much lower than the spatial variations in long-term mean temperature across the contiguous United States.

The ability of the MOS-based system to reproduce daily variability in precipitation and maximum temperature is presented in Fig. 4 for the months of January and July. The presentation is identical to Fig. 2, where each station in the contiguous United States is represented by a colored dot. Spatial variations in the accuracy of the MOS-based precipitation and maximum temperature predictions are similar to the raw NCEP predictions. Comparing Fig. 4 with Fig. 2, note again the modest skill for January precipitation predictions in California and the upper Midwest, the high skill for January temperature predictions over the eastern United States, and the low skill for July precipitation predictions throughout the country. There are, however, several regions where the MOS-based forecasts are more accurate than the raw NCEP forecasts. Higher forecast accuracy is readily apparent for January precipitation in the northeastern United States and for maximum temperature in January and July over the entire country. The skill of the maximum temperature forecasts over the Rocky Mountains and Appalachians in January, and over the east coast of the United States in July (Fig. 2), improves when applying statistical MOS guidance (Fig. 4).

To bring these results together, Fig. 5 compares the median skill of the raw NCEP and MOS-based precipitation and temperature forecasts for the four midseason months of January, April, July, and October. The median forecast skill is computed using all stations in the cooperative network that had sufficient data to develop MOS equations and compute skill scores (i.e., the median is computed for all stations in Figs. 2 and 4). As in Figs. 2 and 4, the skill of the maximum and minimum temperature forecasts is measured by the explained variance (i.e., the r^2 value), and the skill of precipitation forecasts is measured by Spearman rank correlations. Kupier's skill score (Wilks 1995) is used to evaluate the accuracy of precipitation occurrence predictions, that is, how well forecasted wet (dry) days match observed wet (dry) days (see appendix). Raw NCEP forecasts are shown in Fig. 5 as squares, and MOS-based forecasts are shown as triangles.

The MOS-based maximum and minimum temperature forecasts are, in almost all cases, more accurate than the raw NCEP forecasts (top two rows in Fig. 5). This is most apparent at the beginning of the forecast cycle in January (e.g., day+0), where the MOS-based predictions explain approximately 20% more variance than the raw NCEP predictions. For precipitation occurrence, the MOS guidance has substantially greater accuracy than the raw NCEP predictions. This mostly reflects the frequent ''drizzle'' in global-scale models

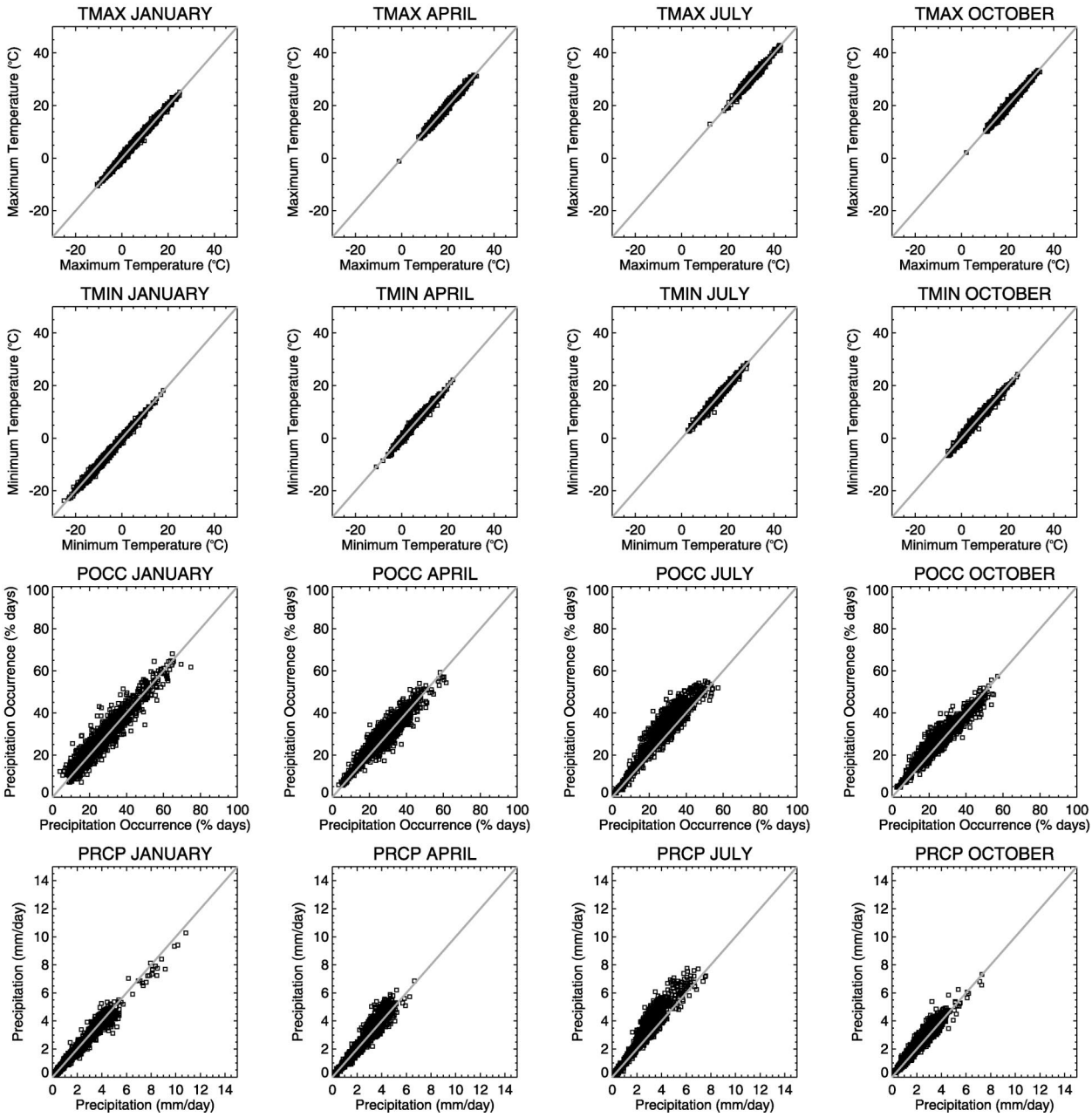

FIG. 3. Scatterplots of the observed and modeled MOS-based long-term (1977–98) mean for the four midseason months of Jan (column 1), Apr (column 2), Jul (column 3), and Oct (column 4) for max and min temperature (°C; top two rows), precipitation occurrence (expressed as a percentage of precipitation days; third row), and precipitation amounts (mm day^{-1}; bottom row). Each point illustrates the observed and modeled mean for an individual station.

that is accentuated by interpolating NCEP data from all grid points within a 500-km search radius to point station locations. MOS-based predictions of precipitation amounts (bottom row in Fig. 5) are less accurate than the raw NCEP predictions. While this result is initially surprising, it is most likely due to an inadequate number of precipitation days used to develop the MOS equations in dry regions. Recall that the regression equations for precipitation amounts are developed using the subset of days when there is precipitation at the station. The MOS-based day+0 precipitation predictions in January (Fig. 3) are of much higher skill in California and the northeastern United States than the corresponding raw NCEP predictions (Fig. 5). Work is continuing to determine the minimum sample size necessary to develop stable MOS equations, and the effects of different methods to artificially increase the sample size (e.g., ''borrowing'' data from adjacent months and nearby sta-

TABLE 2. Median absolute bias, computed over all stations in the contiguous United States, for the MOS-based predictions of max temperature (TMAX), min temperature (TMIN), precipitation occurrence (POCC), and precipitation amounts (PRCP). Median absolute bias for max and min temperature is expressed in terms of °C. Median absolute bias for precipitation occurrence is expressed as a percentage of days with precipitation, and the median absolute bias for precipitation amounts is expressed as a percentage of the long-term precipitation mean.

	TMAX	TMIN	POCC	PRCP
Jan	0.4020	0.4950	2.3000	9.3218
Feb	0.4610	0.4530	2.2000	10.2492
Mar	0.4640	0.4160	2.1000	8.8520
Apr	0.3800	0.3850	2.1000	8.8729
May	0.4070	0.3740	2.5000	11.5224
Jun	0.4140	0.3430	3.6000	13.3811
Jul	0.4210	0.3360	4.4000	16.7464
Aug	0.3850	0.3620	4.2000	15.9492
Sep	0.3680	0.3990	3.3000	14.7059
Oct	0.3520	0.4290	2.2000	8.8704
Nov	0.3480	0.3830	2.1000	7.7243
Dec	0.4180	0.5650	2.1000	9.2237

tions). The results in Fig. 5 show that improvements obtained from MOS are most pronounced for short forecast lead times. After about 4–5 days, correlations between MOS predictions and station data are of similar magnitude to correlations between raw NCEP output and station data. For these longer lead times, the main benefit of the MOS approach is to correct the systematic biases in the NCEP MRF output.

6. Use of MRF output to produce forecasts of streamflow

Based on the results presented thus far, we assess if the use of downscaled MRF output (MOS) be used in a hydrologic model to improve upon the traditional practice of forecasting streamflow using the climatic ESP procedure (Day 1985). Potential increases in forecast skill are assessed when the ESP ensemble inputs are replaced with MOS-based ensemble forecasts of precipitation and temperature. The streamflow forecast experiments are constructed as follows: Basin initial conditions for both the MOS-based and ESP forecasts are estimated by running our hydrologic model with station observations up to the start of each forecast period, and then with the forecast ensemble. The forecast ensemble for ESP comprises station observations from matching dates in the historical record [see Day (1985) for more details], and the forecast ensemble from MOS is generated though the regression-based estimates and stochastic modeling of the residuals in the regression equation (see section 5b for more details). In the ESP approach there is essentially zero skill in the inputs—the skill in the forecasts of streamflow is due to specification of the basin conditions at the start of the forecast and the influence of those conditions on the basin hydrologic response.

Hydrologic model simulations for these forecast experiments are performed using the U.S. Geological Survey's Precipitation-Runoff Modeling System (PRMS), which is described in detail in Leavesley et al. (1983), Leavesley et al. (1996), and Hay et al. (2002). The hydrologic simulations generated using station observations (precipitation and maximum and minimum temperature) were used as a measure of ''truth'' to assess the skill of the hydrologic forecasts. This focuses attention on the hydrologic impact of errors in the inputs, instead of possible errors in the hydrologic model itself. The ESP and MOS-based forecasts use the same initial conditions, which are estimated for each forecast by running PRMS with station observations. The ranked probability skill score (RPSS) is used to assess the skill of the ESP and MOS-based streamflow forecasts (appendix).

Results are examined in the following four basins: 1) Animas River at Durango, Colorado (Animas); 2) East Fork of the Carson River near Gardnerville, Nevada (Carson); 3) Cle Elum River near Roslyn, Washington (Cle Elum); and 4) Alapaha River at Statenville, Georgia (Alapaha). The surface hydrology of the first three basins (Animas, Carson, and Cle Elum) is dominated by snowmelt. The Carson and Cle Elum basins are also characterized by frequent rain-on-snow events in the winter months. The Alapaha basin is a low-elevation rainfall-dominated basin. Table 3 lists some of the defining features of each basin, and Fig. 6 shows the location of each.

The probabilistic skill of the 8-day streamflow forecasts produced using statistically downscaled MRF output (MOS) and the climatic ESP technique are presented in Fig. 7. The contour plots show the month along the *x* axis, the forecast day along the *y* axis, and the RPSS as the contoured variable. Increases in forecast skill from MOS-based forecasts are most pronounced during the peak snowmelt season in the three western basins (the Animas, the East Fork of the Carson, and the Cle Elum). At this time of the year, daily variations in streamflow are more closely tied to variations in temperature than precipitation, and the high skill in predictions of temperature translates into high skill in predictions of streamflow. The MOS-based forecasts and ESP perform equally well in the rainfall-dominated basin (Alapaha), where skillful predictions of streamflow are hampered by the poor predictions of precipitation. The conclusion gleaned from these results is that skillful short-term predictions of runoff are possible in snowmelt situations, when knowledge of the accumulated snowpack is available. Further work on a larger set of basins is required to verify this statement.

7. Summary and discussion

This paper examined an archive containing more than 40 years of 8-day atmospheric forecasts from the NCEP reanalysis project to assess the possibilities for using medium-range NWP model output for predictions of

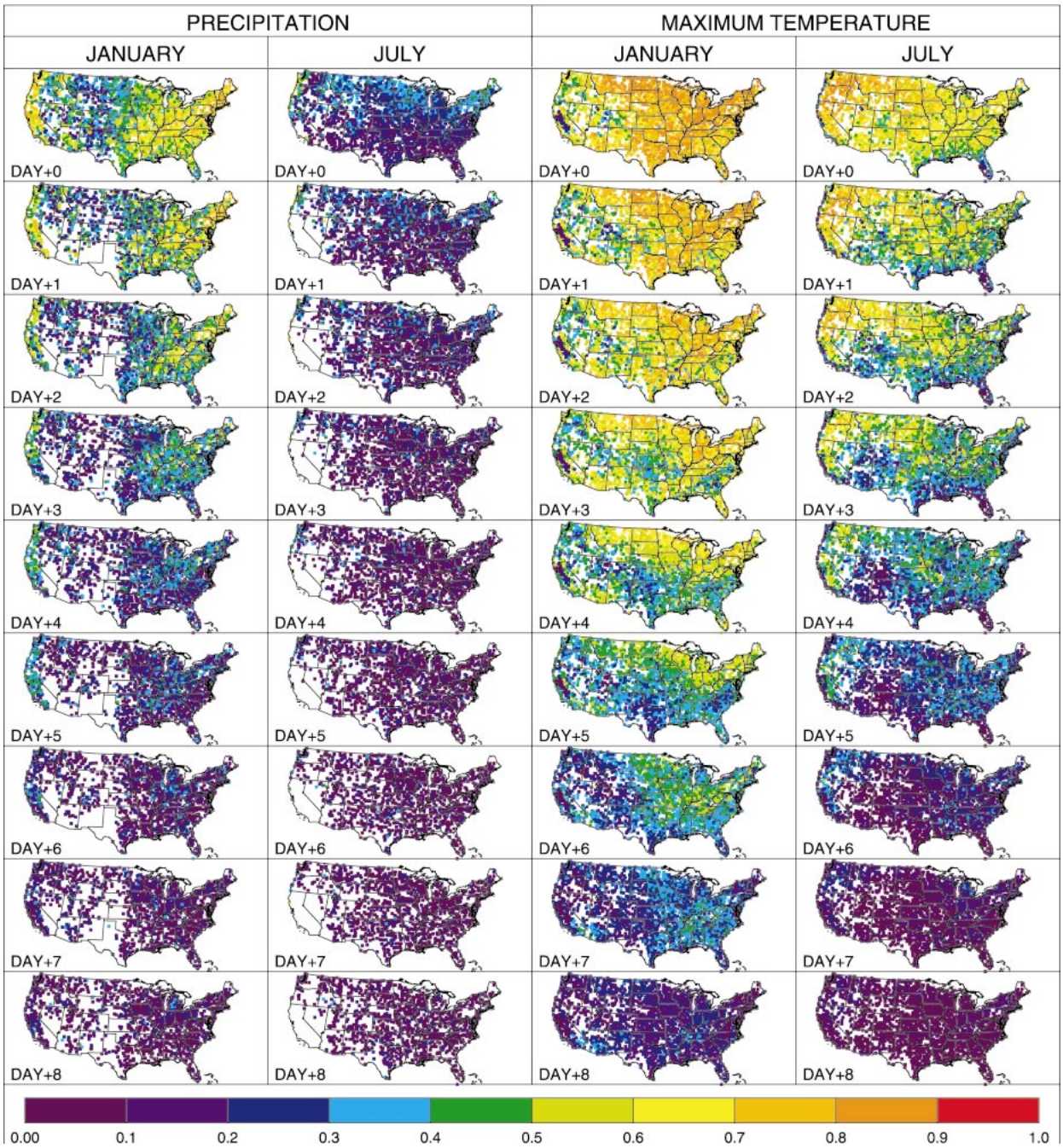

FIG. 4. Accuracy of the MOS-based precipitation and 2-m max temperature forecasts, showing forecast skill for day+0 (top row) and then skill for each subsequent forecast lead time. Forecast skill for precipitation forecasts is assessed using Spearman rank correlations, and forecast skill for 2-m max temperature forecasts is assessed using squared Pearson correlations (r^2).

streamflow. Systematic biases in the NCEP forecasts are often large. In many regions, precipitation biases are in excess of 100% of the mean, and temperature biases are in excess of 3°C. In some locations, biases are even higher. In addition, the accuracy of the NCEP forecasts is rather low in many areas of the country. Most apparent are the generally low skill in precipitation forecasts (particularly in July) and the low skill in temperature forecasts over the Rocky and Appalachian Mountains in January and over the eastern seaboard in July. These results outline a clear need for additional processing of the NCEP Medium-Range Forecast Model output before it is used for hydrologic predictions.

Techniques of model output statistics (MOS) were used to improve the raw NCEP forecast model output. In our MOS technique, atmospheric variables included

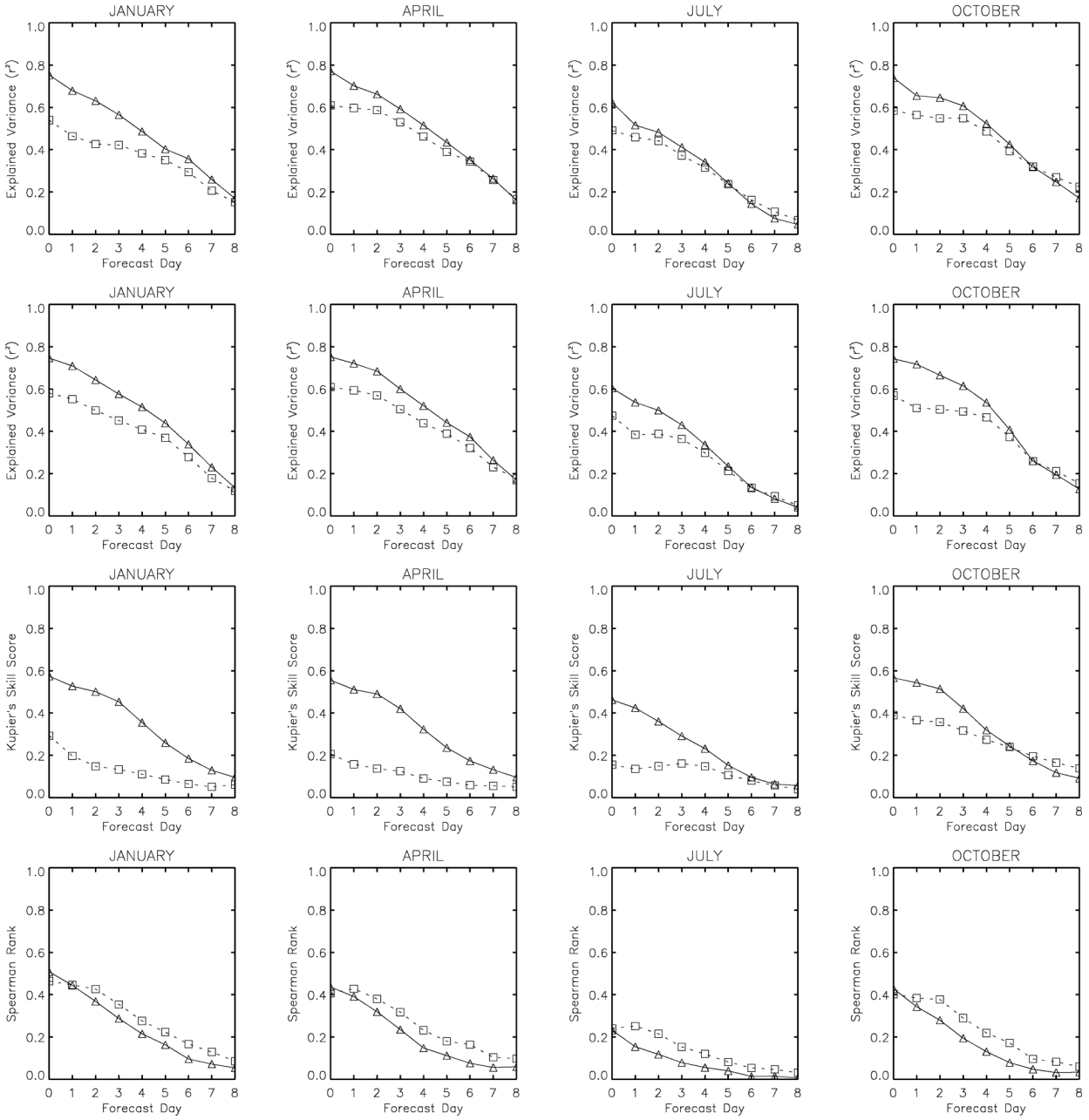

FIG. 5. Accuracy of the raw NCEP and the MOS-based precipitation and temperature forecasts for the four midseason months of Jan (column 1), Apr (column 2), Jul (column 3), and Oct (column 4). Shown are skill scores for max and min temperature (squared Pearson correlation; top two rows), precipitation occurrence (Kupier's skill score; third row), and precipitation amounts (Spearman rank correlation; bottom row). Raw NCEP predictions are expressed with a dotted line (squares), and the MOS-based predictions are expressed with a solid line (triangles).

in the NCEP forecast archive (e.g., total column precipitable water, 2-m air temperature) were used as predictors in a forward screening multiple linear regression approach to forecast precipitation occurrence, precipitation amounts, maximum temperature, and minimum temperature for over 11 000 stations in the National Weather Service cooperative network. This procedure effectively removes all systematic biases in the raw NCEP precipitation and temperature forecasts. In addition, the MOS guidance results in substantial improvements in the accuracy of maximum and minimum temperature forecasts throughout the country, most apparent in Rocky and Appalachian Mountains where the skill of the raw 2-m temperature forecasts was low. MOS guidance also substantially improves predictions of precipitation occurrence. Forecast improvements of pre-

TABLE 3. Study basins.

	Study basin:			
	Animas River at Durango	East Fork of the Carson River near Gardnerville	Cle Elum River near Roslyn	Alapaha River at Statenville
State	Colorado	California/Nevada	Washington	Georgia
Gauging station identification	09361500	10309000	12479000	02317500
Drainage area (km^2)	1792	922	526	3626
Elevation range (m)	2000–3700	1600–3000	680–1800	40–125
Number of hydrologic response units*	121	96	124	180
Nash–Sutcliffe goodness-of-fit statistic between measured and simulated runoff using station observations	0.85	0.83	0.80	0.75
Number of stations	37	16	14	28

* Hydrologic Response Units are the subcatchment areas modeled in PRMS.

cipitation amounts are more modest than for temperature and precipitation occurrence. The MOS guidance did result in increased forecast accuracy over the northeastern United States in January, but overall the accuracy of MOS-based forecasts of precipitation amounts is slightly lower than the raw NCEP forecasts. This may be due to an inadequate number of precipitation days to develop stable MOS equations in dry regions. Nevertheless, the raw NCEP precipitation forecasts contain biases and need some statistical correction before they can be used directly in hydrologic forecasting applications.

Statistically downscaled MRF output (NCEP atmospheric forecasts) are found to provide realistic predictions of streamflow in the snowmelt-dominated basins examined in the western United States. Short-term variations in streamflow in snowmelt-dominated river systems are influenced more by variations in temperature than variations in precipitation. If the volume of snowpack is estimated correctly for the winter season, reliable short-term forecasts of streamflow are possible through a good representation of the effects of temperature on the rates of snowmelt (see also Hay et al. 2002). The accuracy of the MOS-based temperature forecasts are in fact much higher than the MOS-based precipitation forecasts (Fig. 4), and the predictions of streamflow in

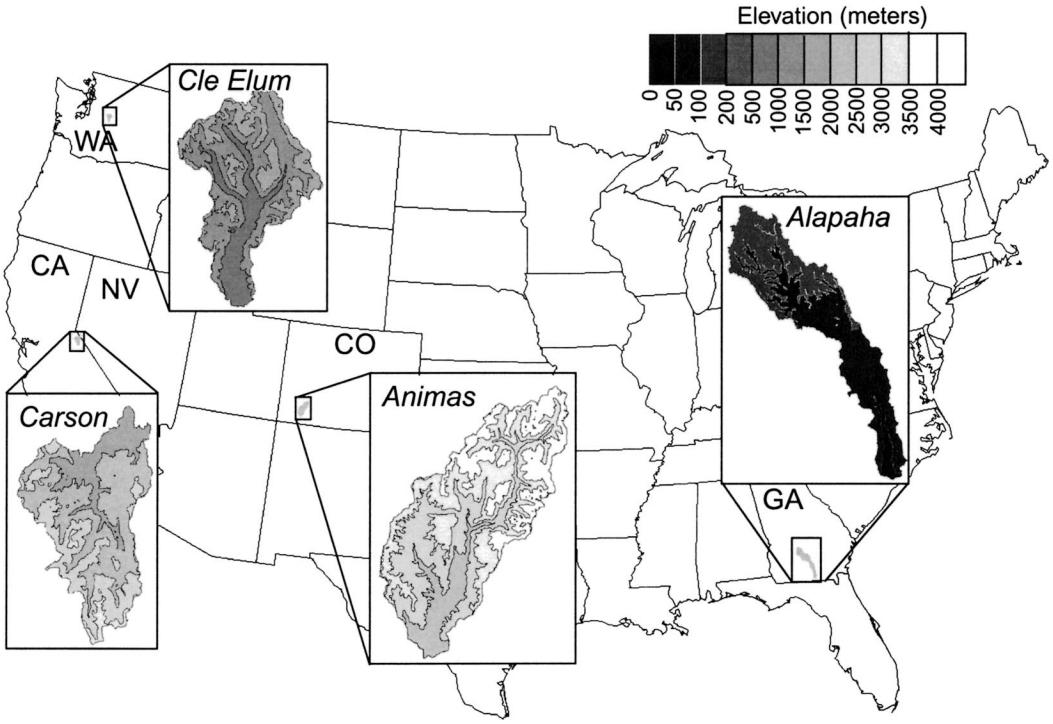

FIG. 6. Location of study basins.

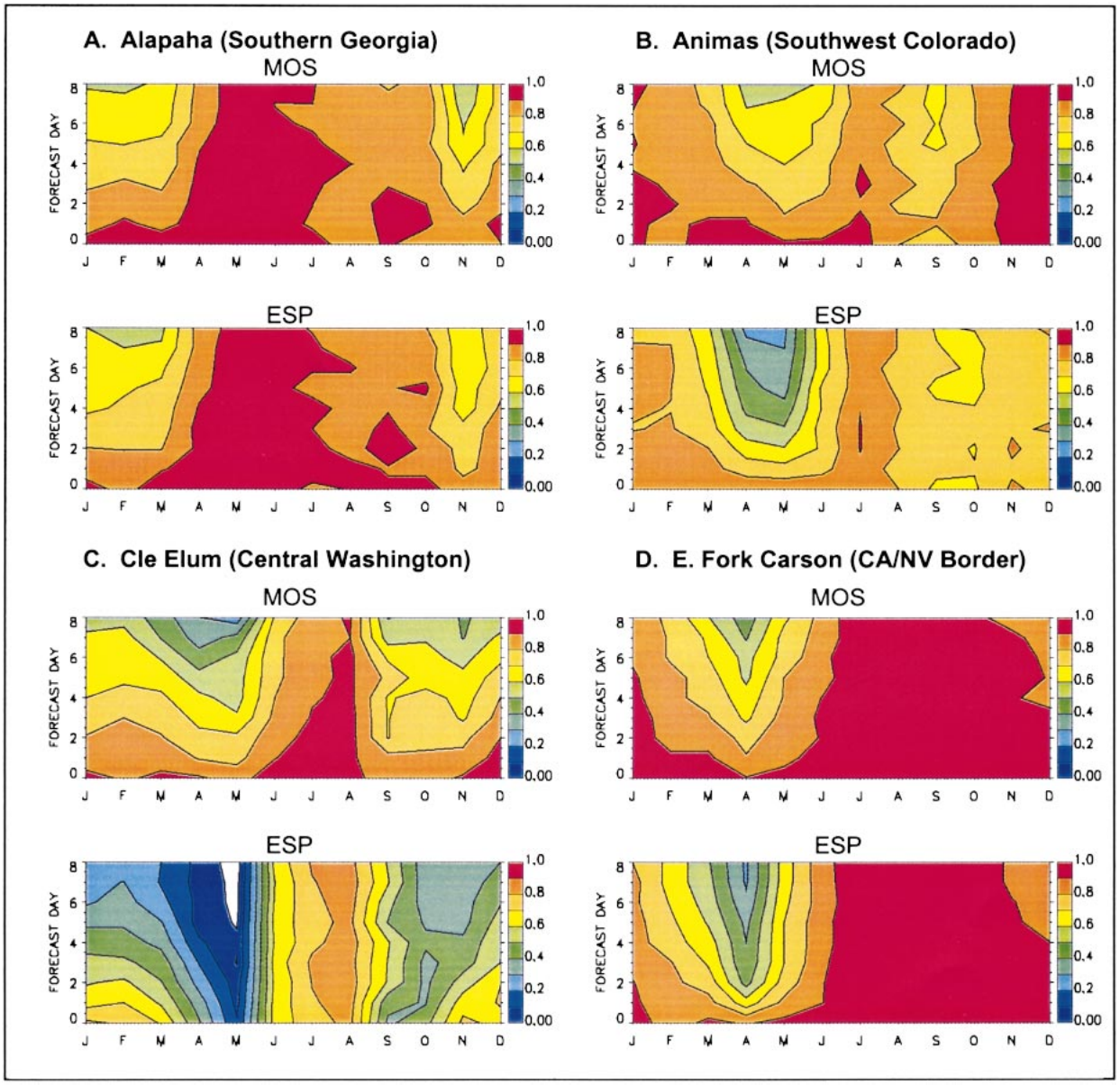

FIG. 7. RPSS calculated for each forecast day and month, using (top) MOS-based precipitation and temperature forecasts and (bottom) the climatological ESP approach. See text for further details.

snowmelt-dominated river basins (the Animas, East Fork of the Carson, and the Cle Elum) do exhibit greater improvement over ESP than in the rainfall-dominated basin (the Alapaha). The poor forecasts of precipitation limit the use of atmospheric forecast output in river basins where the surface hydrology is dominated by rainfall.

Further improvement in the skill of streamflow forecasts depends not only on the local-scale forecasts of precipitation and temperature, but also on the specification of basin initial conditions and on hydrologic model simulations of streamflow. All three of these issues need more attention. On the atmospheric side, forecast skill at local scales is limited by the coarse horizontal resolution of the MRF (e.g., precipitation occurs on the subgrid scale) and deficiencies in model physics (e.g., summertime precipitation may be poorly represented because of inadequacies in convective parameterizations). It is likely that nesting a series of regional atmospheric models to finer scales may be necessary to adequately resolve the subgrid-scale variations and physical atmospheric processes important for hydrologic modeling. Of course, at longer lead times forecast skill depends on the prediction of large-scale climate features (e.g., the 500-hPa height field). In terms of estimating basin initial conditions, opportunities for forecast im-

provements are largest in snowmelt-dominated basins where it is possible to estimate the spatial variability of snowpack from satellites. New satellite missions to estimate soil moisture and subsurface storage offer some promise, but satellite estimates of soil moisture are currently only reliable for the top soil layers on nonvegetated surfaces, and estimates of subsurface storage are currently only reliable on large spatial scales. Further work is needed on this topic. Improved model simulations of streamflow are possible through both advances in parameter estimation methodologies and improvements in model structure. Recent work has shown the value of multimodel approaches to improve hydrologic predictions, and these approaches can be implemented operationally through close coordination between different modeling groups. Focused attention on improving streamflow forecasts will help water managers optimize the use of water resources and thus satisfy the increasingly competitive demands for water.

Acknowledgments. This work was supported by both the NOAA GAPP Program (Award NA16GP2806) and the NOAA RISA Program (Award NA17RJ1229). The authors are grateful to Mark C. Serreze and Robert L. Wilby for comments on an earlier draft of this manuscript.

APPENDIX

Measures of Forecast Skill

a. The Pearson and Spearman correlation coefficients

The equation for the Pearson correlation coefficient is

$$r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}, \qquad \text{(A1)}$$

where x is station data (in this case maximum and minimum temperature observations), $\bar{x}$ is the mean of the station observations, y is the predicted value (in this case raw NCEP output or the MOS-based predictions), and $\bar{y}$ is the mean of the predicted value. The equation for the Spearman ''rank'' correlation coefficient is identical to Eq. (A1), except the rank of the values is used in place of the actual values. For both the Pearson and Spearman correlation coefficients, a value of 1.0 represents a perfect forecast.

b. Kupier's skill score

Kupier's skill score is used to assess the skill of binary predictions (in this case precipitation occurrence). It is calculated from a 2×2 contingency table as

$$\text{KSS} = \frac{ad - bc}{(a + c)(b + d)}, \qquad \text{(A2)}$$

where a represents the number of cases when a wet day was forecast and a wet day was observed, b represents the number of cases when a wet day was forecast and a dry day was observed, c represents the number of cases when a dry day was forecast and a wet day was observed, and d represents the number of cases when a dry day was forecast and a dry day was observed (see also Wilks 1995). Similar to the Pearson and Spearman correlation coefficients, a value of 1.0 represents a perfect forecast for Kupier's skill score.

c. The ranked probability skill score

The RPSS is used to provide a measure of the probabilistic skill of the ensemble streamflow forecasts. The RPSS is based on the ranked probability score (RPS) computed for each forecast-observation pair:

$$\text{RPS} = \sum_{m=1}^{J} (Y_m - O_m)^2, \qquad \text{(A3)}$$

where Y_m is the cumulative probability of the forecast for category m, and O_m is the cumulative probability of the observation for category m. This is implemented as follows (see also Wilks 1995): First, the observed time series is used to distinguish 10 (J) possible categories for forecasts of precipitation and temperature (i.e., the minimum value to the 10th percentile, the 10th percentile to the 20th percentile . . . the 90th percentile to the maximum value). These categories are determined separately for each month and basin. Next, for each forecast-observation pair, the number of ensemble members forecast in each category is determined and their cumulative probabilities are computed. Similarly, the appropriate category for the observation is identified and the observation's cumulative probabilities are computed (i.e., all categories below the observation's position are assigned ''0,'' and all categories equal to and above the observation's position are assigned ''1''). Now, the RPS is computed as the squared difference between the observed and forecast cumulative probabilities, and the squared differences are summed over all categories [Eq. (A3)]. The RPSS is then computed as

$$\text{RPSS} = 1 - \frac{\overline{\text{RPS}}}{\overline{\text{RPS}}_{\text{rand}}}, \qquad \text{(A4)}$$

where $\overline{\text{RPS}}$ is the mean ranked probability score for all forecast-observation pairs, and $\overline{\text{RPS}}_{\text{rand}}$ is the mean ranked probability score for randomly shuffled forecast-observation pairs.

REFERENCES

Allen, D. M., 1971: Mean square error of prediction as a criterion for selecting variables. *Technometrics,* **13,** 469–475.

Antolik, M. S., 2000: An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts. *J. Hydrol.,* **239,** 306–337.

Briggs, P. R., and J. G. Cogley, 1996: Topographic bias in mesoscale precipitation networks. *J. Climate,* **9,** 205–218.

Charba, J. P., 1977: Operational system for predicting thunderstorms two to six hours in advance. NOAA Tech. Memo. NWS TDL-64, 24 pp.

Chelliah, M., and C. F. Ropelewski, 2000: Reanalyses-based tropospheric temperature estimates: Uncertainties in the context of global climate change detection. *J. Climate,* **13,** 3187–3205.

Connelly, B. A., D. T. Braatz, J. B. Halquist, M. M. DeWeese, L. Larson, and J. J. Ingram, 1999: Advanced hydrologic prediction system. *J. Geophys. Res.,* **104** (D16), 19 655–19 660.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.,* **33,** 140–158.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *ASCE J. Water Res. Plann. Manage.,* **111,** 157–170.

Dey, C. H., and L. L. Morone, 1985: Evolution of the NMC Global Assimilation System: January 1982–December 1983. *Mon. Wea. Rev.,* **113,** 304–318.

DiMego, G., 1988: The National Meteorological Center Regional Analysis System. *Mon. Wea. Rev.,* **116,** 1137–1156.

Eischeid, J. K., P. A. Pasteris, H. F. Diaz, M. S. Plantico, and N. J. Lott, 2000: Creating a serially complete, national daily time series of temperature and precipitation for the western United States. *J. Appl. Meteor.,* **39,** 1580–1591.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Hagemann, S., and L. D. Gates, 2001: Validation of the hydrological cycle of ECMWF and NCEP reanalyses using the MPI hydrological discharge model. *J. Geophys. Res.,* **106** (D2), 1503–1510.

Hamlet, A. F., and D. P. Lettenmaier, 1999: Columbia River streamflow forecasting based on ENSO and PDO climate signals. *ASCE J. Water Res. Plann. Manage.,* **125,** 333–341.

Hay, L. E., M. P. Clark, R. L. Wilby, W. J. Gutowski, R. W. Arritt, E. S. Takle, Z. Pan, and G. H. Leavesley, 2002: Use of regional climate model output for hydrologic simulations. *J. Hydrometeor.,* **3,** 571–590.

Janowiak, J. E., A. Gruber, C. R. Kondragunta, R. E. Livezey, and G. J. Huffman, 1998: A comparison of the NCEP–NCAR reanalysis precipitation and the GPCP rain gauge–satellite combined dataset with observational error considerations. *J. Climate,* **11,** 2960–2979.

Jensenius, J. S., Jr., 1992: The use of grid-binary variables as predictors for statistical weather forecasting. Preprints, *12th Conf. on Probability and Statistics in the Atmospheric Sciences,* Toronto, ON, Canada, Amer. Meteor. Soc., 225–230.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.,* **77,** 437–471.

——, S. J. Lord, and R. D. McPherson, 1998: Maturity of operational numerical weather prediction: Medium range. *Bull. Amer. Meteor. Soc.,* **79,** 2753–2769.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Leavesley, G. H., R. W. Lichty, B. M. Troutman, and L. G. Saindon, 1983: Precipitation-runoff modeling system: User's manual. U.S. Geological Survey Water Investment Rep. 83-4238, 207 pp.

——, P. J. Restrepo, S. L. Markstrom, M. Dixon, and L. G. Stannard, 1996: The modular modeling system—MMS: User's manual. U.S. Geological Survey Open File Rep. 96-151, 142 pp.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, 1997: A Pacific Interdecadal Climate Oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.,* **78,** 1069–1079.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast models. *J. Climate Appl. Meteor.,* **26,** 1589–1600.

Pan, H.-L., and W.-S. Wu, 1994: Implementing a mass flux convection parameterization package for the NMC medium range forecast model. Preprints, *10th Conf. on Numerical Weather Prediction,* Portland, OR, Amer. Meteor. Soc., 96–98.

Panofsky, H. A., and G. W. Brier, 1963: *Some Applications of Statistics to Meteorology.* Mineral Industries Continuing Education, College of Mineral Industries, The Pennsylvania State University, 224 pp.

Peppler, R. A., and P. J. Lamb, 1989: Tropospheric static stability and central North American growing season rainfall. *Mon. Wea. Rev.,* **117,** 1156–1180.

Reek, T., S. R. Doty, and T. W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bull. Amer. Meteor. Soc.,* **73,** 753–762.

Reid, P. A., P. D. Jones, O. Brown, C. M. Goodess, and T. D. Davies, 2001: Assessments of the reliability of NCEP circulation data and relationships with surface climate by direct comparisons with station based data. *Climate Res.,* **17,** 247–261.

Serreze, M. C., and C. M. Hurst, 2000: Representation of mean Arctic precipitation from NCEP–NCAR and ERA reanalyses. *J. Climate,* **13,** 182–201.

Trenberth, K. E., and C. J. Guillemot, 1998: Evaluation of the atmospheric moisture and hydrologic cycle in the NCEP/NCAR reanalyses. *Climate Dyn.,* **14,** 213–231.

Vislocky, R. L., and J. M. Fritsch, 1997: Performance of an advanced MOS system in the 1996–97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc.,* **78,** 2851–2857.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Woollen, J. S., E. Kalnay, L. Gandin, W. Collins, S. Saha, R. Kistler, M. Kanamitsu, and M. Chelliah, 1994: Quality control in the reanalysis system. *Bull. Amer. Meteor. Soc.,* **75,** 13–14.