Short-Range Precipitation Forecasts from Time-Lagged Multimodel Ensembles during the HMT-West-2006 Campaign

HUILING YUAN, JOHN A. McGINLEY, PAUL J. SCHULTZ, CHRISTOPHER J. ANDERSON,* AND CHUNGU LU*

NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 19 January 2007, in final form 30 September 2007)

ABSTRACT

High-resolution (3 km) time-lagged (initialized every 3 h) multimodel ensembles were produced in support of the Hydrometeorological Testbed (HMT)-West-2006 campaign in northern California, covering the American River basin (ARB). Multiple mesoscale models were used, including the Weather Research and Forecasting (WRF) model, Regional Atmospheric Modeling System (RAMS), and fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5). Short-range (6 h) quantitative precipitation forecasts (QPFs) and probabilistic QPFs (PQPFs) were compared to the 4-km NCEP stage IV precipitation analyses for archived intensive operation periods (IOPs). The two sets of ensemble runs (operational and rerun forecasts) were examined to evaluate the quality of high-resolution QPFs produced by time-lagged multimodel ensembles and to investigate the impacts of ensemble configurations on forecast skill. Uncertainties in precipitation forecasts were associated with different models, model physics, and initial and boundary conditions. The diabatic initialization by the Local Analysis and Prediction System (LAPS) helped precipitation forecasts, while the selection of microphysics was critical in ensemble design. Probability biases in the ensemble products were addressed by calibrating PQPFs. Using artificial neural network (ANN) and linear regression (LR) methods, the bias correction of PQPFs and a cross-validation procedure were applied to three operational IOPs and four rerun IOPs. Both the ANN and LR methods effectively improved PQPFs, especially for lower thresholds. The LR method outperformed the ANN method in bias correction, in particular for a smaller training data size. More training data (e.g., one-season forecasts) are desirable to test the robustness of both calibration methods.

1. Introduction

Because of significant socioeconomic impacts, quantitative precipitation forecasts (QPFs) become critical weather elements in numerical weather prediction (Fritsch et al. 1998; Pielke and Downton 2000). Winter storms can cause severe damage and hazards in western coastal areas in North America (e.g., Ralph et al. 1999). In addition, wintertime precipitation provides most of freshwater supply for the western United States (Palmer 1988). Therefore, improvements in QPFs are extremely important for this area.

Water managers in western states are looking toward more efficient and economical operations by considering forecast-based operations (FBOs; Pugner 2003; G. Estes 2006, California Department of Water Resources, personal communication) that would allow water management decisions to be made on the basis of precipitation forecasts, which would be a significant departure from the current practice of using current and recent quantitative precipitation estimates (QPEs). Water management in the 0-to-5-day time frame is a complex process that must consider a wide range of obligations from public safety (flood avoidance), to agriculture (irrigation storage), to recreation (lake water levels), and to ecology (risk to fisheries). A move toward more accurate methods of water management through FBO would allow cost–benefit criteria to be combined with QPF and probabilistic QPF (PQPF) to allow improved risk management to support decisions on water storage or release. With sound QPF as input, hydrological models can be used to predict runoff by both deterministic and probabilistic methodologies. The work discussed in this paper is pertinent to this goal.
The Hydrometeorological Testbed (HMT; http://hmt.noaa.gov/) program, operated by the National Oceanic and Atmospheric Administration (NOAA), is intended to infuse new technologies, models, and scientific results into weather and river forecasts, beginning by emphasizing U.S. coastal weather, since 1997. The HMT-West-2006 winter campaign continued to focus on hydrological issues including quantitative precipitation estimation, observing, and forecasting over the American River basin (ARB; Fig. 1) in northern California. The ARB extends eastward from the low elevations of the Central Valley into the foothills and high peaks of the Sierra Nevada, with Lake Tahoe on the east side of the ridge. The ARB is a critical region, where the runoff during the cool season supplies most of the annual freshwater for the metropolitan Sacramento area. It also has high flooding potential during winter storms due to steep terrain. Winter weather over the ARB is dominated by synoptic-scale systems from the east Pacific ocean strongly modulated by orographic forcing (Dettinger et al. 2004).

In terms of hydrological applications, this paper mainly addresses QPFs and POPFs for the HMT-West-2006 campaign. Numerous early studies (e.g., Eckel and Walters 1998; Mullen and Buizza 2001) indicate that probabilistic forecasts may provide more information for decision-making systems. In the past decade, short-range ensemble forecasting (SREF, Brooks et al. 1995) has been developed and implemented in different operational centers worldwide (Lewis 2005, and references within), such as the breeding-perturbed SREF (e.g., Toth and Kalnay 1997; Du et al. 2006) at NOAA/National Centers for Environmental Prediction (NCEP) and the singular-vector-perturbed ensemble forecasts at the European Centre for Medium-Range Weather Forecasts (ECMWF; e.g., Molteni et al. 1996; Hersbach et al. 2000). Examples of multimodel ensembles include the superensemble method (Krishnamurti et al. 1999) and the NCEP SREF system. Time-lagged ensemble forecasts have been discussed for synoptic variables important for short-range (1–3 h) winter weather prediction (Lu et al. 2007), such as geopotential height, temperature, relative humidity, and wind at 40-km resolution.

Time-lagged ensembles can take advantage of forecasts from previous cycles to enlarge the ensemble size
without additional computational cost, and multimodel ensembles may advance weather forecasts by increasing ensemble spread. Therefore, a time-lagged multimodel ensemble system was created to provide weather forecasts during the HMT-West-2006 campaign by the NOAA/Earth System Research Laboratory (ESRL)/Global Systems Division (GSD) [formerly the NOAA/Forecast Systems Laboratory (FSL)]. Because high-resolution models can better represent heterogeneous topography over the ARB area, the ensembles were run at a very high resolution of 3 km out to 12 h. The 4-km NCEP stage IV precipitation analyses are used as the observations to evaluate 6-h QPFs and PQPFs during HMT-West-2006. Our goal is to examine the performance and forecast accuracy of such a high-resolution time-lagged multimodel ensemble system on predicting QPFs and PQPFs for winter storms.

The impacts of ensemble configurations on QPFs are discussed in the concept of selecting models, microphysical schemes, and initial–boundary conditions. For archived intensive operation periods (IOPs; Tables 1, 2), the two sets of ensembles, including operational runs and retrospective runs (rerun) with different combinations of models–physics, were analyzed. The impacts of microphysical schemes in high-resolution forecasts on rainfall volume (areal precipitation) over the ARB were investigated (Jankov et al. 2007) using a factor separation method (Stein and Alpert 1993). Based on their results, new microphysical schemes were included in the rerun ensembles for the selected IOPs, in order to reduce the wet biases shown in the HMT-West-2006 operations. Also, the diabatic initialization by the Local Analysis and Prediction System (LAPS; Albers 1995; Jian et al. 2003; Jian and McGinley 2005) can reduce the “spinup” problem of QPFs, and it was used in the ensemble members.

Besides ensemble configurations, postprocessing is a critical procedure to enhance forecast skill in a biased forecast system and can, in some applications, perform as well as current human forecasters (Mass 2003). Calibration of PQPFs is employed to reduce the forecast bias. Calibration in this paper refers to reliability, that is, the statistical consistency between the distributions of forecasts and the observations (Jolliffe and Stephenson 2003). Early studies applied model output statistics (MOS; Glahn and Lowry 1972), a multiple linear regression method, to calibrate PQPFs (e.g., Vislocky and Fritsch 1997). Other calibration methods of PQPFs include interpreting and adjusting rank histograms (e.g., Eckel and Walters 1998; Hamill and Colucci 1997), which is sensitive to the shape of rank histograms and the sample size, a logistic regression technique (Hamill et al. 2004), and an analog technique (Hamill et al. 2006) based on a 23-yr reforecast dataset. The analog method could improve both reliability and resolution of PQPFs. The analog and MOS methods require long-term training data of at least several years of historical forecasts. Limitations to the value of training samples arise from insufficient numbers (Atger 2003) and interdependence of the samples (Eckel and Walters 1998).

Artificial neural network (ANN) techniques, one type of nonlinear statistical regression, have also been used to improve PQPFs (e.g., Mullen et al. 1998; Mullen and Buizza 2004). For example, large wet biases of PQPFs (Yuan et al. 2005, 2007b) from a 12-km Regional Spectral Model (RSM; Juang and Kanamitsu 1994) ensemble system were greatly reduced through an ANN (Yuan et al. 2007a), which was trained using a 4-month dataset to calibrate 1-month forecasts on the 4-km NCEP stage IV grids over the southwest United States. Similar to HMT-West-2006, the RSM experiments had the complex terrain, orographically forced

<table>
<thead>
<tr>
<th>Event</th>
<th>Start</th>
<th>End</th>
<th>Blue Canyon precipitation (liquid equivalent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>IOP1</td>
<td>0000 UTC 1 Dec 2005</td>
<td>0900 UTC 2 Dec 2005</td>
<td>6.98 in.</td>
</tr>
<tr>
<td>IOP2</td>
<td>0100 UTC 18 Dec 2005</td>
<td>1700 UTC 19 Dec 2005</td>
<td>4.69 in.</td>
</tr>
<tr>
<td>IOP3</td>
<td>1800 UTC 20 Dec 2005</td>
<td>1200 UTC 22 Dec 2005</td>
<td>4.43 in.</td>
</tr>
<tr>
<td>IOP4</td>
<td>1200 UTC 30 Dec 2005</td>
<td>0000 UTC 1 Jan 2006</td>
<td>9.58 in.</td>
</tr>
<tr>
<td>IOP5</td>
<td>2100 UTC 1 Jan 2006</td>
<td>0000 UTC 3 Jan 2006</td>
<td>2.2 in.</td>
</tr>
<tr>
<td>IOP6</td>
<td>0000 UTC 11 Jan 2006</td>
<td>2100 UTC 11 Jan 2006</td>
<td>0.57 in.</td>
</tr>
<tr>
<td>IOP7</td>
<td>0300 UTC 14 Jan 2006</td>
<td>0300 UTC 15 Jan 2006</td>
<td>1.05 in.</td>
</tr>
<tr>
<td>IOP8</td>
<td>0000 UTC 18 Jan 2006</td>
<td>0100 UTC 19 Jan 2006</td>
<td>1.32 in.</td>
</tr>
<tr>
<td>IOP9</td>
<td>1200 UTC 28 Jan 2006</td>
<td>0900 UTC 29 Jan 2006</td>
<td>1.4 in.</td>
</tr>
<tr>
<td>IOP10</td>
<td>1200 UTC 1 Feb 2006</td>
<td>2000 UTC 2 Feb 2006</td>
<td>2.23 in.</td>
</tr>
<tr>
<td>IOP11</td>
<td>0300 UTC 4 Feb 2006</td>
<td>2100 UTC 4 Feb 2006</td>
<td>0.71 in.</td>
</tr>
<tr>
<td>IOP12</td>
<td>0000 UTC 27 Feb 2006</td>
<td>1500 UTC 28 Feb 2006</td>
<td>5.75 in.</td>
</tr>
<tr>
<td>IOP13</td>
<td>0600 UTC 2 Mar 2006</td>
<td>0300 UTC 3 Mar 2006</td>
<td>2.73 in.</td>
</tr>
</tbody>
</table>
winter precipitation, as well as high-resolution models and observations. Motivated by the successful RSM calibration, the ANN technique is adopted in this study to calibrate PQPFs over the ARB area. However, the sample size for this study is much smaller than that used for the RSM training. Lu et al. (2007) effectively corrected the biases for a time-lagged ensemble system using a linear regression (LR) method. Hence, the LR method is also examined in this paper. The bias correction is performed via a cross-validation procedure for the two sets of ensembles, and its benefit for this high-resolution time-lagged multimodel ensemble system is discussed.

The models and data are explained in section 2. Verification and calibration methods are introduced in section 3. In sections 4 and 5, respectively, the ensemble system is evaluated and calibration results described. Conclusions and summaries are given in section 6.

### 2. Model and data

From December 2005 to March 2006 during the HMT-West-2006 campaign, real-time weather forecasts were operated using three mesoscale models, including the Weather Research and Forecasting (WRF; http://www.mmm.ucar.edu/wrf/users/) model with the advanced research WRF (ARW) dynamic core version 2.2.1, the Regional Atmospheric Modeling System (RAMS; Cotton et al. 2003) version 6.0, and the fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5, version 3.7; e.g., Grell et al. 1995). The top pressure level was 100 mb and vertical levels were 31, 42, and 31 for the ARW, RAMS, and MM5 models, respectively. The Schultz (1995) microphysical scheme was employed in the MM5 model, while the Lin et al. (1983) microphysics was used in the ARW model and the RAMS model used its own microphysics. Boundary conditions (BCs) of the three models were provided by the 40-km North American Mesoscale (NAM; the former NCEP Eta Model; http://www.meted.ucar.edu/nwp/pcu2/NAMMay2005.htm) model. The RAMS and ARW models were also driven by the 40-km Rapid Update Cycle (RUC; Benjamin et al. 2004) model BCs. Hereafter we refer to these five model runs (ARW with NAM and RUC BCs, RAMS with NAM BCs, and MM5 with NAM BCs) as “operational” for the three IOPs (Table 2). Convective schemes were deactivated, which is appropriate for models operating horizontal resolution of 3 km. The large domain (Fig. 2) covers an area of about 450 km × 450 km including the ARB area [hereafter refers to the internal box (38.1°–39.6°N, 121.5°–119.6°W) in Fig. 2, the area east of 121.5°W in Fig. 1].

Except for the RAMS model, the ARW and MM5 models used the LAPS initialization, the diabatic “hot” start. The conventional “cold” start simulation usually suffers from spinup problems due to a lack of moisture information in the initial background conditions. LAPS incorporates a variety of real-time data from ground

### Table 2. Model configurations of 6-h ensemble forecasts for archived operational and rerun IOPs during HMT-West-2006.

<table>
<thead>
<tr>
<th>Event: No. of validation time start–end</th>
<th>Model (model–physics,BCs)</th>
<th>Time-lagged forecasts</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Operational IOPs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP10: 6</td>
<td>ARW—Lin.NAM</td>
<td>0–6 h</td>
</tr>
<tr>
<td>1200 UTC 1 Feb–1800 UTC 2 Feb 2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP12: 7</td>
<td>ARW—Lin.RUC</td>
<td>3–9 h</td>
</tr>
<tr>
<td>0000 UTC 27 Feb–1200 UTC 28 Feb 2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP14: 10</td>
<td>MM5—Schultz.NAM</td>
<td>6–12 h</td>
</tr>
<tr>
<td>0600 UTC 5 Mar–1200 UTC 7 Mar 2006</td>
<td>RAMS,NAM, no LAPS</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RAMS.RUC, no LAPS</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>3 IOPs: 23 6 h</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Five models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A total of 15 members</td>
<td></td>
</tr>
<tr>
<td><strong>Rerun IOPs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP1: 11</td>
<td>ARW—Ferrier.NAM</td>
<td>0–6 h</td>
</tr>
<tr>
<td>0600 UTC 30 Nov–1800 UTC 2 Dec 2005</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP4: 10</td>
<td>ARW—Thompson.NAM</td>
<td>3–9 h</td>
</tr>
<tr>
<td>1800 UTC 29 Dec 2005–0000 UTC 1 Jan 2006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IOP10: 7</td>
<td>MM5—Schultz.NAM</td>
<td>6–12 h</td>
</tr>
<tr>
<td>1200 UTC 1 Feb–0000 UTC 3 Feb 2006</td>
<td>RAMS.NAM</td>
<td>9–15 h</td>
</tr>
<tr>
<td>IOP12: 8</td>
<td>0600 UTC 26 Feb–0000 UTC 28 Feb 2006</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>Four IOPs: 36 6 h</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Four models</td>
<td></td>
</tr>
<tr>
<td></td>
<td>A total of 20 members</td>
<td></td>
</tr>
</tbody>
</table>

© 2008 American Meteorological Society. All rights reserved.
FIG. 2. The 6-h precipitation accumulations validated at 1800 UTC 27 Feb 2006 during IOP12 for (a) the stage IV observations and the 0–6-h operational forecasts from the WRF model with the Lin microphysics driven by the (b) RUC and (c) NAM BCs; the RAMS model without LAPS driven by the (d) RUC and (e) NAM BCs; and (f) the MM5 model with the Schultz microphysics driven by the NAM BCs. The 0–6-h rerun forecasts driven by the NAM BCs are from the WRF model with the (g) Ferrier and (h) Thompson microphysics and (i) the RAMS model. The time-lagged multimodel ensemble mean is for (j) the operational runs with the 15 members and (k) the rerun forecasts with the 20 members. The internal square box covers the ARB area. The RMSE (mm) and spatial correlation coefficients (cor) between each forecasted field and the stage IV observations for the large domain are shown in the titles.
observations, radar, aircraft, and satellite as well as background analyses provided by the 40-km NAM analyses. Diagnosed cloud information is input into a one-dimensional cloud model that recovers microphysical mixing ratios and vertical motions. LAPS then produces adjusted three-dimensional divergence, consistent mass fields, and complete moisture information at the model initial time and reduces the spinup time in the model integration. With 3-h initialization and 12-h projection, three time-lagged members (the 0–6-, 3–9-, and 6–12-h QPFs) were available from each of five models at the 6-h lead time. Therefore, a total of 15 ensemble members make up the operational ensemble forecasts.

Since the operational ARW model with the Lin microphysics showed huge wet biases, the ARW with different microphysics were rerun in order to improve ensemble forecasts. Four models driven by the NAM BCs, including the RAMS, the MM5 with the Schultz microphysics, and the ARW models with the Ferrier (Ferrier et al. 2002) and Thompson (Thompson et al. 2004) microphysics, were implemented for four IOPs (Table 2). We refer to these as “reruns.” All rerun models were initialized with LAPS every 3 h and run out to 18 h. Similarly, a total of 20 members composed the ensemble forecasts for 6-h QPFs at the 6-h lead time.

Verification data for this study are the 4-km NCEP stage IV 6-h precipitation analyses (http://www.emc.ncep.noaa.gov/mmib/ylin/pclpan/lstage4, hereafter stage IV), which are used as “truth” regardless of the observation uncertainty on the verification accuracy (Yuan et al. 2005). Model results were interpolated to the stage IV grids by inverse-distance weighting. Verification was performed by comparing concomitant observations and forecasts over the land. The hourly stage IV data showed larger discrepancies compared to gauge data than did the 6-h stage IV data over the ARB area (not shown). As a result, only 6-h QPFs were evaluated.

3. Evaluation and calibration methodology

a. Verification metrics

Using the verification framework by Murphy and Winkler (1987), verification metrics of QPFs and POPPFs include the root-mean-square error (RMSE), spatial correlation coefficients, Brier skill score (BrSS), ranked probability skill score (RPSS), area under relative characteristic curve (ROC), rank histograms (RH), and attributes diagrams (Jolliffe and Stephenson 2003; Wilks 2006; and in references therein). The skill score is computed based on sample climatology at each grid pixel in order to avoid false forecast skill (Hamill and Juras 2006). For example, the BrSS measures dichotomous events at a given threshold and are expressed as

\[ BrSS = 1 - \frac{BrS}{BrS_c} \]  

where the Brier score \( BrS = \frac{1}{n} \sum_{j=1}^{n} (\hat{p}_j - o_j)^2 \) is the difference between the forecast probability \( \hat{p}_j \) and the observed frequency \( o_j \) equals 1 for an occurred event, otherwise 0 for \( n \) samples (grid pixels). The referenced Brier score \( BrS_c = \sigma(1 - \sigma) \) is derived from sample climatology frequency \( \sigma \) computed at each grid pixel. The positive BrSS indicates skillful forecast with the perfect value of 1. The resampling method (Hamill 1999), that is, the bootstrapping method, is used to estimate the uncertainty ranges of a skill score. By randomly selecting the statistics 1000 times from all verification periods (with replacement) and computing a new score, the upper and lower bounds (95% and 5%) of a score are obtained to provide 90% confidence bounds (CBs).

b. Calibration methods

The ANN and LR techniques were adopted to calibrate POPPFs using seven ordered probabilities as the input data. For each grid pixel, a series of probabilities for sequential thresholds were calculated \([0.25, 1, 5, 10, 15, 20, 25, 50, 75, 100, 125, 150 \text{ mm (6 h)}^{-1}]\). Calibration was separately conducted for each selected threshold by training the closest seven probabilities from the probability series, which was designed to capture the essence of the probability density function (PDF) centered at the target probability threshold. For example, at the 5 mm \((6 \text{ h})^{-1}\) threshold, the probabilities of 0.25, 1, 5, 10, 15, 20, 25 mm \((6 \text{ h})^{-1}\) were the input data; at 25 mm \((6 \text{ h})^{-1}\), the input data were probabilities at the thresholds 10, 15, 20, 25, 50, 75, 100 mm \((6 \text{ h})^{-1}\), where the selected threshold is shown in italic. As in Eq. (3.1), the target output data is a dichotomous observed probability (0 if the observed precipitation is less than a given threshold, otherwise 1). The output data derived in the ANN (LR) model is a probability specific to a target threshold for corresponding grid pixels. To derive a stable relationship between the input and output data (observations), the training data included all available grid pixels over the ARB area during the training period. The trained weights by the ANN (LR) model were saved and applied to the validation data to compute new POPPFs.

Cross validation was performed as part of the calibration procedure. For the three operational IOPs (Table 2), the verification data from two of these cases were used to calibrate the third. For example, IOP10 and IOP12 were used to calibrate IOP14; IOP10 and
IOP14 to calibrate IOP12, etc. Similarly, three of the four rerun IOPs were used to calibrate the fourth (IOPs 1, 4, 10 to calibrate IOP12, etc.). The calibration results shown in section 5 combined the three operational IOPs in one set of figures and the four rerun IOPs in another set of figures. After the bias correction for all selected thresholds, the calibrated probabilities were examined to ensure that probabilities at lower thresholds were not smaller than probabilities at higher thresholds to enforce a monotonic property. Outliers were corrected by setting the lower probability value at a lower threshold to the higher probability at the adjacent higher threshold. More outliers were found for relatively rare cases due to the sample limitation. The LR calibration over the ARB area revealed that of the three operational IOPs, the calibration for IOP10 (0.6% grid pixels) and IOP14 (1.5% grid pixels) had far fewer outliers than IOP12 (19% grid pixels). The calibrated rerun IOP4 (13% grid pixels) showed many more outliers than the rerun IOP1, IOP10, and IOP12 (0.055%, 0.02%, and 0 grid pixels, respectively).

The ANN (Hsu et al. 1995; Yuan et al. 2007a) used in the bias correction is a three-layered, feed-forward, and nonlinear model, which includes a linear least squares simplex algorithm with global optimization searching. The input layer (Ninput = 7 nodes) takes in the input data, the above-mentioned seven probabilities. The output layer (Noutput = 1 node) functions to create the probability value between 0 and 1. The hidden layer (varying nodes, e.g., Nhidden = 4 nodes) connects the input and output nodes through a set of weights, which are computed through a logistic sigmoid function by minimizing the RMSE between the output data and the target observation. The weights are applied to the validation dataset for obtaining new probabilities using the same ANN structure.

The LR method is an alternative simplified method and the LR equation is as follows:

$$P(x, t) = a + \sum_{i=1}^{M} b_if_i(x, t),$$

where $M = 7$, $f_i(x, t)$ is the ordered input probabilities for $i = 1, 2, \ldots, 7$; $P(x, t)$ is the observed probability; and $a$ is a constant interpreting error residual. The coefficients $b_i$ and $a$ are estimated by minimizing the errors of the target observed probabilities and the derived probabilities in the LR using $N$ training data samples, that is, to meet the least squares function:

$$\min_{a, b_i} \sum_{j=1}^{N} \left( P(x, t) - \left[ a + \sum_{i=1}^{M} b_if_i(x, t) \right] \right)^2.$$
differ in the dynamics, physics, and many other aspects related to QPF. Second, the microphysical scheme plays an important role in QPFs. With the same ARW model, the simulation using Lin microphysics (Figs. 2b,c) demonstrates a much higher wet bias than using the Ferrier (Fig. 2g) and Thompson (Fig. 2h) microphysics, with especially strong intensities along the windward slopes of the mountainous areas in terms of the RMSE and spatial correlation coefficients. Third, the initialization is critical for producing QPFs. Inclusion of the LAPS initialization in the rerun RAMS modeling (Fig. 2i) enhanced precipitation compared to the operational RAMS run without LAPS (Fig. 2e), especially for mountainous precipitation cells. Last, BCs also lead to different forecasts. The NAM BCs show fewer biases than the RUC BCs in both the ARW (Fig. 2b versus Fig. 2c) and RAMS models (Fig. 2d versus Fig. 2e). However, the magnitude of differences is smaller compared to that caused by different microphysical schemes (Fig. 2c versus Figs. 2g,h). Jankov et al. (2007) examined the sensitivity of three factors (microphysics, planetary boundary layer schemes, and initial conditions) on the ARB-averaged QPFs for HMT-2006-West. The study showed that changing microphysical schemes caused significant changes in QPFs. More tests over the ARB area are needed to obtain optimal ensemble design.

To compare the two sets of ensembles, the RMSE of 0–6-h QPFs was computed and averaged from the 11 simultaneous valid periods during IOP10 and IOP12 (Fig. 3). The errors are generally larger over the ARB area than over the large domain, especially for the ARW model with the Lin microphysics. Consistent with Fig. 2, the ensemble mean of the rerun forecasts shows the smallest errors (Fig. 3) of all forecasts, while the operational ensemble mean has generally smaller errors than or comparable to individual operational models. The average spatial correlation coefficients (Fig. 4) for both operational and rerun forecasts show better skill in the ensemble mean than individual models. Unlike the RMSE (Fig. 3), the ARW model with the Lin microphysics is acceptable in spatial distribution (Fig. 4). High-resolution time-lagged multimodel ensembles can improve QPFs in both magnitude (Fig. 3) and spatial distribution (Fig. 4), compared to the 0–6-h forecasts from individual models.

b. Forecast skill of PQPFs

As indicated by 90% CBs, the BrSS (Fig. 5) shows that the operational cases (Fig. 5a) have skill up to the 15 mm (6 h)$^{-1}$ threshold and the CBs of uncertainties extend below zero for higher thresholds [20 and 25 mm (6 h)$^{-1}$], while the rerun cases (Fig. 5b) are skillful to the 25 mm (6 h)$^{-1}$ threshold. The skill over the ARB area has larger variations with wide ranges of CBs than over the large domain. The new ensemble configuration in the rerun cases led to skillful forecasts, especially over the ARB area and for the higher thresholds. Analysis of forecast bias for the individual models (Figs. 3 and 4) also indicates that the improved forecast skill (Fig. 5b) mainly resulted from changing to microphysics with smaller biases.

The RPSS is an extension of the BrSS using multiple
categories. Figure 6 shows the spatial distribution of the RPSS using four thresholds [1, 5, 10, 20 mm (6 h)$^{-1}$] to define five categories. Most areas (Figs. 6a,b) show skillful forecasts with positive values over the large domain. Higher values of the RPSS are generally found along high terrain and coastal areas. Overall, the four rerun IOPs demonstrate more skill over the large domain than the operational cases, especially over Nevada and the northwest part of California. Over the ARB area, the rerun cases exhibit better higher RPSS with good values of $\sim$0.5 and above along the mountain regions and highest values reaching $\sim$0.7 and greater over the region to the northwest of Lake Tahoe. These RPSS results are consistent with the BrSS (Fig. 5).

c. Discriminating ability

The area under the ROC curve can be used to show the forecast ability to discriminate dichotomous events (e.g., precipitation occurrence at a given threshold). Since the ROC curve stratifies forecasts based on observations, that is, conditioned on observations, the ROC area is insensitive to forecast bias (Jolliffe and Stephenson 2003). The perfect ROC area is 1, while unskillful forecasts have the value of 0.5 or less. With an assumption of homoscedasticity, that is, the constant variance across subsets of the data, the value of $\sim$0.75 and higher of the ROC area is an indicator of a good forecast. The discriminating ability is much higher over the ARB area than it is over the large domain (Fig. 7) for both the operational and rerun IOPs. The values of the ROC area are elevated to the range of good forecasts in the rerun cases. The increase is especially large for higher thresholds over the ARB area, for which the ROC area is $\sim$0.9 and above at all thresholds. The enhancement of forecast skill (e.g., BrSS) in the rerun cases is not only because of the reduction of forecast bias (Fig. 2), but also due to an improvement in discriminating ability.

Fig. 5. Brier skill scores for 6-h precipitation accumulations for (a) the three operational IOPs and (b) the four rerun IOPs at 1 to 25 mm (6 h)$^{-1}$ thresholds. The dashed lines with dark circles represent the ARB area and the solid lines with open circles denote the large domain. Error bars indicate 90% confidence bounds. The curves for the large domain are slightly offset to the right of the operational curves for clarity.

Fig. 6. The spatial distribution of the ranked probabilistic skill score for 6-h precipitation accumulations from (a) the three operational IOPs and (b) the four rerun IOPs using four thresholds: 1, 5, 10, and 20 mm (6 h)$^{-1}$. The internal square box covers the ARB area.
d. Forecast bias

Rank histograms depict the rank of the observation data as compared to available ensemble members and show the frequency of each rank for all grid pixels. For example, by fitting into sorted 15 ensemble members at a grid pixel, the possible rank of the observation can be from 1 to 16. Uniform rank frequencies (horizontal solid line in Fig. 8) are expected for a perfect ensemble system. Extreme populations are situated at the lowest ranks (Figs. 8a,b) for the two sets of ensembles. This “L” shape of RH distributions is indicative of an overestimation in precipitation forecasts. That is, the observation ranks low relative to all the ensemble forecasts. Wet biases exist in both operational and rerun cases over the ARB area. A “U” shape of RH distributions in the rerun IOPs (Fig. 8b) suggests an insufficient ensemble spread (i.e., standard deviation of ensemble forecasts), in which observations frequently fall either above or below the range of ensemble members. RH distributions over the large domain (not shown) result in smaller deviations from uniform ranks due to larger data samples and smaller sample variability.

The attributes diagrams for the ARB area (Fig. 9) further demonstrate the forecast bias. Before calibration, the reliability curves for the four rerun IOPs (dark circles) fall below the 1:1 diagonal line (perfect forecast), which is indicative of wet biases in PQPFs considering 90% CBs. Slight dry biases are shown for lower probabilities at 1 and 5 mm (6 h)$^\text{-1}$ (Figs. 9a,b). Internal gray bars show the raw frequency for each probability category in total events, in which the 21 forecast probabilities were binned into 11 categories for plotting clarity. The reliability curves of the three raw operational IOPs (not shown) demonstrate wet biases for 5-mm and higher thresholds, whereas the 90% CBs show larger variation at 25 mm (6 h)$^\text{-1}$.

5. The calibration results

a. Attributes diagrams

With a cross-validation procedure, both the ANN and LR methods were performed over the ARB area for the two sets of ensembles. The LR method exhibited better calibration. Reliability curves of rerun cases show consistent improvement after the LR bias correction (open circles in Fig. 9) for all thresholds with 90% CBs, whereas the calibrated reliability curves are closer to the 1:1 diagonal line. At lower thresholds [1–15 mm (6 h)$^\text{-1}$], dry and wet biases were corrected by decreasing the sharpness, which was accomplished by moving lowest and highest probabilities to midrange probabili-
ties shown in internal histograms before (gray bars) and after calibration (blank bars). At the 20 and 25 mm (6 h)\(^{-1}\) thresholds, the higher probabilities shifted to the lower probabilities after calibration. The calibration of operational IOPs overcorrected PQPFs with a “dry” tendency or the lack of higher probabilities (not shown). More cases along with a better discriminating ability may help to derive relatively more stable weights.
for training rerun IOPs. In general, conditional biases were reduced by the calibration procedure using several IOPs.

b. Reliability, resolution, and uncertainty

The Brier score can be decomposed into three components—reliability, resolution, and uncertainty (Wilks 2006). The reliability term, so-called conditional bias, measures the agreement between the observed frequency and forecast probability, which is negatively oriented (smaller values are better) and corresponds to the squared distance between the diagonal line and the reliability curve weighted by the forecast frequency on an attributes diagram (Fig. 9). The resolution term, which is positively oriented and measures the forecast ability to discriminate occurrence/nonoccurrence of the events from climatology, corresponds to the squared distance between the reliability curve and sample climatology frequency (a horizontal line; not shown) weighted by the forecast frequency. The uncertainty term depends solely on sample climatology frequency, and therefore it is unchanged by calibration. Whenever the resolution is greater than the reliability, the forecast is skillful with a positive BrSS as compared to the sample climatology.

The decomposition of the BrS (Fig. 10) shows that for the three operational IOPs (Fig. 10a), the reliability term is greatly reduced after calibration with a slightly decreased resolution term. For the four rerun IOPs (Fig. 10b), with subtle change in the reliability term, the resolution term decreases slightly for 1 and 5 mm and increases above the 10-mm threshold. The ROC area also changes slightly after calibration (not shown), since the ROC area is more closely related to the discriminating ability (resolution) than to the conditional bias. Comparing the change of the reliability and resolution terms, the skill has been enhanced at higher thresholds, but not at 1 mm. With different model configurations, the calibration processes may improve the different aspects of forecast skill. A large reduction of reliability in the operational IOPs and a small increase of resolution in the rerun IOPs contribute to improving forecast skill.

c. Brier skill score

The BrSS (Fig. 11) shows that the skill evidently increases at 10-mm and higher thresholds after calibration using the LR and ANN methods. The skill decreases or changes only slightly for PQPFs at 1 and 5 mm (6 h)$^{-1}$, which possess smaller biases (Figs. 9 and 10) and thus gain little improvement in calibration. The promising result is that forecast skill improves for heavier events, which are more important for hydrological forecasts. The BrSS does not show significant difference with respect to the overlapping of the CBs before and after calibration, but the uncertainty range decreases as the low error bound greatly increases to the skillful range or higher values. This indicates that the calibrated results may provide more reliable forecasts. The LR method outperformed the ANN method, with fewer differences in the rerun cases (Fig. 11b) but distinct discrepancies for the operational cases with rare events at higher thresholds (Fig. 11a). By training four years of forecasts data in a season for selected gauges, Kuligowski and Barros (1998) determined that...
the calibrated QPFs had lower RMSE by the LR model than by a backpropagation ANN and the latter had better threat scores for heavy precipitation events. Short data length in this study limits the ANN advantage at higher thresholds. In addition, the RPSS (not shown) was mostly improved over relatively low elevations of the ARB area, with little variation over high terrains. This suggests that using longer historical data from small local regions may improve the calibration as opposed to blending all data over a large complex topography.

d. Training data

The sensitivity of calibration to the selection of the localized and nonlocalized training data was examined. By expanding the training data to include grid pixels in the large domain, outliers associated with alteration of the monotonic probability distribution (section 3) in the calibrated operational IOP12 decreased from 19% to 14% grid pixels over the ARB area, while outliers in the calibrated rerun IOP4 decreased from 13% to 5.3%. The other cases had no outliers. Conditional biases in the attributes diagrams and forecast skill scores, however, were not improved (not shown). The operational IOPs at 20- and 25-mm thresholds were an exception with rare rainfall events in the training. More training data (such as one-season forecasts or multiple-year forecasts) for local calibration with similar climatology is expected to improve PQPFs. The nonlinear ANN requires more data than the LR method. The selected input variables may follow a more linear pattern, and this also favors the LR technique. This can be seen in the attributes diagrams (Fig. 9). At higher thresholds (Figs. 9b,d), the reliability curves show similar patterns and the slope shifts toward the same direction after calibration, while at 1- and 5-mm thresholds (Figs. 9a,b), the shape of the curves changes after calibration and appears more nonlinear variation.

6. Summary and discussion

QPFs and PQPFs from time-lagged multimodel ensembles were produced during HMT-West-2006 over the ARB area, which possesses complex topography and heterogeneous surfaces. Forecast skill highly depends on the choice of the model, parameterization, initialization, and BCs. In general, the rerun forecasts with the new configuration possessed better forecast skill and discriminating ability than the operational ones. Judicious selection of microphysical schemes is very critical for high-resolution simulations, and the biases of orographically forced precipitation were more sensitive to microphysical schemes than to the selection of BCs. For example, the Lin microphysics in the ARW model showed extremely wet biases over the mountainous domain compared to the Ferrier and the Thompson microphysics. The “spinup” problem was alleviated by the LAPS initialization in the RAMS simulations with slight wet biases. Multimodel ensembles showed improved skill in ensemble-mean QPFs over the forecasts from individual models. With limited computational resources, good choice of ensemble configurations and model formulations are critical in improving ensemble forecasts.

Postprocessing can also improve ensemble forecasts. With a cross-validation procedure, both the LR and ANN methods improved forecast skill of PQPFs over the ARB area, reduced uncertainties of skill scores, and mitigated conditional biases by decreasing the reliability term for the operational IOPs and increasing the resolution term for the return IOPs. Improvements by the LR calibration were greater than those produced by the ANN calibration, especially for the operational cases at higher thresholds. The fewer parameters in the LR model and more nearly linear character of the bi-

![Fig. 11. Brier skill scores for 6-h precipitation accumulations over the ARB area from (a) the three operational IOPs and (b) the four rerun IOPs before (dashed line with dark circles) and after the LR (solid line with open circles) and ANN (open triangles) calibration. Error bars indicate 90% confidence bounds. The calibrated curves are slightly offset to the right of the original curves for clarity.](image-url)
ases in the input data also could favor the LR calibration. The sample size of the training data poses challenges for both calibration methods, especially for using data collected during a short period with respect to the interdependence of data samples over a small region. By training using samples from the large domain, the number of outliers altering monotonic PQPFs over the ARB area was decreased but the skill was not improved [20 and 25 mm (6 h)^{-1} were an exception in the operational IOPs]. Classification of the training data by topography or smaller regions with similar climatology is expected to improve bias correction.

The use of high-resolution time-lagged multimodel ensembles can provide valuable QPF and PQPF for a wide variety of water management applications. As stated in the introduction the move to FBO is a concept requiring a new approach to forecasts and forecast postprocessing. We have demonstrated with our work during HMT-2006 that calibrated QPF and PQPF can be provided by a small ensemble. Coupled with hydrological models, these can be used to derive a range of water flow scenarios and probabilities that can serve the needs of FBO. In addition, methodologies developed using the time-lagged multimodel ensemble and postprocessing framework are applicable to other weather forecast requirements. Methods applied in this paper can be extended to other HMT applications including temperature and precipitation type, and to support HMT in other seasons and regions such as the planned HMT-East effort in 2009. HMT-2006 cases will be rerun with the new ensemble configuration to be used in HMT-2007 (December 2006 to March 2007) to build up the training dataset. Several WRF-based model configurations with different microphysics and dynamic cores will be implemented. The impacts of sample size on the calibration and other calibration methods for correcting the ensemble mean and second moment will be examined.

Besides enhancing short-range weather forecasts, another goal of the HMT program is to improve QPE. The choice and quality of verification datasets highly affect the apparent verification skill of ensemble forecasts. Since 1-km precipitation estimation data will be available over the ARB in the future, cross-validation of the stage IV data and ensemble forecasts using this QPE and other data will be investigated for studying observational uncertainty and improving model forecasts.

Acknowledgments. This research was performed while the first author held a National Research Council Research Associateship Award at NOAA/ESRL. Thanks to Dr. Kuo-lin Hsu for providing the neural network code. The computational facilities were provided and maintained by the NOAA/ESRL/GSD. The authors wish to thank Dr. Ed Tollerud for his insightful review and valuable suggestions. Thanks to Ms. Annie Reiser for editing this manuscript and Mr. John Osborn for producing the river basin map. We also thank the three anonymous reviewers for their valuable suggestions.

REFERENCES


——, and J. Juras, 2006: Measuring forecast skill: Is it real skill or


