Effects of Spatial and Temporal Aggregation on the Accuracy of Statistically Downscaled Precipitation Estimates in the Upper Colorado River Basin

SUBHRENDU GANGOPADHYAY

Cooperative Institute for Research in Environmental Sciences, and Department of Civil, Environmental, and Architectural Engineering, University of Colorado, Boulder, Colorado

MARTYN CLARK

Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado

KEVIN WERNER AND DAVID BRANDON

Colorado Basin River Forecast Center, Salt Lake City, Utah

BALAJI RAJAGOPALAN

Department of Civil, Environmental, and Architectural Engineering, University of Colorado, Boulder, Colorado

(Manuscript received 20 October 2003, in final form 19 May 2004)

ABSTRACT

To test the accuracy of statistically downscaled precipitation estimates from numerical weather prediction models, a set of experiments to study what space and time scales are appropriate to obtain downscaled precipitation forecasts with maximum skill have been carried out. Fourteen-day forecasts from the 1998 version of the National Centers for Environmental Prediction (NCEP) Medium-Range Forecast (MRF) model were used in this study. It has been previously found that downscaled temperature fields have significant skill even up to 5 days of forecast lead time, but there is practically no valuable skill in the downscaled precipitation forecasts. Low skill in precipitation forecasts revolves around two main issues. First, the (intermittent) precipitation variability on daily and subdaily time scales could be too noisy to derive meaningful relationships with atmospheric predictors. Second, the model parameterizations and the coarse spatial resolution of the current generation of global-scale forecast models might be unable to resolve the local-scale variability in precipitation. Both of these issues may be addressed by spatial and temporal averaging.

In this paper the authors present a diagnostic study using a set of numerical experiments to understand how spatial and temporal aggregations affect the skill of downscaled precipitation forecasts in the upper Colorado River basin. The question addressed is, if the same set of predictor variables from numerical weather prediction models is used, what space (e.g., station versus regional average) and time (e.g., subdaily versus daily) scales optimize regression-based downscaling models so as to maximize forecast skill for precipitation? Results in general show that spatial and temporal averaging increased the skill of downscaled precipitation estimates. At subdaily (6 hourly) and daily time scales, the skill of downscaled estimates at spatial scales greater than 50 km was generally higher than the skill of downscaled estimates at individual stations. For the 6-hourly time scale both for stations and for mean areal precipitation estimates the maximum forecast skill was found to be approximately half that of the daily time scale. At forecast lead times of 5 days, when there is very little skill at daily and subdaily time scales, useful skill emerged when station data are aggregated to 3- and 5-day averages.

1. Introduction

Numerical weather prediction (NWP) models are among the most important tools used by operational agencies to make accurate and meaningful quantitative precipitation forecasts (QPFs; Antolik 2000). Medium-range forecasts from the current generation of global-scale NWP models are laden with biases and have poor skill in many regions (e.g., Clark and Hay 2004). Model output statistics (MOS) guidance is thus necessary to postprocess NWP output to produce reliable operational QPFs (Antolik 2000). Several approaches have been used for MOS (e.g., Glahn and Lowry 1972) and range from simple bias corrections to developing complex models based on parametric and nonparametric statistical methods. These methods have also been used by

Corresponding author address: Subhrendu Gangopadhyay, CSTPR/CIRES, University of Colorado, Campus Box 488, Boulder, CO 80309-0488.
E-mail: subhrendu.gangopadhyay@colorado.edu

the global change community to estimate the climate impacts associated with enhanced atmospheric concentrations of greenhouse gasses and are referred to as statistical downscaling (SDS).

MOS and SDS methods can be broadly classified into three categories: 1) transfer functions, 2) weather generators, and 3) weather typing. A common transfer function downscaling approach is multiple linear regression (MLR), where grid-cell values from atmospheric models are used as predictors for surface variables such as local precipitation and temperature (e.g., Clark and Hay 2004). Regression models can also be developed using principal components of the predictor fields (Hewitson and Crane 1992), through canonical correlation analysis (CCA) and singular-value decomposition (SVD) analysis (von Storch and Zwiers 1999). In addition to linear regression, nonlinear models such as the ones based on nonlinear interpolation (Brandsma and Buishand 1997), geostatistics (Biau et al. 1999), and artificial neural networks (Hewitson and Crane 1996), have also been developed. Weather generators on the other hand are stochastic models of climate variability (Wilks 1999). Weather generators are typically developed for daily time scales (e.g., Rajagopalan and Lall 1999; Buishand and Brandsma 2001; Yates et al. 2003), but subdaily models are also available (Katz and Parlange 1995). The weather generators can also be conditioned upon large-scale atmospheric states to translate output from global-scale models into useful information at local scales (e.g., Katz and Parlange 1996; Wilby 1998). The third methodological group, weather typing, is a synoptic downscaling approach where weather classes are related to local and regional climate variations. The weather classes can be defined synoptically or derived specifically for downscaling purposes, for example, by constructing indices of airflow (Conway et al. 1996). Another weather typing approach is based on the nonparametric classification and regression trees (CART; Breimann et al. 1984) analysis. CART is a complex classification scheme based simultaneously on large-scale and local climate variables and has been applied primarily to simulate local daily rainfall (Schnur and Lettenmaier 1998). Nonparametric SDS approaches based on analogues have also been developed to estimate local precipitation (Zorita and von Storch 1999; Gangopadhyay et al. 2002).

Though there are several methods to perform SDS, all methods strive to extract the signal from global-scale models that is useful to describe variability in surface climate at local scales. For precipitation, the noisy character of precipitation fields often masks the signal in global-scale models. It is possible to reduce the noise in precipitation fields through both temporal aggregation and spatial averaging. This may help maximize the skill that is extracted from global-scale forecast models and increase the skill in downscaled precipitation estimates. The question we address is, if we use the same set of predictor variables, what space (e.g., station versus regional average) and time (e.g., subdaily versus daily) scales optimize downscaling models so as to maximize forecast skill for precipitation?

The paper next describes the datasets and downscaling experiments (section 2). Section 3 presents the downscaling methodology and the skill measure used to compare the downscaling experiments. Results and discussions on the space–time aggregation experiments are presented in section 4. A summary of the research and conclusions ends the presentation (section 5).

2. Datasets and downscaling methodology

We use reanalysis datasets from the National Centers for Environmental Prediction (NCEP) 1998 Medium-Range Forecast (MRF) model in this study. This section briefly describes the data archives from the NCEP 1998 MRF model available from the National Oceanic and Atmospheric Administration (NOAA) Climate Diagnostic Center (CDC), the predictor variables used in this study, precipitation datasets for stations and mean areal precipitation (MAP) fields in the upper Colorado River basin from the Colorado Basin River Forecast Center (CBRFC), and the downscaling experiments.

a. The CDC forecast archive

The NOAA CDC has generated a ''reforecast'' dataset (1979–present) using a fixed version (circa 1998) of the NCEP operational MRF model (Hamill et al. 2004). Output variables used in this study are (a) the accumulated precipitation for a 12-h period (e.g., 0000 UTC–1200 UTC), (b) 2-m air temperature, (c) relative humidity at 700 hPa, (d) 10-m zonal wind speed, (e) 10-m meridional wind speed, (f) total column precipitable water, and (g) mean sea level pressure. These 7 variables were selected from a potential list of over 300 model variables, and were shown by Clark and Hay (2004) to be particularly useful for downscaling precipitation.

b. Precipitation datasets and downscaling experiments

Hydrologic models run by the River Forecast Centers (RFCs) in the United States are based upon National Weather Service River Forecasting System (NWSRFS) guidelines and use precipitation and temperature data that are interpolated to basin subareas. River basins are divided into subbasins, and are further subdivided typically into three catchment areas based on elevation bands. Six-hourly MAP and mean areal temperature (MAT) fields are estimated for each of the catchment areas following preprocessing and calibration guidelines outlined by the NWS (Anderson 2004; Larson 2004). The distribution of stations and MAP areas in the upper Colorado basin is shown in Fig. 1. There are 10 hourly stations, 60 daily stations, and 64 subbasins in the study

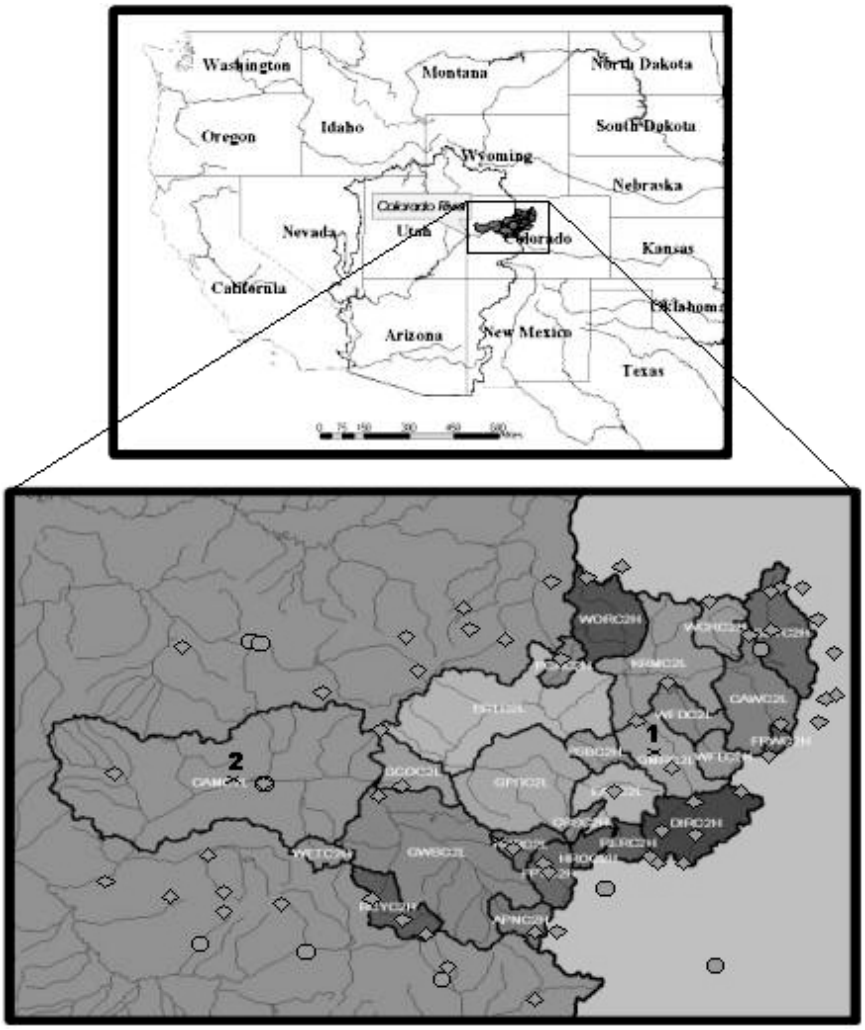

FIG. 1. Location of the upper Colorado River basin and a map (not to scale) showing subbasins (MAP areas), daily stations (diamonds), and hourly stations (circles).

area. The hourly and daily stations were selected on the basis that at least 90% of data were available for the period 1979–98. The estimated MAP fields for each of the 64 subbasins are calibrated using 6-hourly mean station precipitation values (see, Anderson 2004; Larson 2004).

From the raw station datasets (hourly data at hourly stations and daily data at daily stations, dly_dly) and MAPs (6-hourly values, map_6hy), several datasets were derived for the downscaling experiments (Table 1). These derived datasets are 1) 6-hourly precipitation values at the hourly stations each averaged over the four 6-hourly periods, 0000–0600, 0600–1200, 1200–1800, and 1800–2400 UTC (hly_6hy); 2) daily precipitation

TABLE 1. Datasets used in downscaling experiments.

Notation	Description
hly_dly	Daily data derived at the hourly stations from raw hourly station data.
dly_dly	Raw daily data at the daily stations.
map_dly	Daily data derived for the MAP regions from calibrated 6-hourly MAP data.
hly_6hy	6-hourly data derived at the hourly stations from raw hourly station data.
map_6hy	Calibrated 6-hourly MAP data for the MAP regions.
dly_3da	Daily data at the daily stations derived using a 3-day moving average of dly_dly.
dly_5da	Daily data at the daily stations derived using 5-day moving average of dly_dly.
map_3da	Daily data at the MAP regions derived using a 3-day moving average of map_dly.
map_5da	Daily data at the MAP regions derived using a 5-day moving average of map_dly.

estimates at the hourly stations (hly_dly); 3) daily precipitation values at each of the 64 subbasins (map_dly); 4) 3-day moving average at the daily stations (dly_3da); 5) 5-day moving average at the daily stations (dly_5da); 6) 3-day moving average at the MAPs; and 7) 5-day moving average at the MAPs. These precipitation datasets were used to test the effect of spatial and temporal aggregation on the accuracy of downscaled precipitation estimates.

To assess the effect of spatial aggregation on the skill of downscaled precipitation forecasts we consider two time scales: daily and 6 hourly. We downscale and generate 14-day forecasts for the three daily cases—hly_dly, dly_dly, and map_dly—as well as for the two 6-hourly cases—hly_6hy and map_6hy. The datasets and acronyms are summarized in Table 1. We also analyzed the sensitivity of downscaled precipitation estimates to spatial averaging of daily station data up to a radius of 150 km. This experiment was restricted to the daily time scale, as there was not sufficient data to carry out this experiment at the subdaily time scale.

To study the effect of temporal aggregation, we considered two spatial scales—point scale and regional scale. The point scale consists of hourly and daily stations where downscaling is compared for the cases hly_6hy, hly_dly, and dly_dly. The regional scale includes map_6hy and map_dly. Also, to look into the effects of temporal aggregation at longer forecast lead times, we downscale to the four cases dly_3da, dly_5da, map_3da, and map_5da, where the suffixes 3da and 5da represent 3- and 5-day averages, respectively (see Table 1).

3. Statistical downscaling methodology and model evaluation

a. Statistical downscaling methodology

We developed our statistical downscaling models based on multiple linear regressions (von Storch and Zwiers 1999). Predictors (see section 2a) from the NCEP MRF model archive for the period 1979–2001 were used to develop regression equations and to forecast precipitation at individual sites (stations or MAP regions). A separate equation was developed for each space (hourly stations, daily stations, and mean areal averages) and time (6-hourly, daily, 3- and 5-day averages) case, and for each forecast lead time by month. Forecasts were generated for a lead time of up to 14 days, and all forecasts were initialized at 0000 UTC on that day. For a given day, we use forecasts from three adjacent 12-h periods, 0000–1200 UTC from day+0 (5 P.M.–5 A.M. local time in Colorado), forecasts from 1200–0000 UTC on day+0 (5 A.M.–5 P.M. local time in Colorado), and forecasts from 0000–1200 UTC on day+1 (5 P.M.–5 A.M. local time in Colorado). This provides a total of 21 potential predictor variables in the regression equations for the subdaily and daily time scales, and helps account for possible temporal phase errors in the model output. For downscaling to 3- and 5-day averages of station and mean areal precipitation values, a set of seven predictor variables were derived using a 3- and 5-day moving window respectively on the original predictor set.

The intermittent and skewed character of the daily and subdaily precipitation data make it necessary to preprocess these data prior to developing the regression equations. The site time series of precipitation was first disaggregated into a time series of occurrence (1 = wet days and 0 = dry days) and precipitation amounts (only wet days). Logistic regression was used to model precipitation occurrence, and ordinary least squares regression was used to develop models for precipitation amounts.

The time series of occurrence is used as the response variable for the logistic regression model, and the time series of precipitation amounts is used as the response variable for the ordinary least squares model. For precipitation amounts, the station/MAP precipitation data (only wet days) are transformed to a normal distribution using a nonparametric probability transform (Panofsky and Brier 1963). For each data point, the cumulative probability of observed precipitation is computed. This is matched with the cumulative probability from a standard normal distribution (mean of zero and standard deviation of one), and the normal deviate corresponding to the cumulative probability in the standard normal distribution is used to replace the original precipitation value.

The regression models have the form

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k + \cdots + \alpha_n x_n + \varepsilon, \quad (1)$$

$$p = 1 - \frac{1}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \cdots + \beta_n x_n)}, \quad (2)$$

where, Eq. (1) is for ordinary least squares regression and Eq. (2) is for logistic regression. In these equations, y is the response variable in the ordinary least squares model (precipitation amount at a station/MAP), p is the response variable in the logistic regression model (e.g., the probability of precipitation at a station location), and n is the total number of predictors. In these equations, α_0 and β_0 are the regression constants: α_k and β_k are the slope coefficients for the kth explanatory variable (x_k, $k = 1, \ldots, n$), and ε is the error term. The ex-

planatory variables (x_k, x_2, . . . , x_n) are forecasted outputs from the NCEP MRF model (e.g., 700-hPa relative humidity, mean sea level pressure). The solution of ordinary least squares equation was done using the SVD algorithm (Press et al. 1992), and the logistic regression equation was solved iteratively using the method (and code) presented by Agterberg (1989). The forward selection approach was used to identify the variables used in the regression equations (Antolik 2000; Clark et al. 2004).

Once a regression model has been developed, a value of y can then be predicted from the regression equations ($\hat{y}$) for each data point in the time series [Eq. (1)]. The residuals ($\varepsilon = y - \hat{y}$) were then tested for independence, normality, and constant variance. We used the turning point test (Clarke 1984), and the skewness test of normality (Snedecor and Cochran 1989) respectively to test the assumptions that the residuals are independent and normally distributed. Since the regression models were developed in normal space it was expected that the residuals will follow a normal distribution, and we found that the hypothesis of normality was accepted in all cases at the 5% significance level. The residuals also passed the test of independence at the 5% significance level. To test whether the residuals have a constant variance we used a graphical diagnosis. Residuals were plotted against the model predictions, and examined if the variability between the residuals remained relatively constant across the range of the predicted values. We observed a similar degree of variability around the mean for the residuals across the range of predicted precipitation amounts (not shown). This gives confidence that the residuals indeed come from a population with constant variance. Generally, nonnormality and nonconstant variance are a related problem, and from all the tests and graphical diagnosis we can conclude that the residuals are independent and come from a normal distribution with a constant variance.

The next step was to generate ensembles. To account for the intermittent properties of precipitation, precipitation is modeled in a two-stage process. Logistic regression is used to estimate precipitation occurrence [Eq. (2)], and ordinary least squares regression [Eq. (1)] is used to estimate precipitation amounts. Precipitation was modeled (in normal space) as follows:

$$y_{\text{iens}} = \begin{cases} 0, & \text{when } \hat{p} < u \\ \hat{y} + N\sigma_e, & \text{when } \hat{p} \geq u. \end{cases} \quad (3)$$

In Eq. (3), $u \sim U(0, 1)$ is a random number from a uniform distribution ranging from 0 to 1, and $\hat{p}$ is the probability of precipitation occurrence predicted from the logistic regression model [Eq. (2)]. If $\hat{p} < u$, then we assume there is no precipitation. If $\hat{p} \geq u$, precipitation is set to occur and the precipitation amount is computed using Eqs. (1) and (3). When $\hat{p} \geq u$, the forecasted normal deviates from Eq. (3) (y_{iens}) are then transformed back to the original skewed distribution of observed precipitation using the nonparametric probability transform technique described above. The stochastic modeling of the regression residuals inflates the variance of precipitation-reducing problems of variance underestimation that are typical of regression-based models. One hundred ensembles were generated using this approach.

b. Model evaluation

Model evaluation and all comparisons are carried out using the probabilistic skill measure, ranked probability skill score (RPSS). RPSS is a measure of categorical forecast skill and is computed as follows (Wilks 1995). The RPSS is based on the ranked probability score (RPS) computed for each forecast–observation pair:

$$\text{RPS} = \sum_{m=1}^{J} (Y_m - O_m)^2, \quad (4)$$

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: First, the observed time series is used to distinguish 10 possible categories (J) for forecasts of precipitation (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, forecast lead time, and station. Next, for each forecast–observation pair, the number of ensemble members forecast in each category is determined (out of 100 ensemble members), 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. (4)]. The RPSS is then computed as

$$\text{RPSS} = 1 - \frac{\overline{\text{RPS}}}{\overline{\text{RPS}}_{\text{clim}}}, \quad (5)$$

where $\overline{\text{RPS}}$ is the mean ranked probability score for all forecast–observation pairs, and $\overline{\text{RPS}}_{\text{clim}}$ is the mean ranked probability score for climatological forecasts [i.e., where there is an equal probability in each of the m categories; Eq. (4)].

RPSS is calculated for each of the downscaling experiments described in Table 1, and the different cases are compared through plots of cumulative distribution functions (CDF). The RPSS value indicate the fraction of times (or equivalently as a percentage) we can gen-

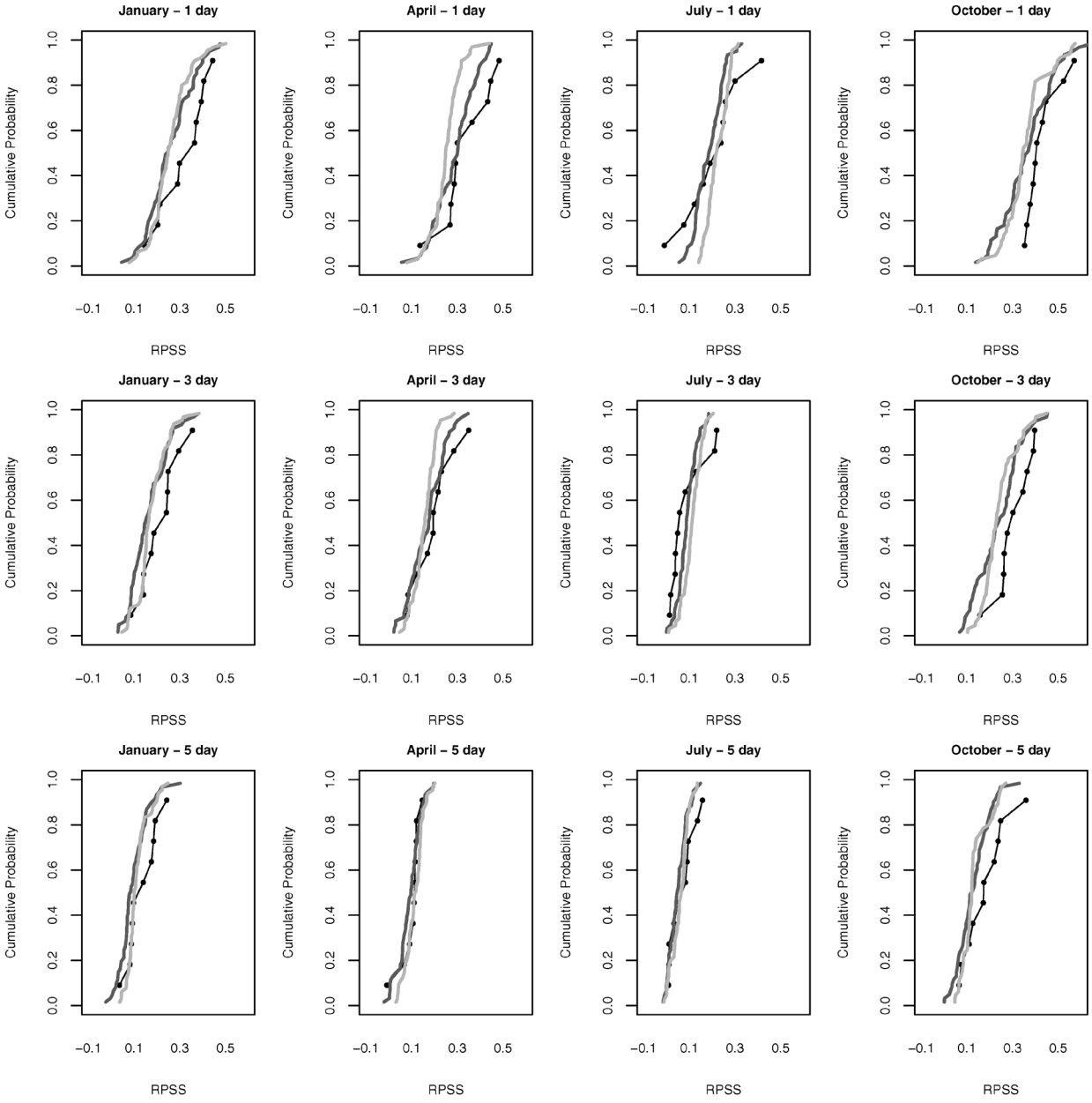

FIG. 2. Cumulative probability distribution of RPSS for hly_dly (points and lines), dly_dly (black line), and map_dly (gray line) for Jan, Apr, Jul, and Oct at forecast lead times of 1, 3, and 5 days.

erate forecasts better than the climatology. A negative RPSS value implies that we perform worse than climatology. Another point to note is the number of ensemble members used in calculating the RPSS. Here we used 100 ensemble members, as presently we do not have the computing capabilities to use a large number of ensembles. However, we did test (not shown) the sensitivity of the forecast skill to sets of 30, 40, 50, . . . , 100 realizations to identify the asymptotic behavior and successive improvements in skill. We found that the differences in RPSS between different combinations of 30 ensembles were quite large, but this difference was small for different combinations of 100 ensembles.

4. Results and discussions

a. Spatial aggregation

To analyze the effect of spatial aggregation, we fix the time scales to daily and 6 hourly (subdaily). We

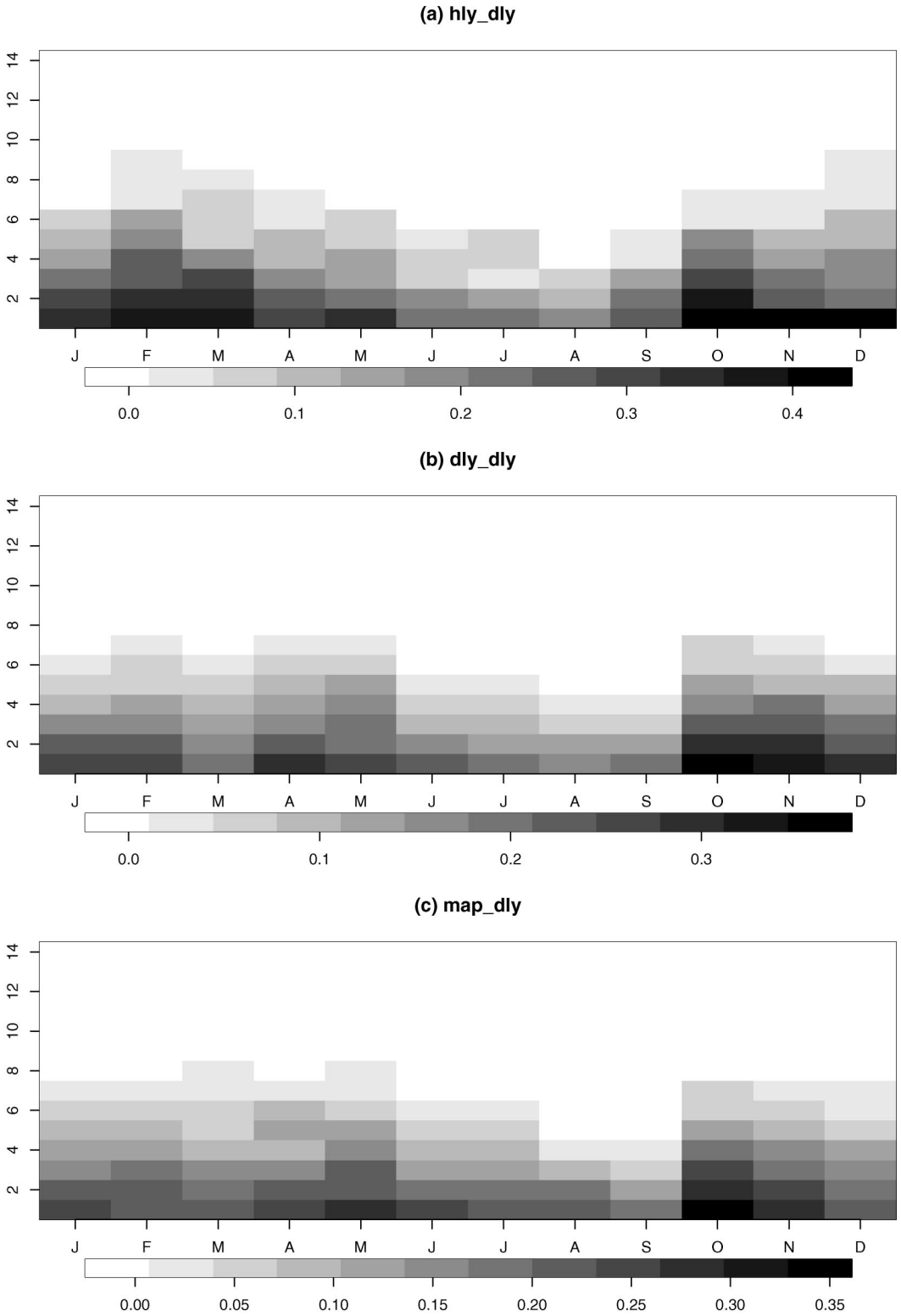

FIG. 3. Distribution of median basin RPSS by month (horizontal axis) for forecast lead times of 1 through 14 days (vertical axis) in the three daily cases: (a) hly_dly, (b) dly_dly, and (c) map_dly.

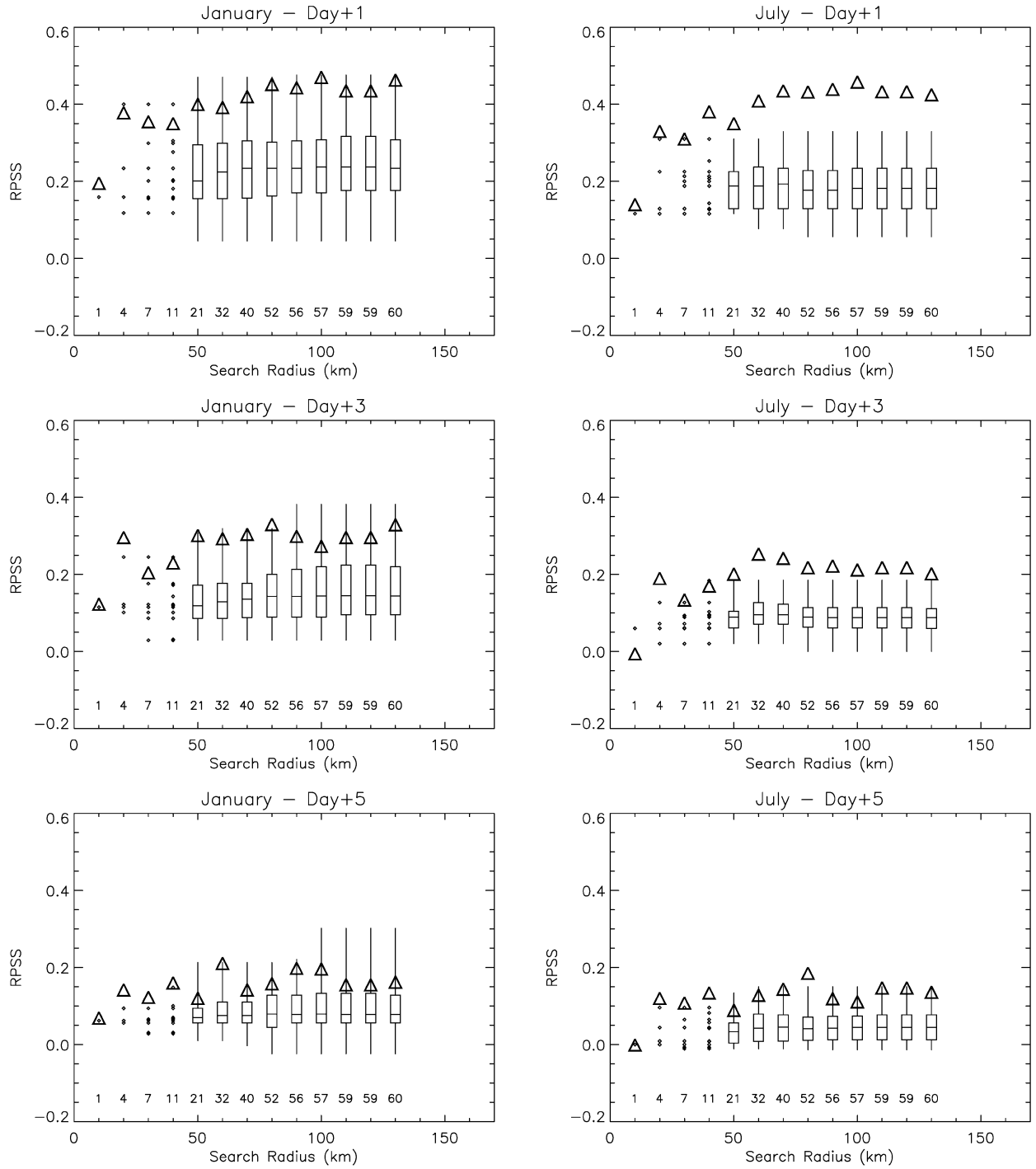

FIG. 4. Sensitivity of RPSS to averaging distance and number of stations (shown above bottom axis) for reference point 1 (refer to Fig. 1) for (left) Jan and (right) Jul for forecast lead times of 1, 3, and 5 days. Box plots are plotted when number of stations averaged are ≥20, otherwise stations are represented with single points. Triangles show the RPSS for reference point 1 for each of the averaging cases.

then compute RPSS at spatial scales of stations and for mean areal precipitation regions and compare the values for these two time scales. The sensitivity of RPSS to spatial averaging of precipitation stations is also analyzed.

1) DAILY TIME SCALE

Figure 2 shows the comparison of RPSS for the three cases—hly_dly (points and lines), dly_dly (thick black line), and map_dly (gray line)—in each plot (see Table

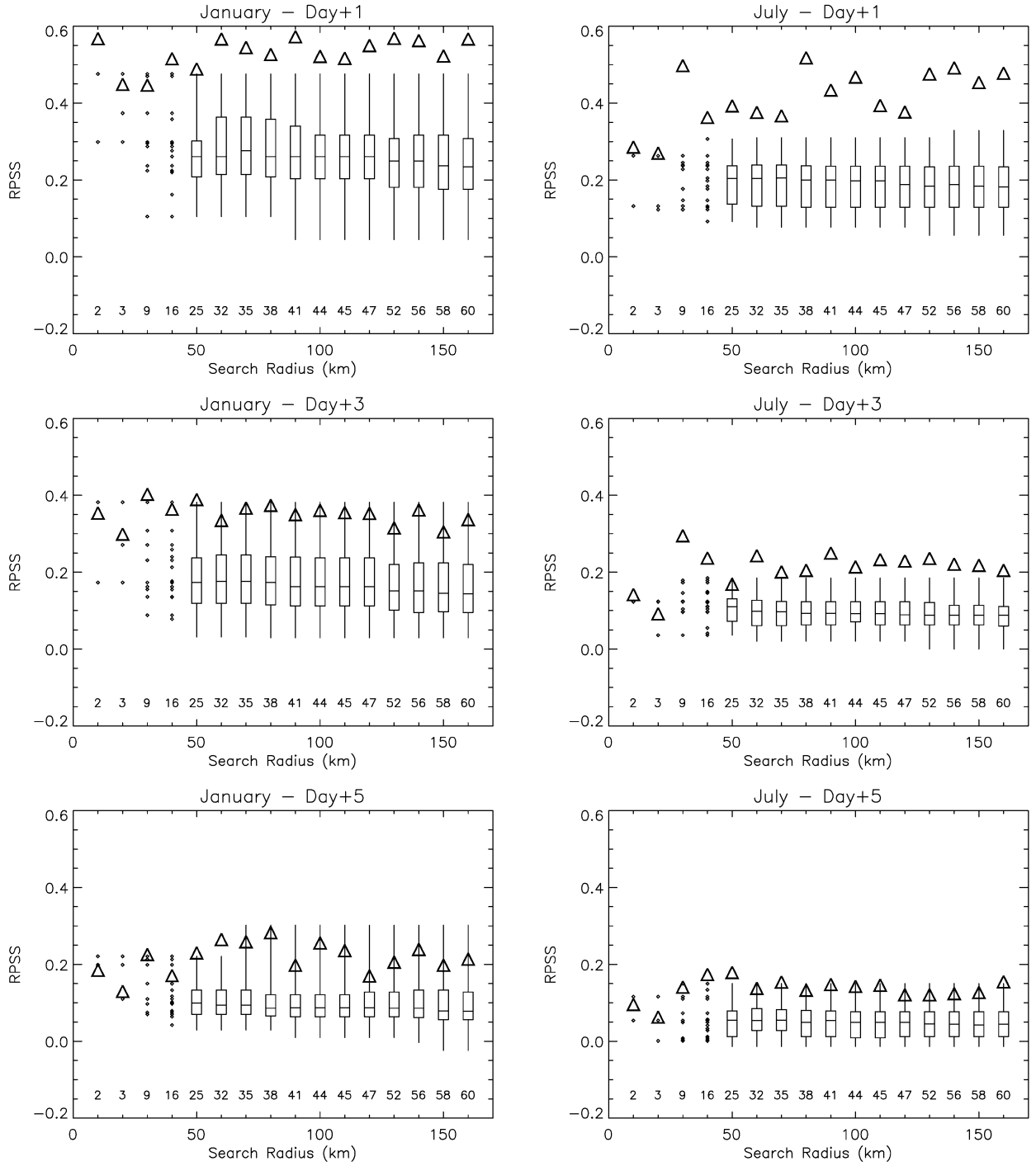

FIG. 5. Same as Fig. 4, but for reference point 2 (refer to Fig. 1).

1 for the definition of acronyms). Results are plotted for the months of January, April, July, and October for forecast lead times of 1, 3, and 5 days. The results show that, in general, skill is greater for the hly_dly case over the dly_dly and map_dly cases. There is no apparent reason as to why the skill at the daily time scale for the hourly stations should be greater. Because of limited number of hourly stations (only 10) in this case, this is likely a sampling bias. However, higher skill for hly_dly may also be due to better measurements of precipitation at the hourly stations. The skill of daily station data and daily MAPs is similar. At short forecast lead times (1

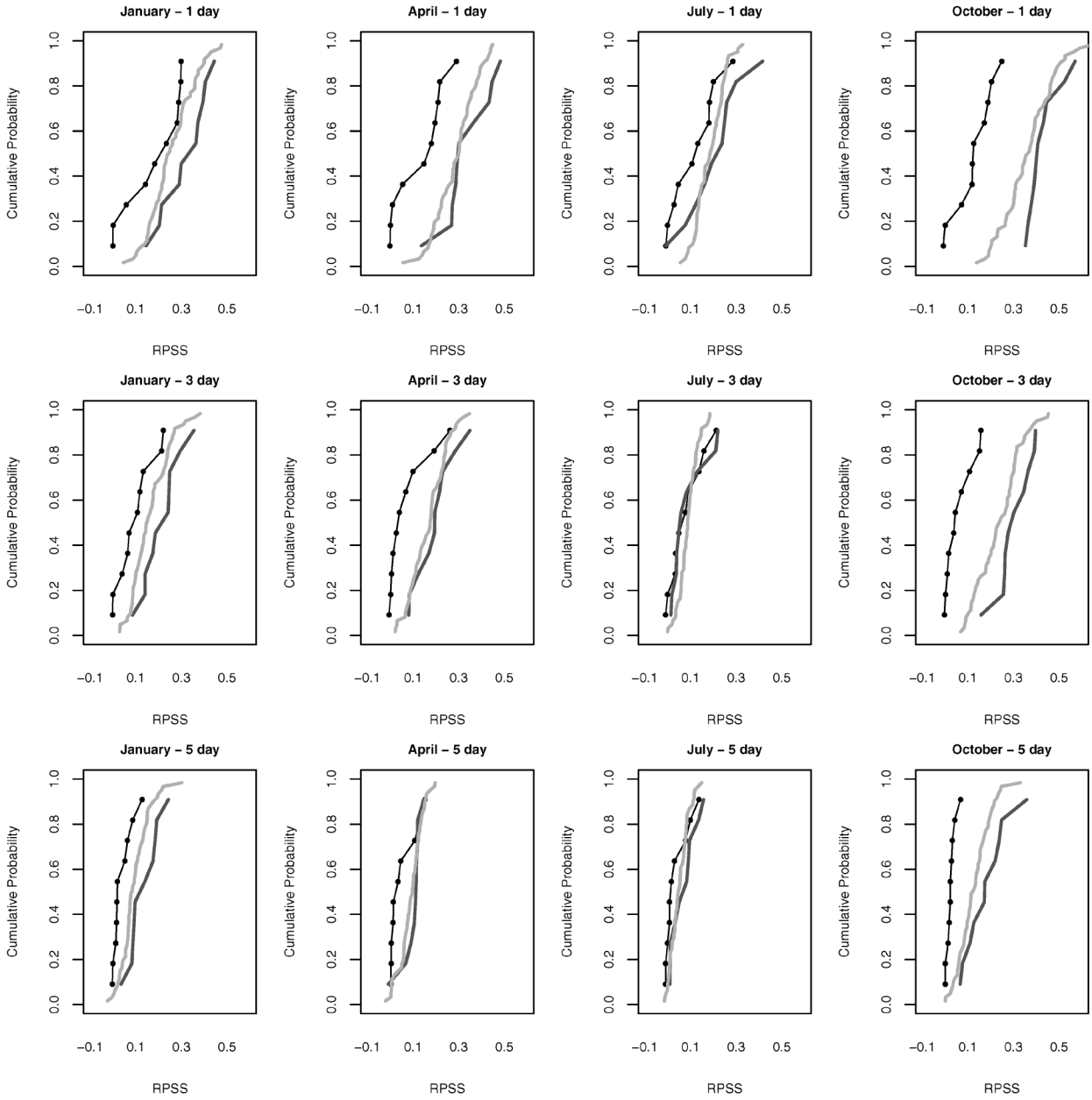

FIG. 6. Same as Fig. 2, but for hly_6hy (points and line), hly_dly (black line), and dly_dly (gray line).

and 3 days), the RPSS CDFs are more strongly separated than at longer forecast lead times (5 days). This is expected, as there is very little skill at 5 days, whatever dataset is used.

An initially surprising result from Fig. 2 is the close agreement between the station (dly_dly) and MAP (map_dly) CDFs. A priori, one may expect the MAPs to have higher skill because of the noise reduction associated with spatial averaging. However, this small difference between stations and the basin subareas is related to the method used to construct mean areal precipitation values. MAP for the basin subareas is constructed as a weighted average of the surrounding stations. In contrast to other interpolation methods (e.g., where the weights are based on inverse distance), the weights for each station are determined in the hydrologic model calibration process. That is, various station weights are tried, and optimal station weights are identified when the difference between observed and modeled runoff is minimized. In this method it is common that only one or two of the surrounding stations have very high weights, and the other surrounding stations have very low weights. Thus, the MAP estimates can, in some cases, simply be considered as single-station estimates.

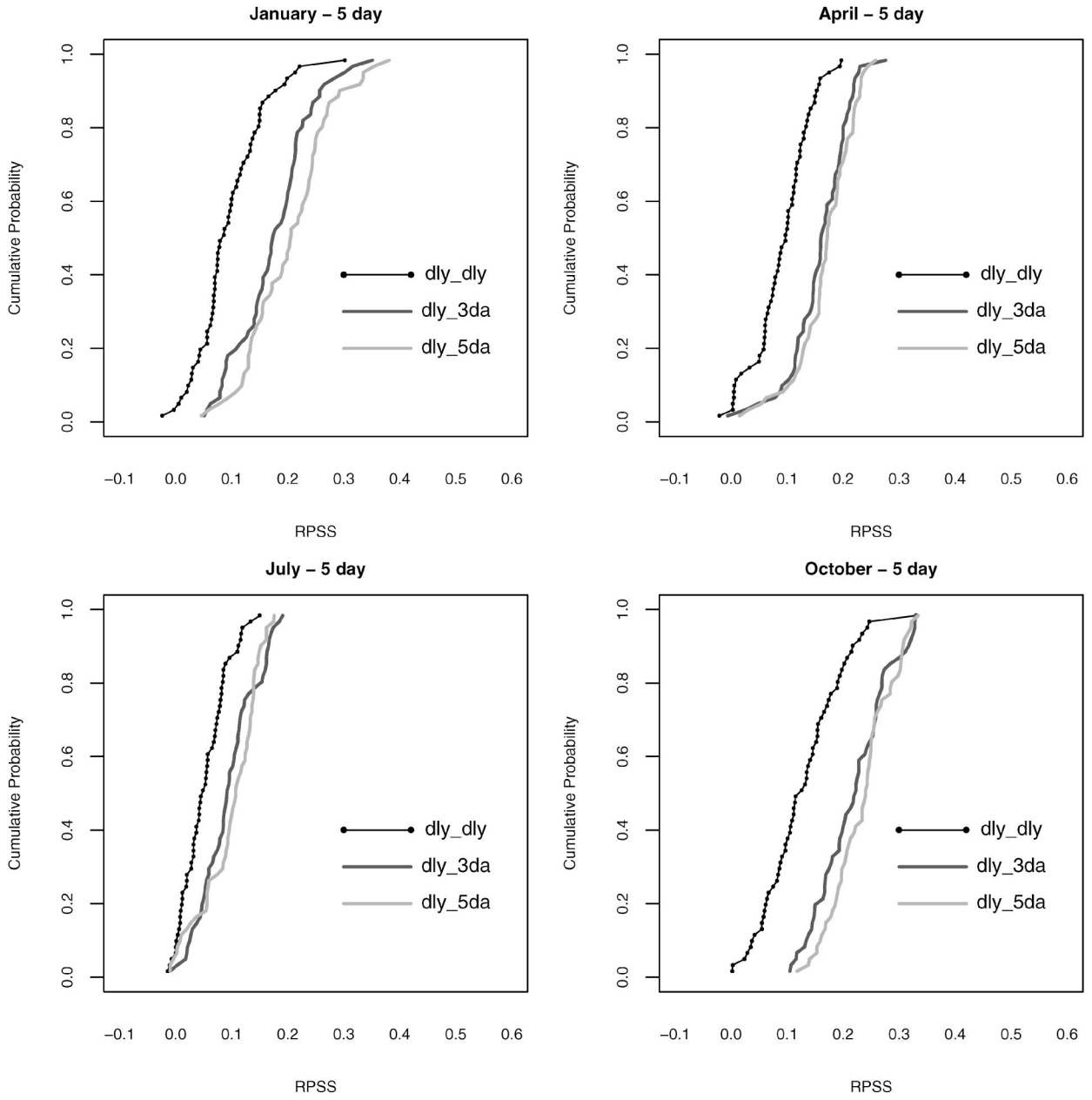

FIG. 7. Cumulative probability distribution of RPSS at 5-day forecast lead time for dly_dly, dly_3da, and dly_5da in Jan, Apr, Jul, and Oct.

Figure 3 shows the distribution of forecast skill by month for up to 14 days of forecast lead time in the three cases, (a) hly_dly, (b) dly_dly, and (c) map_dly. These are representative median skills for the entire basin. The median skill is calculated in each case from the RPSS values over all stations or subareas and for a given month and forecast lead time. Essentially, there is no valuable skill in the downscaled forecast beyond a lead time of 5 days. Also the skill in these three cases is the least during the summer months and is highest during winter. This can be interpreted as the variability in atmospheric circulation and precipitation is more spatially coherent during winter. In summer, most of the precipitation results from convective storms, which are much more localized.

Similar results (not shown) were also obtained for the subdaily time scale (6 hourly). Overall, downscaled estimates of MAP have higher skill than downscaled estimates for individual stations. However, these results should be viewed cautiously because of the small number of the hourly stations. Also, the forecast skill was generally found to be higher for the daily time scale

TABLE 2. Distribution of daily stations for the two reference points.

Reference point 1 (dense) lat = 39.75°N, lon = 106.25°W, elev = ~8800 ft			Reference point 2 (sparse) lat = 39.50°N, lon = 108.00°W, elev = ~5600 ft		
Category (d in km)	no. of stations (cumulative)	Distance range (km)	Category (d in km)	no. of stations (cumulative)	Distance range (km)
1 ($d < 10$)	1 (1)	5.549	1 ($d < 10$)	2 (2)	5.979–9.714
2 ($10 \leq d < 20$)	3 (4)	15.563–18.756	2 ($10 \leq d < 20$)	1 (3)	18.019
3 ($20 \leq d < 30$)	3 (7)	20.841–29.778	3 ($20 \leq d < 30$)	6 (9)	23.218–29.635
4 ($30 \leq d < 40$)	4 (11)	30.579–34.676	4 ($30 \leq d < 40$)	7 (16)	31.941–38.256
5 ($40 \leq d < 50$)	10 (21)	40.769–48.522	5 ($40 \leq d < 50$)	9 (25)	43.679–49.964
6 ($50 \leq d < 60$)	11 (32)	50.216–58.857	6 ($50 \leq d < 60$)	7 (32)	51.672–59.953
7 ($60 \leq d < 70$)	8 (40)	61.913–69.522	7 ($60 \leq d < 70$)	3 (35)	61.424–68.334
8 ($70 \leq d < 80$)	12 (52)	70.684–79.820	8 ($70 \leq d < 80$)	3 (38)	70.941–79.483
9 ($80 \leq d < 90$)	4 (56)	80.405–88.947	9 ($80 \leq d < 90$)	3 (41)	86.457–88.567
10 ($90 \leq d < 100$)	1 (57)	92.115	10 ($90 \leq d < 100$)	3 (44)	90.043–98.294
11 ($100 \leq d < 110$)	2 (59)	102.367–108.287	11 ($100 \leq d < 110$)	1 (45)	103.612
12 ($110 \leq d < 120$)	0 (59)	—	12 ($110 \leq d < 120$)	2 (47)	111.563–115.312
13 ($120 \leq d < 130$)	1 (60)	123.430	13 ($120 \leq d < 130$)	5 (52)	122.455–129.138
			14 ($130 \leq d < 140$)	4 (56)	131.404–139.348
			15 ($140 \leq d < 150$)	2 (58)	147.316–147.569
			16 ($150 \leq d < 160$)	2 (60)	150.375–155.053

than for the subdaily time scale. The effect of temporal aggregation on forecast skill is examined further in section 4b.

2) RPSS SENSITIVITY TO SPATIAL AVERAGING

To further understand the effects of spatial averaging on the accuracy of downscaled precipitation estimates we carried out experiments using strong, better-measured, and more spatially coherent storm events. The approach used to identify such events is discussed in the appendix. These experiments were carried out for two locations within the basin marked as 1 and 2 in Fig. 1. Reference point 1 is located in a higher elevation (~8800 ft) within a dense neighborhood of daily stations, while reference point 2 (elevation: ~5600 ft) was selected within a sparse neighborhood of daily stations. Neighboring stations were included with an incremental search radius of 10 km until all 60 daily stations were included. The details of this averaging process are given in Table 2.

We next separately downscaled to each point, 1 and 2, for all regional averages summarized in Table 2. RPSS from this downscaling experiment is shown in Fig. 4 (reference point 1) and Fig. 5 (reference point 2). Plots are made for winter (January) and summer (July) up to a forecast lead time of 5 days. RPSS for each category (see Table 2) is plotted with the averaging distance (search radius). The search radius was incremented progressively by 10 km until all the 60 daily stations were covered. The number of stations averaged within a given distance is also shown in the plots. The skill from each of the individual stations averaged is shown as points when the number of stations was less than 20 and as box plots when the number of stations averaged was 20 or more. The RPSS for the 13 spatial averages for point 1 and the 16 spatial averages for point 2 are depicted with triangles. For reference point 1, the RPSS clearly increases as more stations are averaged (triangles in Fig. 4). This is most pronounced in July for the forecast lead time of 1 day, but is also evident for other months and lead times. Reference point 2 is located in a lower-elevation region characterized by generally higher downscaling skill (not shown). The improvement in skill with distance is less pronounced for reference point 2 (Fig. 5), as the regional average at larger spatial scales includes the high-elevation stations, which, by themselves, have lower downscaling skill. Nevertheless, the skill for the regional average (triangles) is much higher than the skill at most of the individual stations. Note in particular that for the forecast lead time of 5 days when the RPSS for most individual stations is generally less than 0.1, the RPSS for the regional average is about 0.2. Because of limited data from hourly stations, this exercise was carried out only with daily stations.

b. Temporal aggregation

Analogous to spatial aggregation, we first fix the space scales to points (i.e., stations) and regions (i.e., mean areal precipitation estimated over basin subareas) to analyze the effect of temporal aggregation. Then we compare and contrast RPSS at 6-hourly and daily time scales for these two spatial scales. Finally, we analyze the effects of temporal averaging over 3 and 5 days at large forecast lead times.

1) POINT SCALE

The three cases compared are hly_6hy, hly_dly, and dly_dly (see Table 1). Figure 6 shows the CDFs for these three cases in January, April, July, and October at forecast lead times of 1, 3, and 5 days. The skill for the 6-hourly station data (hly_6hy) is consistently lower than

the skill for the daily station datasets (hly_dly and dly_dly), demonstrating the higher skill for longer averaging times. The hourly station data aggregated to the daily time scale (hly_dly) has slightly higher skill than the raw daily station data (dly_dly), which likely is a sample bias. The three CDFs in summer are closer to each other (mostly at longer lead times), reflecting lack of skill in downscaled precipitation at this time of the year.

Also, for the regional scale we compared the cases map_6hy and map_dly (not shown). Similar to the point scale, at the regional scale, the daily time scale provides consistently higher forecast skill.

2) RESULTS FOR LONGER FORECAST LEAD TIMES

We next analyzed the effect of temporal aggregation at the forecast lead time of 5 days at the point scale (Fig. 7). The results are shown for January, April, July, and October and show that skill is much higher (i.e., nonzero) when data are aggregated to 3- and 5-day averages. These results demonstrate that even if there is no useful skill at the daily time scale, useful skill can be obtained at longer averaging times. These results may be useful to assess the probability of significant precipitation several days into the future. The 3- and 5-day averages show equivalent skill. Once again in summer, the skill is much lower, but averaging does improve the forecast skill at longer forecast lead times. Similar results were obtained for experiments carried using the MAP areas.

5. Summary and conclusions

In this paper we have analyzed the effect of spatial and temporal aggregation on the skill of downscaled precipitation estimates in the mountainous upper Colorado River basin. Previous research by the authors (e.g., Clark and Hay 2004) has shown that skill in precipitation forecasts through statistical downscaling is quite limited, and this study was undertaken to understand the appropriate space–time scales that should be used in a downscaling procedure. The downscaling model used is based on a transfer function approach developed using multiple linear regressions. The predictors (a total of seven variables) for the multiple linear regression models were selected from the ''reforecast'' analysis carried out using the operational version of the NCEP 1998 MRF model. We carry out several experiments combining space scales ranging from points to areal averages (up to 150 km) and time scales ranging from 6-hourly to 5-day averages.

As expected, spatial and temporal averaging increased the skill of downscaled precipitation estimates. At subdaily and daily time scales, the skill of downscaled estimates at spatial scales greater than 50 km was generally higher than the skill of downscaled estimates at individual stations. Also, since total accumulated precipitation from the MRF model was used as one of the predictor variables, and is itself an area average, it is expected to be a better predictor for area-averaged precipitation than for station precipitation. Furthermore, at forecast lead times of 5 days, when there is very little skill at daily and subdaily time scales, useful skill emerged when station data are aggregated to 3- and 5-day averages. Extending these findings to the operational setting is not straightforward. For example, the NWS operational streamflow models require 6-hourly inputs for each of the subbasins. Further work is required to assess if downscaled estimates at larger spatial scales and longer averaging times, when disaggregated to 6-hourly values, produce improved streamflow forecasts.

Acknowledgments. This work was supported by the NOAA GEWEX Americas Prediction Program (GAPP) and the NOAA Regional Integrated Science and Assessment (RISA) Program. We thank Jeffery Whittaker and Tom Hamill at the NOAA Climate Diagnostics Center in Boulder Colorado for providing output from historical runs of the NCEP MRF model, and Brent Bernard from CBRFRC for his GIS support. We also thank the three anonymous reviewers for their helpful comments.

APPENDIX

Estimation of Stronger and Spatially Coherent Storm Events

To estimate spatially averaged precipitation at each of the reference points, we performed a two-step preprocessing of the daily station data for the period 1979–2001 (8401 days) to identify the stronger and spatially coherent storm events. First, we assessed the fraction of time (α) when a given percentage (expressed here as decimal percent, p) of all the stations within a given search radius have valid (nonmissing) precipitation data (includes both dry and wet stations). This is computed in the following two steps: (a) we compute the fraction of stations with valid data (p) for all the 8401 days in the time series, and for all averaging points in Table 2, and (b) for each averaging point, we calculate the fraction of time (α) the fraction of stations with valid data is above the thresholds $p = 1.0$, 0.95, 0.90, 0.85, and 0.8. Results are shown in the two top panels of Fig. A1. The most striking result from this exercise is the small fraction of time when all stations have valid precipitation data (i.e., $p = 1.0$). This is most pronounced for reference point 2 for larger spatial averages (e.g., >30 stations), meaning that if all stations are used there will be insufficient data to develop reliable downscaling models. However, for both the reference points, there is data from at least 85% of the stations in slightly over 80% of the days in the time series. That is, $\alpha > 0.8$ for $p \geq 0.85$ when the number of stations averaged is great-

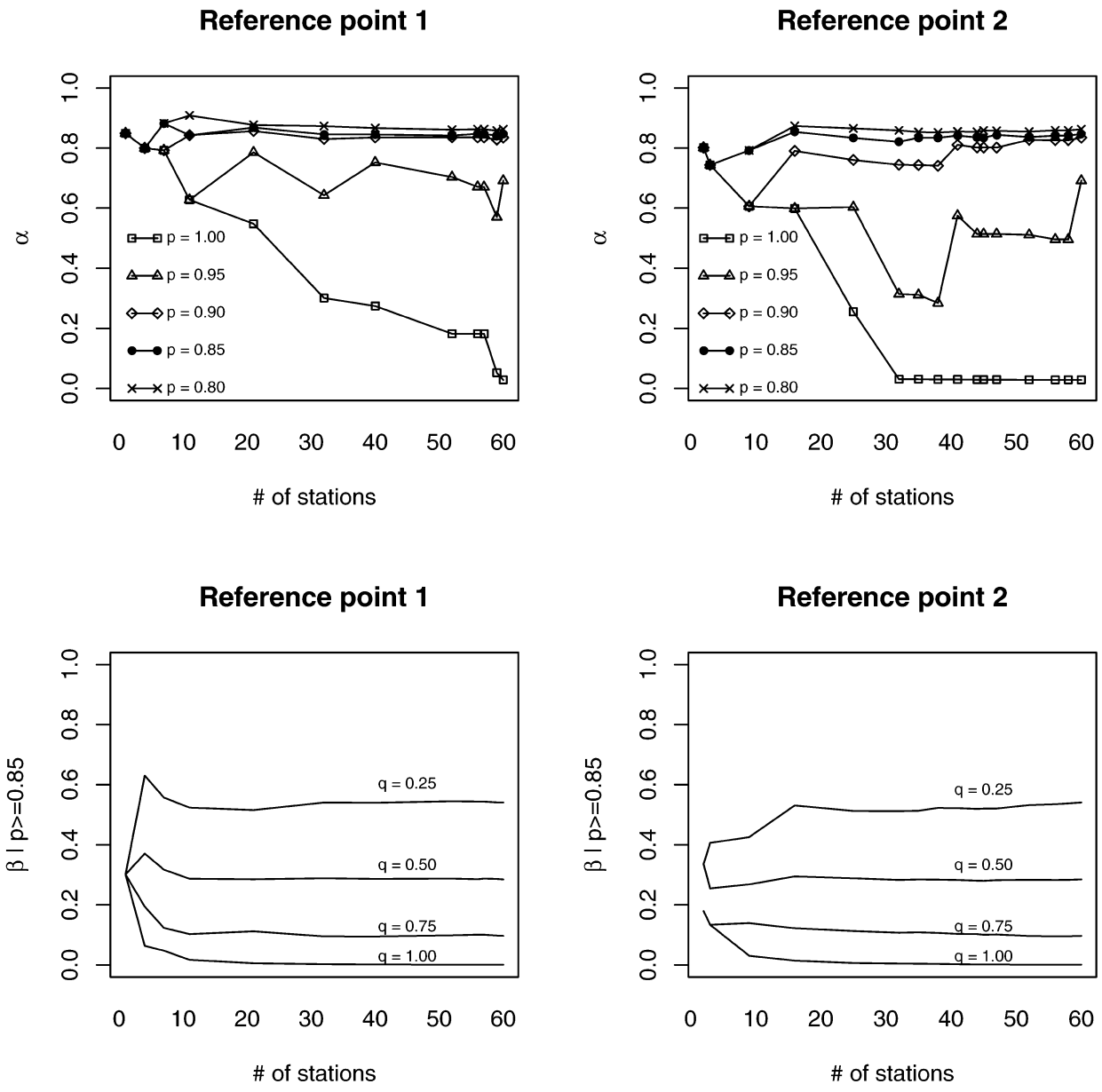

FIG. A1. Spatial and temporal thresholds for averaged daily station data.

er than 10. We thus restrict attention to days when 85% of stations have valid data.

After fixing the threshold $p \geq 0.85$, we compute the fraction of days (β) when there is precipitation in a certain fraction (q) of the stations averaged. This is plotted in the two bottom panels of Fig. A1. We plotted $\beta | p \geq 0.85$ for four q values 1.0, 0.75, 0.50, and 0.25. These plots can be interpreted as follows: $q = 1.0$ is the fraction of time when there is precipitation at all of the stations averaged. When more stations are averaged (e.g., >10), this situation is less common; $q = 0.50$, means that at least 50% of the stations with valid data are wet, and in this case, this occurs for about 30% of the days in the time series (~2520 days).

This analysis helps to derive a meaningful way to spatially average daily precipitation values. For each of the points 1 and 2, we used the thresholds of p and q to be 0.85 and 0.5, respectively. The β fraction is important for developing the spatially averaged precipitation time series. If we simply averaged all of the stations with valid data, we would include situations where there is precipitation at say one or two stations. These situations do not reflect precipitation occurrence and intensity for large regional averages. We thus select an appropriate β value (in this case 0.5) and derive a new time series by simply averaging respective station precipitation values when $p \geq 0.85$ and $q \geq 0.50$, else precipitation for that day was set

to zero. If $p < 0.85$, the data for that day are set to missing.

REFERENCES

Agterberg, F. P., 1989: Logdia–Fortran 77 program for logistic regression with diagnostics. *Comput. Geosci.,* **15,** 599–614.

Anderson, E., 2004: Calibration system Mean Areal Temperature (MAT) computational procedure. Chapter II.7, NWSRFS User's Manual, NOAA/NWS, 9 pp.

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

Biau, G., E. Zorita, H. von Storch, and H. Wackernagel, 1999: Estimation of precipitation by kriging in EOF space. *J. Climate,* **12,** 1070–1085.

Brandsma, T., and T. A. Buishand, 1997: Statistical linkage of daily precipitation in Switzerland to atmospheric circulation and temperature. *J. Hydrol.,* **198,** 98–123.

Breimann, L., J. H. Friedmann, R. A. Olsen, and J. C. Stone, 1984: *Classification and Regression Trees.* Wadsworth, 391 pp.

Buishand, T. A., and T. Brandsma, 2001: Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resour. Res.,* **37,** 2761–2776.

Clark, M. P., and L. E. Hay, 2004: Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *J. Hydrometeor.,* **5,** 15–32.

——, S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake Shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.,* **5,** 243–262.

Clarke, R. T., 1984: Mathematical models in hydrology. FAO Irrigation and Drainage. Paper 19, Food and Agricultural Organization of the United Nations, Rome, Italy, 292 pp.

Conway, D., R. L. Wilby, and P. D. Jones, 1996: Precipitation and air flow indices over the British Isles. *Climate Res.,* **7,** 169–183.

Gangopadhyay, S., M. P. Clark, and B. Rajagopalan, 2002: Statistical downscaling: A comparison of multiple linear regression and k-nearest neighbor approaches. *Eos, Trans. Amer. Geophys. Union,* **83** (Fall Meeting Suppl.), Abstract H12G-02.

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.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

Hewitson, B. C., and R. G. Crane, 1992: Large-scale atmospheric controls on local precipitation in tropical Mexico. *Geophys. Res. Lett.,* **19,** 1835–1838.

——, and ——, 1996: Climate downscaling: Techniques and application. *Climate Res.,* **7,** 85–95.

Katz, R. W., and M. B. Parlange, 1995: Generalizations of chain dependent processes: Application to hourly precipitation. *Water Resour. Res.,* **31,** 1331–1341.

——, and ——, 1996: Mixtures of stochastic processes: Applications to statistical downscaling. *Climate Res.,* **7,** 185–193.

Larson, L., 2004: Calibration system Mean Areal Precipitation (MAP) computational procedure. Chapter II.6, NWSRFS User's Manual, NOAA/NWS, 17 pp.

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.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, 1992: *Numerical Recipes in Fortan.* Cambridge University Press, 963 pp.

Rajagopalan, B., and U. Lall, 1999: A k-nearest neighbor simulator for daily precipitation and other variables. *Water Resour. Res.,* **35,** 3089–3101.

Schnur, R., and D. P. Lettenmaier, 1998: A case study of statistical downscaling in Australia using weather classification by recursive portioning. *J. Hydrol.,* **213,** 362–379.

Snedecor, G. W., and W. G. Cochran, 1989: *Statistical Methods.* The Iowa State University Press, 524 pp.

von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research.* Cambridge University Press, 494 pp.

Wilby, R. L., 1998: Statistical downscaling of daily precipitation using daily airflow and seasonal teleconnection indices. *Climate Res.,* **10,** 163–178.

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

——, 1999: Multisite downscaling of daily precipitation with a stochastic weather generator. *Climate Res.,* **11,** 125–136.

Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek, 2003: A technique for generating regional climate scenarios using a nearest neighbor algorithm. *Water Resour. Res.,* **39,** 1199, doi:10.1029/2002WR001769.

Zorita, E., and H. von Storch, 1999: The analog method as a simple statistical downscaling technique: Comparison with more complicated methods. *J. Climate,* **12,** 2474–2489.