

Statistical downscaling using K -nearest neighbors

Subhrendu Gangopadhyay¹ and Martyn Clark

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

Balaji Rajagopalan

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

Received 25 June 2004; revised 30 October 2004; accepted 30 November 2004; published 18 February 2005.

[1] Statistical downscaling provides a technique for deriving local-scale information of precipitation and temperature from numerical weather prediction model output. The K -nearest neighbor (K -nn) is a new analog-type approach that is used in this paper to downscale the National Centers for Environmental Prediction 1998 medium-range forecast model output. The K -nn algorithm queries days similar to a given feature vector in this archive and using empirical orthogonal function analysis identifies a subset of days (K) similar to the feature day. These K days are then weighted using a bisquare weight function and randomly sampled to generate ensembles. A set of 15 medium-range forecast runs was used, and seven ensemble members were generated from each run. The ensemble of 105 members was then used to select the local-scale precipitation and temperature values in four diverse basins across the contiguous United States. These downscaled precipitation and temperature estimates were subsequently analyzed to test the performance of this downscaling approach. The downscaled ensembles were evaluated in terms of bias, the ranked probability skill score as a measure of forecast skill, spatial covariability between stations, temporal persistence, consistency between variables, and conditional bias and to develop spread-skill relationships. Though this approach does not explicitly model the space-time variability of the weather fields at each individual station, the above statistics were extremely well captured. The K -nn method was also compared with a multiple-linear-regression-based downscaling model.

Citation: Gangopadhyay, S., M. Clark, and B. Rajagopalan (2005), Statistical downscaling using K -nearest neighbors, *Water Resour. Res.*, 41, W02024, doi:10.1029/2004WR003444.

1. Introduction

[2] Statistical downscaling provides a way to utilize output of climate models for local-scale applications. Typical grid size for global-scale simulations are of the order of 100–200 km, and the raw global-scale model output is of limited use when information is required at local scales. The objective of downscaling is to overcome this scale mismatch and to use the skill in atmospheric forecasts at local scales.

[3] In short, statistical downscaling develops relationships between large-scale atmospheric circulation variables and local climate information (e.g., precipitation and temperature observations at individual stations). Using these observed relationships, forecasts of atmospheric variables can be translated into forecasts of local climate variables. Several methods of varying complexity have been used in performing statistical downscaling. *Zorita and von Storch* [1998] have classified existing statistical methods into three categories: (1) linear methods (e.g., canonical correlation analysis), (2) classification methods (e.g., weather generators and regression tree), and (3) deterministic nonlinear

methods (e.g., neural networks). They also propose an analog method and compare the results with a method chosen from each of the above three categories to reconstruct average December–February (DJF) precipitation over the Iberian Peninsula for the period 1901–1989.

[4] *Widmann et al.* [2003] applied three different statistical downscaling methods that used simulated precipitation fields from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis [*Kalnay et al.*, 1996] as the predictor. These methods are (1) local rescaling of the simulated precipitation, (2) downscaling using singular value decomposition (SVD), and (3) local rescaling with a dynamical correction. The three methods were applied to reconstruct historical (1958–1994) wintertime precipitation over Oregon and Washington and concluded that local rescaling with dynamical correction and SVD-based downscaling yielded comparable skills over the Pacific Northwest region. *Salathé* [2003] forced a hydrologic model of the Yakima River in central Washington with three downscaled precipitation fields to compare the effectiveness of the downscaling methods. One of these methods was an analog method that used a 1000-hPa geopotential height field from the NCEP–NCAR reanalysis as a predictor. *Salathé* [2003] showed that downscaling by local scaling of simulated large-scale precipitation from the NCEP–NCAR model was quite successful in streamflow simulations in the Yakima basin.

¹Now at Hydrosphere Resource Consultants, Boulder, Colorado, USA.

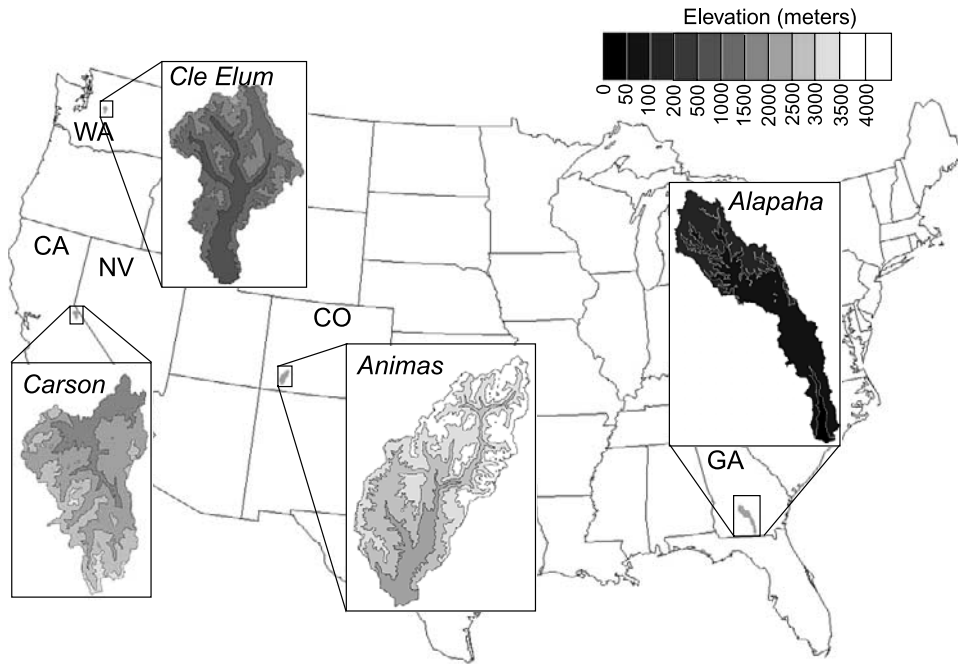


Figure 1. Location and topography of the study basins.

[5] In this paper we present a downscaling methodology based on the K -nearest neighbor (K -nn) algorithm. The K -nn algorithm is described for use in a stochastic weather generator by Lall and Sharma [1996], Rajagopalan and Lall [1999], Buishand and Brandsma [2001], and Yates *et al.* [2003]. The fundamental idea of the K -nn algorithm is to search for analogs of a feature vector (vector of variables for which analogs are sought) based on similarity criteria in the observed time series. In the weather generator model, the day immediately following the analog day is taken as the next day in the generated sequence, and the process is repeated. In the method presented here, local-scale station information is used for analog days selected on the basis of global-scale climate model output.

[6] Though transfer-function-based models (e.g., multiple linear regression, or MLR) are widely in use [Antolik, 2000], the K -nn based approach developed here has several advantages. First, this method is data-driven and makes no assumptions of the underlying marginal and joint probability distributions of variables. For example, to downscale precipitation using MLR, we need a two-step process [e.g., Clark *et al.*, 2004]. We need to account for the intermittent property of precipitation (typically modeled using a logistic regression), and then transform to normal space to satisfy the inherent normality criteria needed in least squares regression to model precipitation amounts. Second, K -nn based downscaling will be shown to intrinsically preserve the spatial covariability and consistency of the downscaled climate fields. Third, ensemble medium-range forecast (MRF) runs can be readily utilized in the downscaling process, and there is no need to use the ensemble mean of MRF predictors, as is normally used in regression models. Finally, the ensemble spread information from MRF runs can be utilized to develop spread-skill relationships, which is not possible in a MLR model [e.g., Clark *et al.*, 2004].

[7] The K -nn downscaling methodology was tested on four example river basins distributed over the continental United States, covering both snowmelt- and rainfall-dominated hydrologic regimes. These four basins are (1) the Animas River in southwestern Colorado, (2) the east fork of the Carson River on the California/Nevada border, (3) the Cle Elum River in central Washington, and (4) the Alapaha River in southern Georgia (Figure 1).

[8] Section 2 provides a description of the data used in the analysis. Section 3 describes the K -nn methodology developed for statistical downscaling. Section 4 presents a discussion of the results from the four example river basins. Section 5 is a summary of the techniques and results.

2. Data Description

2.1. The CDC Forecast Archive

[9] The NOAA-CIRES Climate Diagnostics Center (CDC) has generated a “reforecast” data set (1978 to present) using a fixed version (circa 1998) of the NCEP operational medium-range forecast (MRF) model [Hamill *et al.*, 2004]. This is a spectral model and has a horizontal resolution of approximately 200 km, with 28 vertical layers (T62/L28). The archive consists of one control run plus 14 ensemble members, a total of 15 members. The control run is based on the global analysis from the NCEP-NCAR reanalysis project [Kalnay *et al.*, 1996]. Initial perturbations for ensemble members are generated from the control run with the “breeding method” [Toth and Kalnay, 1993]. Each ensemble member consists of a 14-day forecast starting every day since 1 January 1978, and presently the model continues to be run in real time. The model outputs are saved at 0000 UT and 1200 UT. The 20-year archive data from 1 January 1979 to 31 December 1998 was used in this study.

[10] We used seven output variables [Clark and Hay, 2004] from each of the ensemble members in our analysis.

The model output variables used are (1) the accumulated precipitation for a 12-hour period (e.g., 0000–1200 UT) at the surface, (2) mean sea level pressure, (3) total column precipitable water, (4) relative humidity at 700 hPa, (5) 2-m air temperature, (6) 10-m zonal wind speed, and (7) 10-m meridional wind speed.

2.2. Station Data

[11] This study employs daily precipitation, and maximum and minimum temperature data from the National Weather Service (NWS) manual cooperative (COOP) network of climate observing stations across the contiguous United States. These data were extracted from the National Climatic Data Center (NCDC) Summary of the Day (TD3200) data set [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 (1) absurdly extreme values, (2) internal consistency among variables (e.g., maximum temperature less than minimum temperature), (3) constant temperature (e.g., 5 or more days with the same temperature are suspect), (4) excessive diurnal temperature range, (5) invalid relationships between precipitation, snowfall, and snow depth, and (6) unusual spikes in temperature time series. Records at most of these stations start in 1948 and continue through 1998.

[12] The four example basins (the Animas River, Colorado, referred to in the figures as anmas; the East Carson River, California and Nevada, carsn; the Cle Elum River, Washington, celum; and the Alapaha River, Georgia, alapa) were selected based on their geographical distribution and streamflow characteristics. The Animas, East Carson, and Cle Elum are snowmelt-dominated, and the Alapaha is a rainfall-dominated basin. We select the “best stations” in the COOP network that are located within a 100-km search radius of the center of these four basins: 15 stations in the Animas, 16 stations in the Carson, 18 stations in the Cle Elum, and 10 stations in the Alapaha (Table 1). These “best stations” are defined as those with less than 10% missing or questionable data over the analysis period, 1979–1998.

3. Methodology

[13] The steps in downscaling the atmospheric variables to basin-scale precipitation and temperature using the K -nn algorithm are outlined in this section. The CDC NCEP-MRF forecast archive was retrieved and formatted to form a data matrix consisting of 7305 rows (corresponding to the number of days from 1 January 1979 to 31 December 1998) and 14 columns (corresponding to the number of lead times) for each of the seven variables (see section 2.1). Days similar to each of the 7305×14 days in the archive were identified using the K -nn algorithm. A description of the K -nn algorithm follows.

3.1. K -nn Algorithm

[14] Each of the 15 ensemble members of the MRF archive for each basin was examined individually. The steps of the K -nn algorithm for a given MRF ensemble member and basin are as follows:

[15] 1. Compile a feature vector of MRF model output for a given day and forecast lead time. The feature vector (\vec{F}_f) consists of values for all the climate variables of the day (the feature day, f) for which we are trying to find the K -nearest

Table 1. Stations Used in the Four Study Basins

Number	Latitude, deg	Longitude, deg	Station ID
<i>Animas (Colorado), 37.50°N, 107.50°W</i>			
1	38.40	107.52	CO1609.COOP
2	37.23	108.05	CO3016.COOP
3	37.77	107.13	CO3951.COOP
4	38.03	107.32	CO4734.COOP
5	37.20	108.49	CO5531.COOP
6	38.13	108.29	CO6012.COOP
7	38.02	107.67	CO6203.COOP
8	37.24	107.02	CO6258.COOP
9	37.71	108.04	CO7017.COOP
10	37.73	107.27	CO7050.COOP
11	37.95	107.87	CO8204.COOP
12	37.38	107.58	CO8582.COOP
13	36.83	108.00	NM0692.COOP
14	36.94	107.00	NM2608.COOP
15	36.82	107.62	NM6061.COOP
<i>Alapaha (Georgia), 31.35°N, 83.22°W</i>			
1	31.58	84.17	GA0140.COOP
2	31.53	82.52	GA0211.COOP
3	31.18	84.20	GA1500.COOP
4	31.97	83.78	GA2266.COOP
5	31.52	82.85	GA2783.COOP
6	32.20	83.21	GA2966.COOP
7	31.72	83.25	GA3386.COOP
8	31.03	82.80	GA4429.COOP
9	31.17	83.75	GA6087.COOP
10	31.48	83.53	GA8703.COOP
<i>Cle Elum (Washington), 47.37°N, 121.05°W</i>			
1	47.77	121.48	WA0456.COOP
2	47.17	122.00	WA0945.COOP
3	47.42	121.73	WA1233.COOP
4	47.84	120.04	WA1350.COOP
5	47.18	120.92	WA1504.COOP
6	47.00	120.52	WA2505.COOP
7	47.38	121.97	WA4486.COOP
8	47.13	122.27	WA5224.COOP
9	47.85	121.98	WA5525.COOP
10	47.15	121.93	WA5704.COOP
11	47.30	121.85	WA6295.COOP
12	47.18	119.87	WA6880.COOP
13	47.45	122.30	WA7473.COOP
14	47.54	121.84	WA7773.COOP
15	47.87	121.72	WA8034.COOP
16	47.43	120.31	WA9074.COOP
17	47.40	120.21	WA9082.COOP
18	46.57	120.54	WA9465.COOP
<i>Carson (California-Nevada), 38.55°N, 119.80°W</i>			
1	39.38	120.10	CA0931.COOP
2	38.25	119.23	CA1072.COOP
3	38.28	120.32	CA1277.COOP
4	38.25	120.86	CA1428.COOP
5	37.97	119.92	CA1697.COOP
6	39.32	120.23	CA2467.COOP
7	39.17	120.13	CA8758.COOP
8	38.45	120.50	CA8928.COOP
9	39.33	120.18	CA9043.COOP
10	38.70	120.03	CA9105.COOP
11	37.76	119.59	CA9855.COOP
12	39.15	119.77	NV1485.COOP
13	39.08	119.95	NV3205.COOP
14	39.00	119.75	NV5191.COOP
15	39.08	119.12	NV8822.COOP
16	39.00	119.17	NV9229.COOP

neighbors. Since two model outputs, 0000 and 1200 UT, were available for each of the seven variables, the feature vector \vec{F}_f was assumed to consist of 14 variables.

$$\vec{F}_f = [v_1^1 \ v_2^1 \ \dots \ v_7^1 \ v_1^2 \ v_2^2 \ \dots \ v_7^2] \quad (1a)$$

or

$$\vec{F}_f = [x_1 \ x_2 \ \dots \ x_{14}], \quad (1b)$$

where v_i^j is the value of the climate variable i ($i = 1, \dots, 7$; the seven climate variables, see section 2.1) at time j ($j = 1, 2$; 0000 and 1200 UT) for the feature day f . Explicitly, $x_1 = v_1^1$; $x_2 = v_2^1$, and so on.

[16] 2. Set a window of chosen width centered on the feature day f . We used a 14-day window (7 days lagged and 7 days lead) [Yates et al., 2003] starting with the first day of the archive (1 January 1979). The subset of data for a given variable now consists of all days over the 20-year period (1979–1998) but excluding day f within this 14-day window. Missing data, if any, also need to be accounted for, and let nt be the total number of days with available data within the temporal window centered on the feature day f . So for the 14 variables (refer to step 1), the data matrix was reformatted to have nt rows and 14 columns. The structure of this data matrix ($[A]_{nt \times 14}^f$) is

$$[A]_{nt \times 14}^f = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,14} \\ a_{2,1} & a_{2,2} & \dots & a_{2,14} \\ \dots & \dots & \dots & \dots \\ a_{nt,1} & a_{nt,2} & \dots & a_{nt,14} \end{bmatrix}, \quad (2)$$

where $a_{i,j}$ is the value of the climate variable for time index i ($i = 1, \dots, nt$) and for variable j ($j = 1, \dots, 14$).

[17] 3. Standardize matrix $[A]_{nt \times 14}^f$. The standardized matrix $[S]_{nt \times 14}^f$ is expressed as

$$[S]_{nt \times 14}^f = [\underline{s}_1 \ \underline{s}_2 \ \dots \ \underline{s}_{14}] \quad (3a)$$

$$\underline{s}_j = \frac{1}{\sigma_j} (a_j - \mu_j) \quad (3b)$$

$$\underline{a}_j = [a_{1,j} \ a_{2,j} \ \dots \ a_{nt,j}]^T \quad (3c)$$

$$\mu_j = E[\underline{a}_j] \quad (3d)$$

$$\sigma_j = \left(E[\{\underline{a}_j\}^2] - \{\mu_j\}^2 \right)^{1/2}, \quad (3e)$$

where the underbars represent vectors; \underline{s}_j represents the vector of standardized values of vector \underline{a}_j for variable j . The variable counter j loops from 1 through 14 (the total number of variables); μ_j and σ_j are the mean and standard deviation,

respectively, of variable j estimated from vector \underline{a}_j ; $E[\]$ is the expected value; and superscript T represents the vector-matrix transpose operator.

[18] 4. Perform empirical orthogonal function (EOF) decomposition or principal component analysis (PCA) of matrix $[S]_{nt \times 14}^f$. We first estimate the correlation/covariance matrix $[C]_{14 \times 14}^f$, which is given by

$$[C]_{14 \times 14}^f = \frac{1}{(nt - 1)} [S]^T [S], \quad (4)$$

where $[S]^T$ is the transpose of matrix $[S]$ (the superscript f has been dropped for clarity; see equation (3a)). A singular value decomposition of $[C]_{14 \times 14}^f$ [Press et al., 1992] yields

$$[C]_{14 \times 14}^f = [U][W][V]^T, \quad (5)$$

where $[U]$ and $[V]$ are the orthogonal matrices (order, 14×14), and $[W]$ is a diagonal matrix of the same order whose elements are the eigenvalues ($\lambda_j, j = 1, \dots, 14$ such that $\lambda_1 > \lambda_2 > \dots > \lambda_{14}$; corresponding to the 14 variables). Since $[C]_{14 \times 14}^f$ is symmetric, $[U] = [V]$. Each column of $[U]$ (or $[V]$) represents the eigenvectors corresponding to a given eigenvalue λ_j . Let \underline{u}_j be the eigenvector corresponding to eigenvalue λ_j . So

$$[U]_{14 \times 14} = [\underline{u}_1 \ \underline{u}_2 \ \dots \ \underline{u}_{14}]. \quad (6)$$

The principal components (PCs) are then derived as

$$[P]_{nt \times 14}^f = [S]_{nt \times 14}^f [U]_{14 \times 14} \quad (7a)$$

or

$$[P]_{nt \times 14}^f = [\underline{p}_1 \ \underline{p}_2 \ \dots \ \underline{p}_{14}] \quad (7b)$$

and,

$$\underline{p}_j = \begin{bmatrix} p_{1,j} \\ p_{2,j} \\ \vdots \\ p_{nt,j} \end{bmatrix} \quad (7c)$$

where $[P]_{nt \times 14}^f$ is the principal component matrix for feature day f and column vector \underline{p}_j is the j th principal component ($j = 1, \dots, 14$) of length nt . The principal components that explained more than 1 percent of the total variance (total variance is given by the trace of matrix $[W]$, i.e., $tr[W]$) for feature day f were retained. Let $nret$ be the number of PCs retained, and $nret < 14$. Typically five PCs were retained.

[19] 5. Using summary statistics (mean and standard deviation, equations (3d) and (3e), respectively) from step 3, and eigenvectors from step 4, project the feature vector \vec{F}_f in step 1 on to eigenspace. Let the projected feature vector be \vec{F}'_f , which is given by

$$\vec{F}'_f = \left[\frac{(x_1 - \mu_1)}{\sigma_1} \ \frac{(x_2 - \mu_2)}{\sigma_2} \ \dots \ \frac{(x_{14} - \mu_{14})}{\sigma_{14}} \right]_{1 \times 14} [U]_{14 \times 14} \quad (8a)$$

or

$$\vec{F}'_f = [x'_1 \ x'_2 \ \dots \ x'_{14}], \quad (8b)$$

where x'_j are the elements of the projected feature vector \vec{F}'_f .

[20] 6. For each time element i ($i = 1, \dots, nt$), compute the weighted Euclidian distance between the projected feature vector (equation (8b)) and the PCs (equation (7b)). The distance computation is carried out using only the $nret$ components. Let d_i be the distance metric corresponding to day i , which is calculated as

$$d_i = \left[\sum_{j=1}^{nret} \frac{\lambda_j}{tr[W]} (x'_j - p_{ij})^2 \right]^{1/2}. \quad (9)$$

The ratio $\lambda_j/tr[W]$ is the weight and corresponds to the fraction of variance explained by PC p_j . This gives a set of nt distances as possible neighbors of feature day f .

[21] 7. Sort the distances d_i in ascending order ($d_{(i)}$), and retain only the first K neighbors. The choice of K is based on the prescriptive choice of the square root of all possible candidates (i.e., $K = \sqrt{nt}$) [Rajagopalan and Lall, 1999; Yates et al., 2003]. From the asymptotic arguments of Fukunaga [1990], K should be chosen so as to be proportional to $nt^{4/(d+4)}$, where d is the dimension of the vector (i.e., in our case, number of PCs retained, $nret$) for which the nearest-neighbor density is to be estimated, with the constant of proportionality dependent on the underlying density. One can also use objective criteria such as generalized cross validation (GCV) as proposed by Rajagopalan and Lall [1999] or Lall and Sharma [1996]. For the sample sizes under consideration here, the choice of $K = \sqrt{nt}$ was found to give consistent results for the simulated statistics. With a 14-day window, 20 years of data, and no missing data for days in the temporal window, the maximum number of nearest neighbors in our case was $\sqrt{[(14 + 1) \times 20] - 1} = 17$ (rounded to nearest integer).

[22] 8. Assign weight w_i ($0 < w_i < 1$) to each of the K neighbors using the bisquare weight function [Huber, 2003] based on distance $d_{(i)}$.

$$w_i = \frac{\left[1 - \left(\frac{d_{(i)}}{d_{(K)}} \right)^2 \right]^2}{\sum_{i=1}^K \left[1 - \left(\frac{d_{(i)}}{d_{(K)}} \right)^2 \right]^2}, \quad (10)$$

where $d_{(K)}$ is the distance (sorted) of neighbor K .

[23] 9. Select a neighbor from the K neighbors as an analog for feature day f . A uniform random number [Press et al., 1992], $u \sim U[0, 1]$, is first generated, and if $u \geq w_1$, then the day corresponding to distance $d_{(1)}$ is selected. If $u \leq w_K$, then the day corresponding to $d_{(K)}$ is selected. For $w_1 < u < w_K$, the day corresponding to $d_{(i)}$ is selected for which u is closer to w_i .

[24] 10. Repeat step 9 seven times to generate seven ensemble members.

[25] 11. Repeat steps 1–10 for each of the days (7305) corresponding to a forecast lead time (14 lead times), a total of (7305×14) feature days in the archive.

[26] 12. Repeat steps 1–11 fifteen times corresponding to the 15 MRF runs.

[27] 13. Repeat steps 1–12 four times for the four study basins.

[28] Thus the final output for each of the four basins consisted of analog dates (pointers to physical dates were stored) corresponding to each day in the MRF archive (size, 7305), each forecast lead time (size, 14), and an ensemble of 105 ensemble members (seven realizations from each of the 15 MRF model runs). Note that the downscaling was carried out for the center point of each of the four basins.

[29] The premise of this approach is that the atmospheric model captures the large-scale circulation patterns, which are assumed to be related to the synoptic-scale processes that generate the local weather. Since neighbors are sought only in the PC space of the atmospheric model, the choice of the downscaling location is not influenced by any basin feature (e.g., topography, precipitation shadow, etc.). Our choice of the downscaling location (i.e., centers of basins) was guided primarily by the intended application to hydrologic modeling by River Forecast Centers. Forecasting of precipitation and temperature fields at individual stations adjoining the basins is described in the next section.

3.2. Forecasting Precipitation and Temperature Fields at Individual Stations

[30] A 100-km search radius was used from the center of each basin to pick up the closest stations (see Table 1). The dates derived using the K -nn algorithm for a given basin were used to select from the daily-observed precipitation and temperature values for each of the adjoining stations of that basin. This then constitutes the downscaled precipitation and temperature for each of the stations used in this study. Several statistics were then calculated to analyze these downscaled precipitation and temperature fields, and these are presented in the next section.

4. Results and Discussions

[31] The statistics used to analyze and verify the downscaled precipitation and temperature forecasts are (1) seasonal cycles of precipitation amount and temperature (results are shown only for maximum temperature), (2) bias, (3) spatial correlations, (4) forecast skill, (5) forecast reliability, (6) rank histograms, and (7) spread-skill relationships.

4.1. Seasonal Cycles of Precipitation Amount and Temperature

[32] We first analyzed the variation of the annual cycle of precipitation and temperature for the four study basins. In Figures 2 and 3 the annual cycles (derived from observations for the period 1979–1998) of precipitation and temperature, respectively, for selected COOP stations in the basins along with the ensemble spread (as box plots) for each month are presented. The COOP stations used were CO1609, GA0140, WA0456, and CA0931 for the Animas, Alapaha, Cle Elum, and East Carson, respectively (see Table 1 for locations). The box plots for each month were estimated from the 105 ensemble members and are shown for the forecast lead time of 5 days. The boxes in these plots (e.g., Figure 2) indicate the interquartile range of the simulations, and the whiskers show the 5th and 95th percentile of the simulations, while the open circles indicate values outside this range. The horizontal lines within the

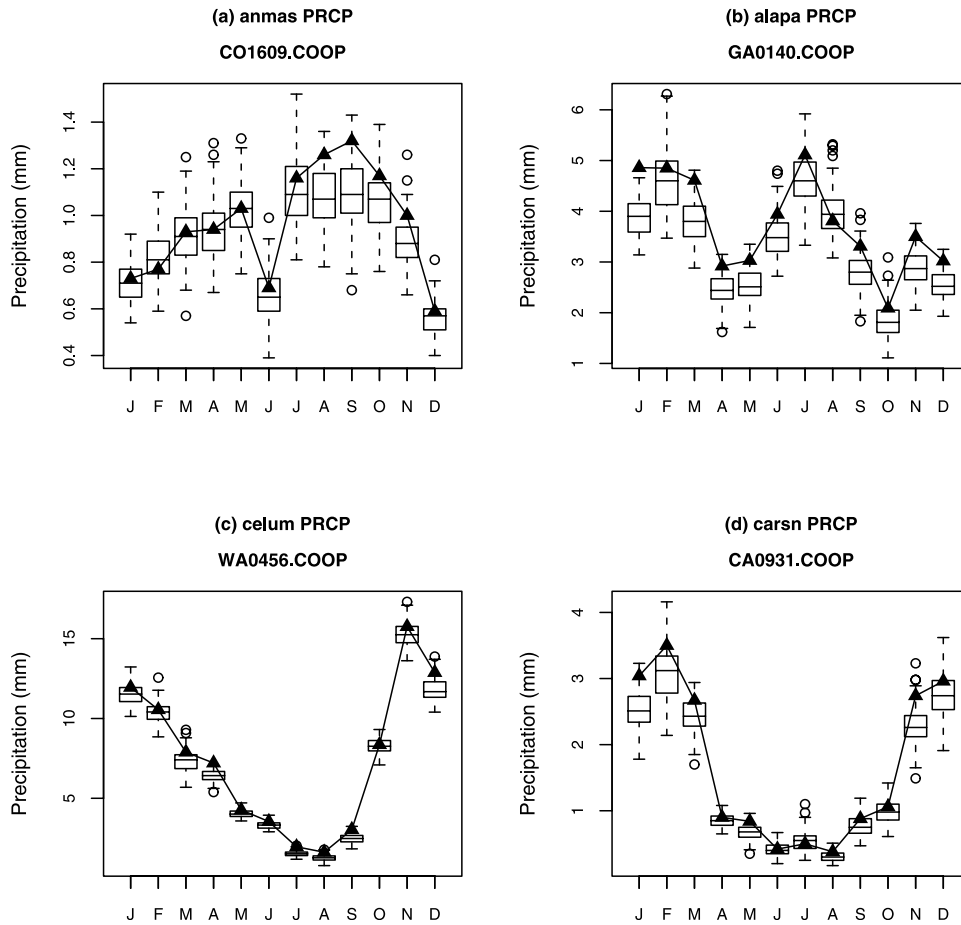


Figure 2. Box plots of total monthly precipitation from the 105 ensemble members for selected stations in the four study basins: (a) CO1609, (b) GA0140, (c) WA0456, and (d) CA0931. Results are shown for lead time 5 days. The solid line and marks are the same statistics derived from the historical data for the period 1979–1998.

box indicate the median value, and the solid lines join values of the statistic from the observed data. Typically, if the statistics of the observed data fall within the box, it indicates that the simulations adequately reproduced the statistics of the historical data.

[33] In case of precipitation (Figure 2), there is a wide regional variation in the amounts and timing of the maximum precipitation occurrences among the basins. The Alapaha, for example, has a precipitation peak in summer, but the Cle Elum is the driest during the summer season (June–July–August). Also, it is well known that in the western United States, in particular during wintertime, the hydroclimate variables have a coherent spatial pattern [e.g., *Rajagopalan and Lall, 1998*]. The atmospheric models generally represent these synoptic scales quite well, and we see better simulations of precipitation amount in the snow-dominated Animas, Cle Elum, and East Carson basins (Figures 2a, 2c, and 2d, respectively) over the rainfall-dominated Alapaha basin (Figure 2b). The K -nn downscaling model in all cases largely captures the seasonal variation of precipitation. Given that the K -nn algorithm was not explicitly designed to preserve monthly statistics, the seasonal cycle is fairly well captured. For maximum temperature (Figure 3), the downscaled values in all cases were able to capture the historical observations. Unlike precipitation, the ensemble spread (interquartile range

in the box plots) was minimal in case of temperature. Similar results were noted for other forecast lead times.

4.2. Bias

[34] Bias is defined as the deviation of the expected value of a given variable from its true value. We estimated the median absolute bias (MAB_l) for each forecast lead time (l) and month as the following:

$$\bar{O}_l = \frac{1}{ndays} \sum_{i=1}^{ndays} O_i^l \quad (11a)$$

$$\bar{Y}_l^e = \frac{1}{ndays} \sum_{i=1}^{ndays} (Y_i^l)^e \quad (11b)$$

$$MAB_l = \text{Median} [|\bar{O}_l - \bar{Y}_l^e|; e = 1, \dots, n], \quad (11c)$$

where, $ndays$ is the total number of days in the time series for a given month (e.g., $ndays = 620$ for January from 20 years of data and with no missing values); \bar{O}_l is the expected value of the observed variable (precipitation or temperature) for lead time l (i.e., climatological mean); and O_i^l is the observation for day i and lead time l . Similarly, \bar{Y}_l^e is the expected value of the downscaled variable for lead

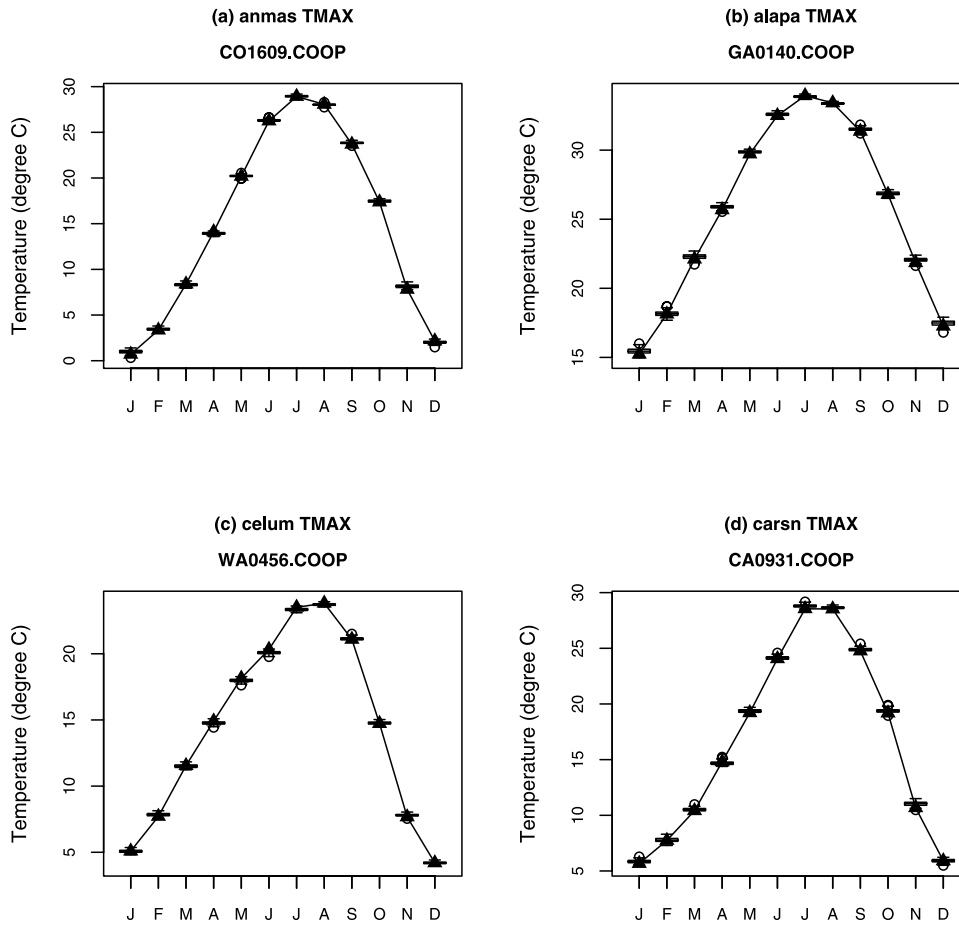


Figure 3. Same as Figure 2 but for temperature.

time l and ensemble member e , and $(Y_i^l)^e$ is the downscaled variable value for day i , lead time l , and ensemble member e . We then calculate the absolute bias for a given ensemble member and use the n (equal to 105) ensemble members to calculate the median absolute bias (MAB_l) for lead time l (equation (11c)). For precipitation, the absolute bias was expressed as a percentage of \bar{O}_l . In other words, the absolute difference term within the square brackets of equation (11c) was expressed as $|\bar{O}_l - \bar{Y}_l^e| \times 100/\bar{O}_l$.

[35] Figure 4 shows the bias for precipitation for each of the four basins for the month of January. Once again, these biases are median absolute biases and are expressed as a percentage of the mean climatology. The box plots correspond to the spread from the number of closest stations (shown in parenthesis) in a given basin. The median bias (estimated from the closest stations for a given basin) for all the basins is within 20%. In some cases, stations have biases greater than 20%. Of the four basins, the biases are largest for the Animas. This is probably because the Animas is the driest of all the four basins with an average January precipitation of about 1.28 mm. The temperature biases (not shown) were quite small and typically were within 0.5°C.

4.3. Spatial Correlations

[36] Spatial autocorrelations are used to check how well the K -nn algorithm performs in preserving the spatial autocorrelation. The Pearson correlation (hereinafter re-

ferred to as correlation) between two example stations, say, 1 and 2, was estimated as follows:

[37] Let \underline{Y}_{1l}^e and \underline{Y}_{2l}^e be the vector of downscaled values for a given variable (e.g., precipitation) for lead time l from ensemble member e . That is,

$$\underline{Y}_{1l}^e = [(Y_{1l}^e)_1 (Y_{1l}^e)_2 \dots (Y_{1l}^e)_{ndays}]^T \quad (12a)$$

$$\underline{Y}_{2l}^e = [(Y_{2l}^e)_1 (Y_{2l}^e)_2 \dots (Y_{2l}^e)_{ndays}]^T, \quad (12b)$$

where $(Y_{1l}^e)_i$ and $(Y_{2l}^e)_i$ are the downscaled variable values for lead time l , ensemble member e , and day i for stations 1 and 2, respectively; and $i = 1, \dots, ndays$. Next we calculate the correlation (ρ_l^e) for a given ensemble member (e) and lead time (l) using the vectors \underline{Y}_{1l}^e and \underline{Y}_{2l}^e . That is,

$$\rho_l^e = \frac{E[\underline{Y}_{1l}^e \underline{Y}_{2l}^e] - E[\underline{Y}_{1l}^e] E[\underline{Y}_{2l}^e]}{\sigma_1 \sigma_2}, \quad (13)$$

where, $E[]$ is the expected value and σ_1 and σ_2 are the standard deviations of \underline{Y}_{1l}^e and \underline{Y}_{2l}^e , respectively.

[38] Figure 5 shows the correlation box plots (for a given l using the 105 ensemble members of ρ_l^e) over 14-day forecast lead time between two example stations in the Animas basin (CO4734 and CO1609), and Figure 6

PRCP January

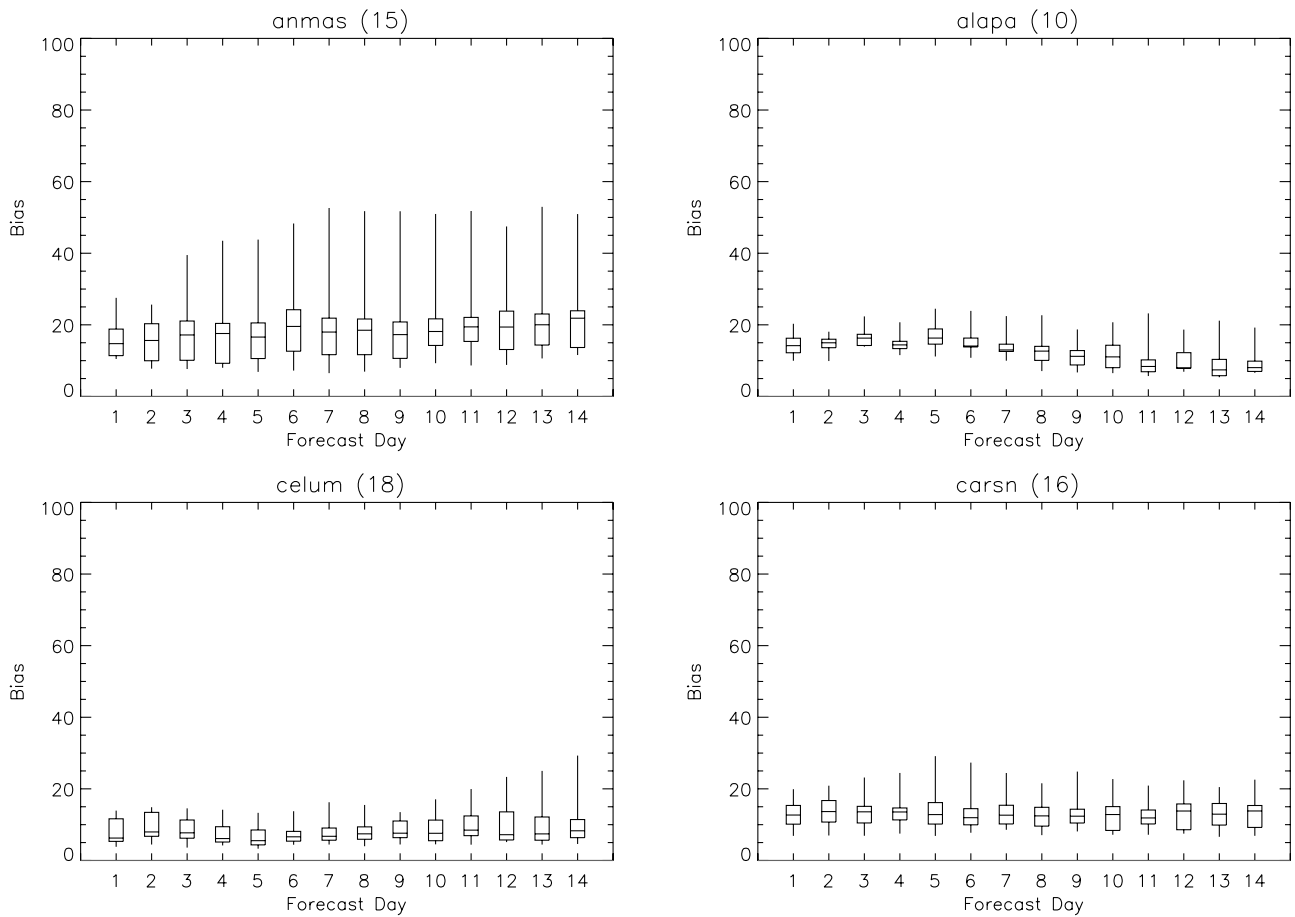


Figure 4. Box plots of median absolute bias (in percentage) for January precipitation for the 14-day forecast lead times in case of the four basins. The box plots are plotted using the number of stations shown in parenthesis following the basin names.

presents similar results for two stations in the Alapaha basin (GA0140 and GA2266) for winter and summer precipitation and temperature. Since we pick the weather data for all the stations simultaneously on the selected neighbor (i.e., day, step 9), the K -nn method intrinsically preserves (1) the spatial correlation structure of the variables (precipitation and temperature) and (2) the correlation between the variables at each station.

[39] For precipitation, in the case of the Animas basin, which overall is a dry basin, the observed spatial correlation is about 0.2 for both January and July. These observed spatial correlations are quite small. Since the Animas basin has significant topographical variations (see Figure 1), elevation differences and measurement errors in precipitation can contribute to low observed spatial correlation values. In the case of Alapaha, which is relatively flat, and wetter, we see a high degree of spatial correlation (about 0.7) between the example stations in January. In July, the spatial correlation diminishes because the precipitation is largely generated by convection, which is generally difficult to capture by the atmospheric models.

[40] For temperature (see Figures 5 and 6), the box plots of downscaled values adequately bracket the observed

spatial correlation. The temperature correlations among the stations are very similar for winter and summer in both the basins, and the biases are quite small in all cases. As mentioned above, the cross correlations (i.e., correlations between variables at each station) are also intrinsically preserved by this downscaling method.

4.4. Forecast Skill

[41] The probabilistic skill of the downscaled precipitation and temperature forecasts was assessed using the ranked probability skill score (RPSS) [Wilks, 1995]. The RPSS is based on the ranked probability score (RPS) computed for each downscaled forecast and observation pair:

$$RPS = \sum_{m=1}^J (Y_m - O_m)^2, \quad (14)$$

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 $10(J)$ possible categories for forecasts of precipitation and

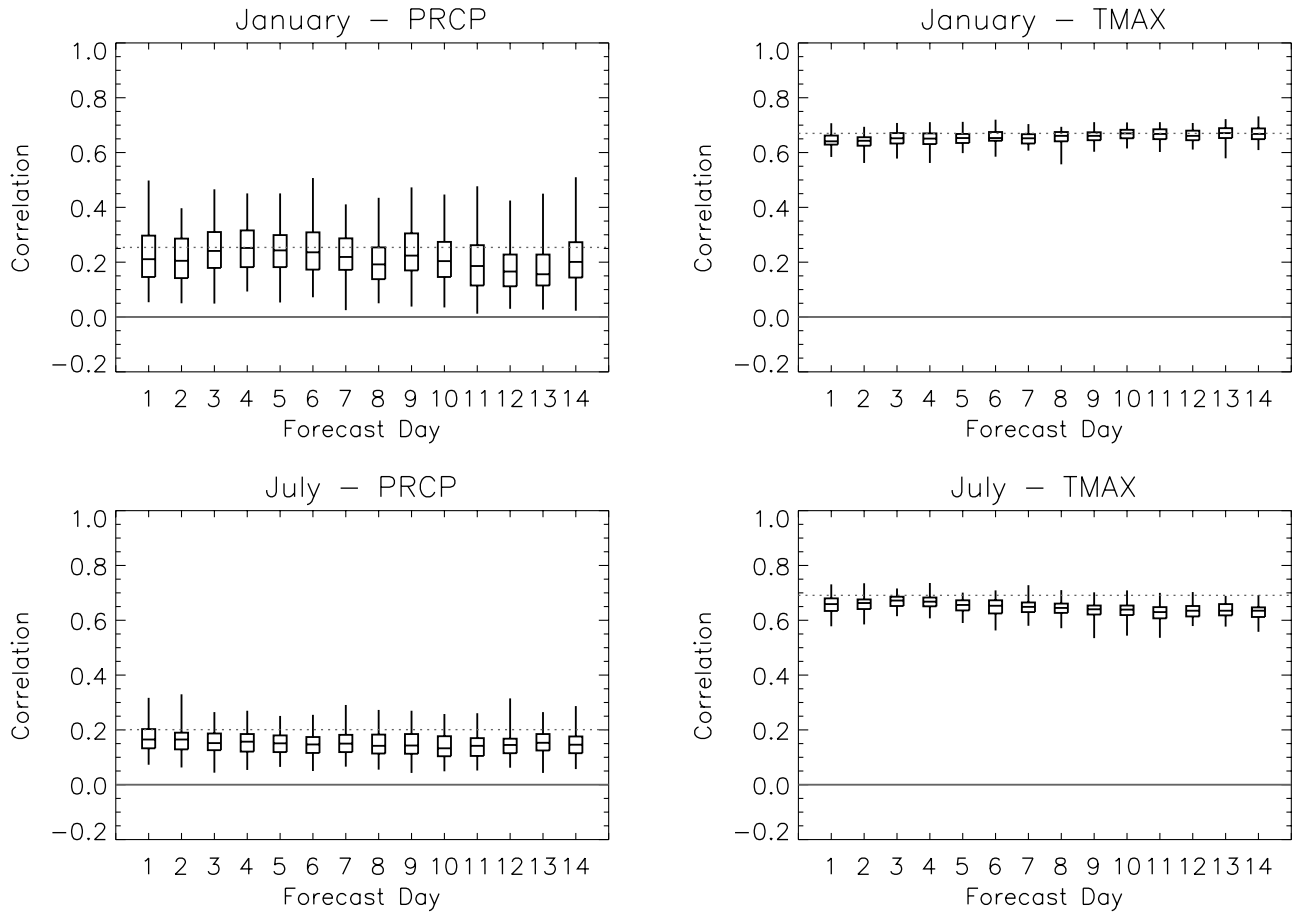


Figure 5. Box plots of spatial autocorrelation from the 105 ensemble members between stations CO1609 and CO4734 in the Animas basin for the 14-day forecast lead times. Precipitation correlations are in the left column for (top) January and (bottom) July, and temperature correlations are in the right column for (top) January and (bottom) July. The dotted horizontal line is the observed spatial correlation between these two stations derived from the historical data for the period 1979–1998.

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, variable, and stations in the basin. Next, for each forecast-observation pair, the number of ensemble members forecast in each category is determined (out of 105 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 less than the observation's position are assigned zero and all categories equal to and greater than 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 (equation (14)).

[42] The RPSS is then computed as

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}_{c_{lim}}}, \quad (15)$$

where \overline{RPS} is the mean ranked probability score for all forecast-observation pairs and $\overline{RPS}_{c_{lim}}$ is the mean ranked probability score for climatological forecast.

[43] For temperature, $\overline{RPS}_{c_{lim}}$ is computed using an equal probability in each of the m categories defined in equation (14) (i.e., $1/J$); for precipitation, the probability for the first category (zero precipitation) is taken as the observed probability of no precipitation, and the probability for all other categories is taken as $1/(J - 1)$ (see equation (14)). An RPSS of 0.0 indicates no difference in skill over the reference climatological forecast ($\overline{RPS}_{c_{lim}}$), and an RPSS of 1.0 indicates a perfect forecast. Negative RPSS implies that the model performs worse than climatology. Here RPSS was estimated separately for each forecast lead time, for each month, and for each station in the basin. The median RPSS was then calculated from the station RPSS values for each of the basins.

[44] Figures 7 and 8 show plots of median RPSS for precipitation and temperature, respectively. These plots show the months along the abscissa and forecast lead times along the ordinate, with darker shading representing regions of higher skill. For precipitation (Figure 7), in all the basins higher skills are obtained during the fall and winter months, and extend for only short forecast lead times (e.g., up to 3 days in the case of Cle Elum). Wintertime skill scores are around 0.4 for all of the basins. This means that the K -nn downscaled forecasts are superior 40% of the time over the

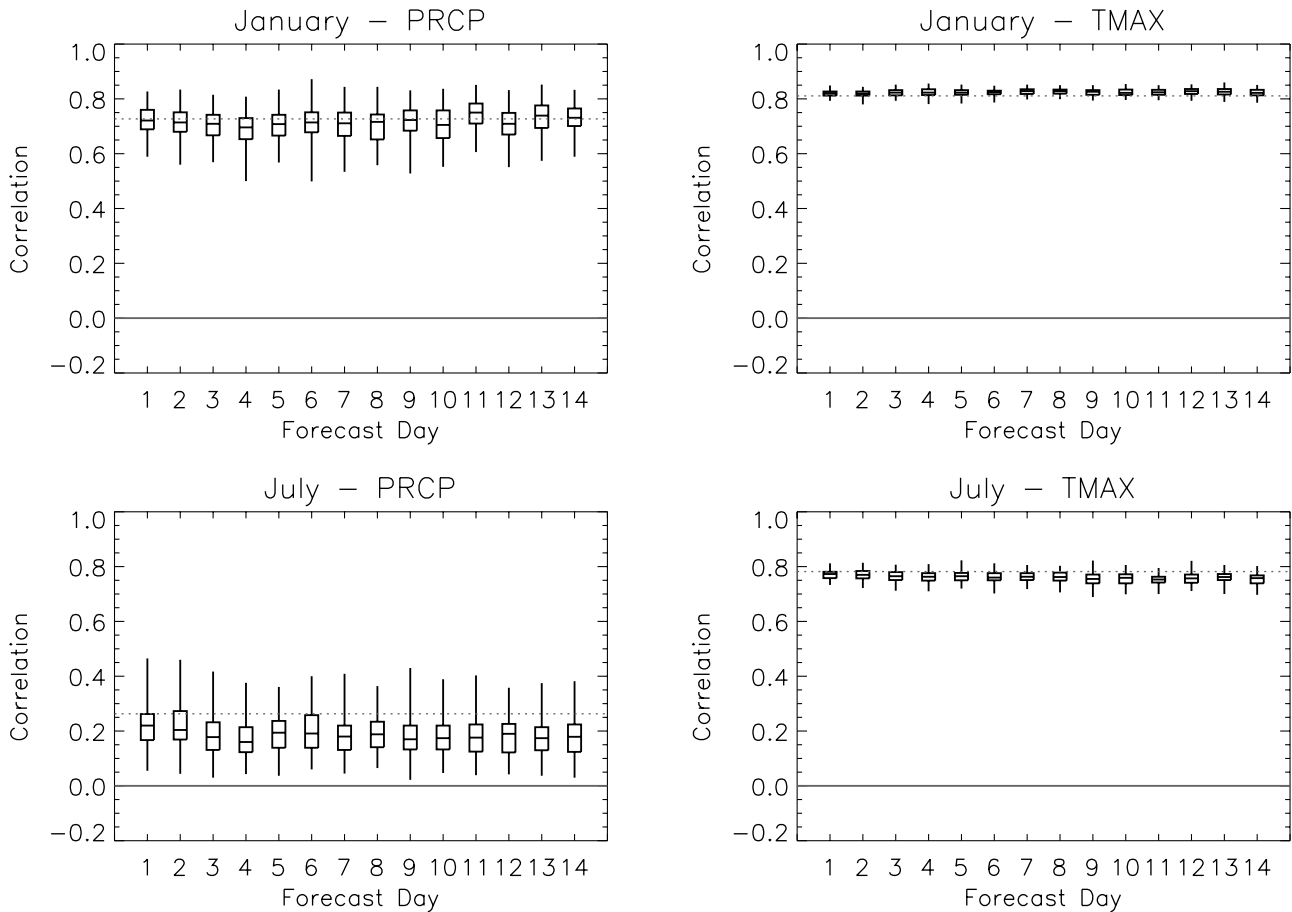


Figure 6. Same as Figure 5 but for stations GA0140 and GA2266.

reference climatological forecasts. In summer, the skills drops down considerably even at short forecast lead times. For Cle Elum, however, we see higher skills even during the summertime. This is because the basin is the driest during the summer months (see Figure 2), and higher skill arises from consistent dry forecasts from the downscaling model.

[45] For temperature (Figure 8) the skills are higher than that of precipitation, with a maximum for all the basins to be around 0.5. Higher skills are generally observed during all the seasons and are valuable up to lead times of 5 days. Also, for both temperature and precipitation, the results overall are very consistent, showing skills diminishing with an increase in forecast lead times.

[46] Since the RPSS is only a single number, it is a useful measure for ranking competing forecasts, but it does not illuminate the underlying basis for the forecast errors. For example [Hamill, 1997], are the forecasts too specific, or biased? Are 25% of the forecasts on average below the 25th percentile of forecast distribution? Thus we need additional forecast verification measures to address such issues. The reliability diagram [Wilks, 1995] is a frequently used tool in probabilistic forecast verification and is discussed in the next section.

4.5. Forecast Reliability

[47] The fundamental interest in forecast verification is to analyze the joint probability distribution of forecasts and observations [Wilks, 1995]. Let y_i denote discrete forecasts that can take one of the any I values y_1, y_2, \dots, y_I ; and let o_j

be the corresponding observations (discrete), which can have any of the J values o_1, o_2, \dots, o_J . Then the joint probability mass function $p(y_i, o_j)$ of the forecasts and observations is given by

$$p(y_i, o_j) = P(y_i \cap o_j); \quad i = 1, \dots, I; \quad j = 1, \dots, J. \quad (16)$$

Using the multiplication rule of probability [e.g., Ang and Tang, 1975, p. 43], equation (16) can be factored as

$$p(y_i, o_j) = p(o_j | y_i)p(y_i), \quad (17)$$

where $p(o_j | y_i)$ is the conditional probability, implying how often each possible event (out of J outcomes) occurred on those occasions when the single forecast y_i was issued; and $p(y_i)$ is the unconditional (marginal) distribution that specifies the relative frequencies of use of each of the forecast values y_i .

[48] The reliability diagram graphically represents the performance of probability forecasts of dichotomous events and depicts the conditional probability that an event occurred (say, o_1), given the different probabilistic forecasts (y_i); that is, the observed relative frequency, $p(o_1 | y_i)$, as a function of the forecast probability $p(y_i)$. This was implemented as follows.

[49] First, the ensemble output (105 ensemble members) for a given basin is converted into probabilistic forecasts (i.e., the probability a specific event occurs). In this case, the

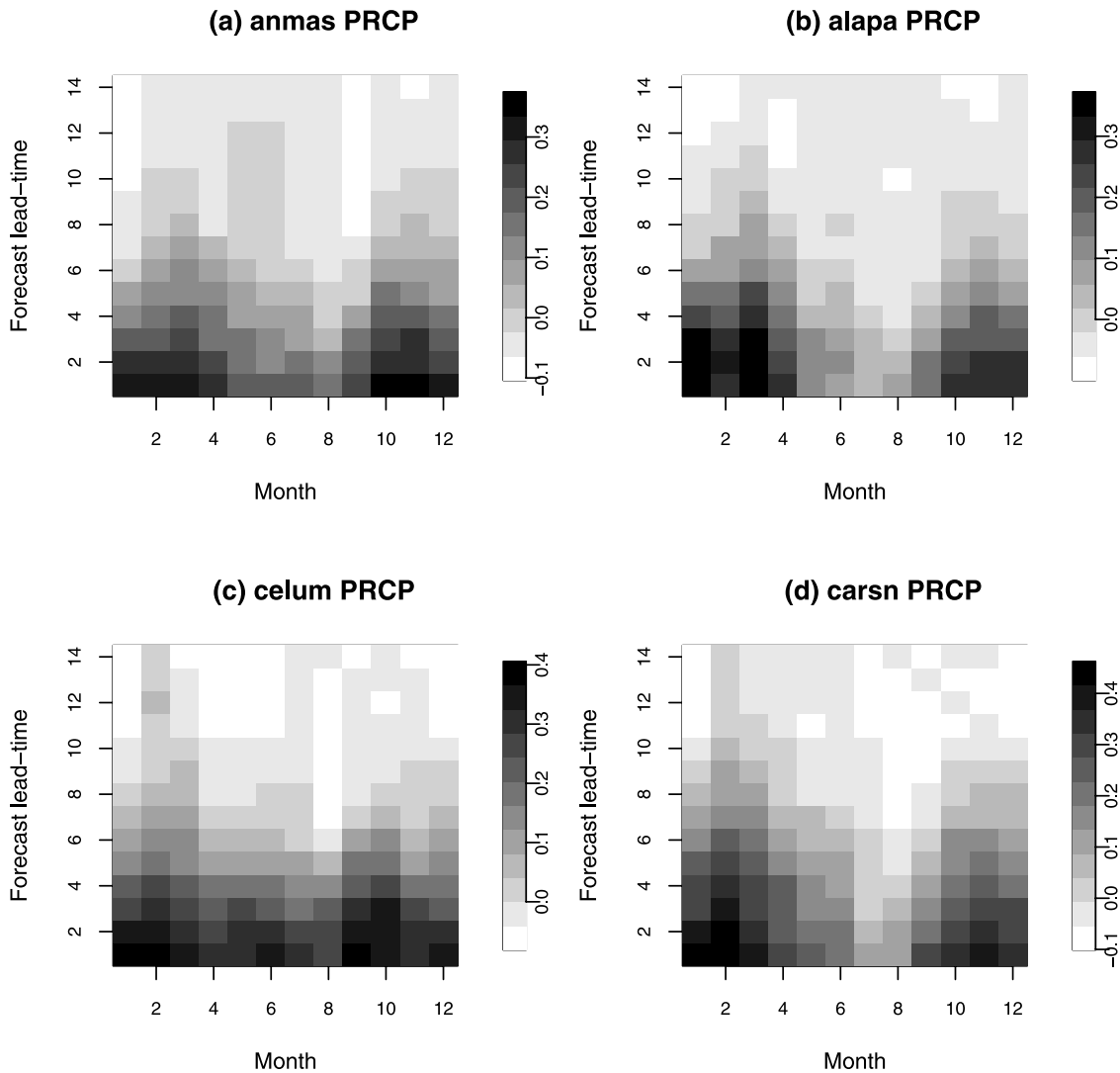


Figure 7. Median RPSS for precipitation in the four basins: (a) Animas, (b) Alapaha, (c) Cle Elum, and (d) East Carson. The months (January–December) are the horizontal axis, and lead times are in the vertical axis.

“event” is that the day is forecasted to lie in the upper tercile of the distribution, and the probability is simply calculated as all ensemble members in the upper tercile divided by the total number of ensemble members. The upper tercile was chosen to focus attention on events such as heavy precipitation and high temperatures that can cause significant changes to streamflow. Next, the observed data are converted to a binary time series: A day is assigned “one” if the data lies in the upper tercile and “zero” if the data does not. The above steps produce a set of probabilistic forecast–observation pairs for each variable, station, month, and forecast lead time. Finally, the forecasted probabilities are classified into I categories (i.e., probabilities between 0.0 and 0.1, between 0.1 and 0.2, . . . , between 0.9 and 1.0, a total of 10 categories), and for each category both the average forecasted probability and the average of the observed binary data are calculated. It should also be noted that the number of categories used affects the forecast resolution (i.e., the ability to distinguish subsample forecast periods with different relative frequencies of the event). These averaged observed relative frequency and forecast

probability values were then plotted to form the basic reliability diagram.

[50] Reliability diagrams for January precipitation and maximum temperature in the four study basins at 5-day forecast lead time are shown in Figures 9 and 10, respectively. For precipitation, if there were fewer than one third of days with precipitation, a value of zero was used for the probabilities in the reliability diagrams. The 1:1 diagonal in these figures represent the perfect reliability line, and the inset histogram shows the frequency of use of each of the forecasts, $p(y_i)$. Also, to construct the reliability diagrams for each basin as a whole, the forecast–observation pairs were lumped together from all stations in that basin (see Table 1). Results show that overall, the forecasted probabilities match the observed relative frequencies remarkably well for both precipitation and temperature. In case of precipitation (Figure 9), for example, in the case of the Alapaha basin, we see some tendency of higher observed relative frequency at lower forecasted probabilities and the opposite at higher forecasted probabilities. In other words, when a low probability of the event is forecasted, the actual

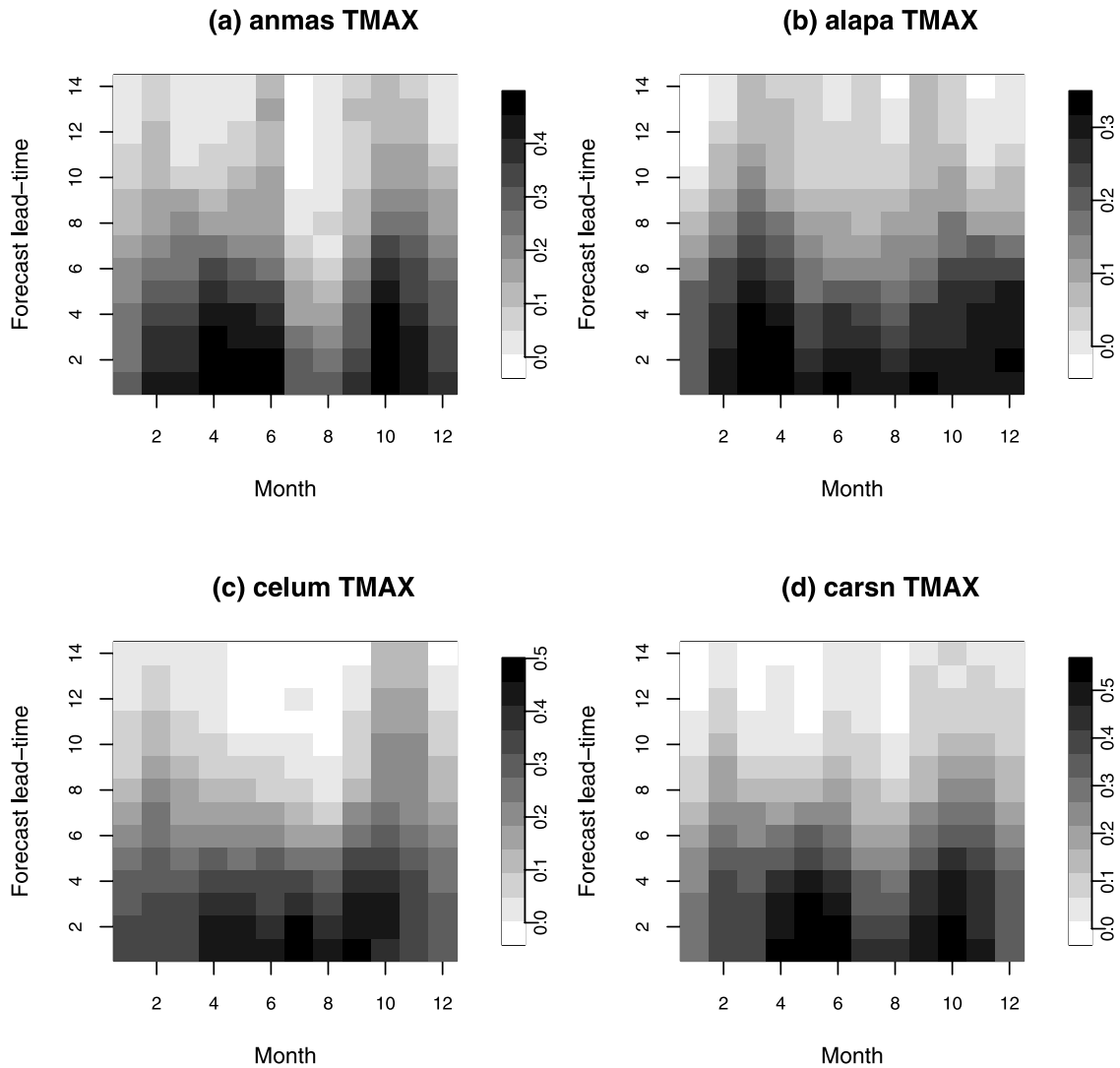


Figure 8. Same as Figure 7 but for temperature.

occurrence of the event is more common, and vice versa. Also note that the sample size at high forecast probabilities is often very small, except in case of the Cle Elum. This basin in the Pacific Northwest receives considerable precipitation in January, and we have enough subsamples in each of the forecasted probabilities (see inset histogram in Figure 9c), i.e., we have excellent resolution and reliability in our downscaled forecasts.

[51] For the case of maximum temperature (Figure 10), in general we have sharper forecasts (high resolution) at the price of some reduced reliability, in particular for the Cle Elum and East Carson, where we can see more frequent occurrences of the event when the forecast probability was slightly lower. Reliability diagrams similar to the above were also plotted for the month of July (not shown). Overall results were similar to January, but for precipitation in the East Carson, practically all the forecasted probabilities (frequency of usage) were within the lowest category (0.0–0.1) and imply the presence of rare events. Though these forecasts were reliable, they exhibit minimal resolution.

[52] Once again, the results are overall quite impressive and demonstrate that the proposed K -nn algorithm can be used to generate reliable forecasts with negligible condi-

tional bias. The reliability of the forecasts was further evaluated using rank histograms.

4.6. Rank Histograms

[53] Rank histograms were used to evaluate the reliability of ensemble forecasts and to diagnose errors in their mean and spread. Rank histograms for a given month and forecast lead time were constructed by repeatedly tallying the ranks of the observed precipitation and temperature values relative to values from the 105 member ensemble. The process to obtain the rank histogram for precipitation is slightly different from that of temperature because of the presence of a large number of zero-precipitation days in the observed and ensemble precipitation time series. For temperature the rank histogram was implemented as follows:

[54] Let for a given forecast day (say, j), and forecast lead time (say, l), let $X = (x_{(1)}, \dots, x_{(n)})$ be the sorted n -member ensemble (recall $n = 105$ in this study) and let V be the observed temperature. Then the rank of V , which can have $(n + 1)$ possible values relative to the sorted ensemble, is obtained. Let this rank be denoted by r_j^l . If, say, there were 20 years of data, then for January (assuming no missing observations), there would be 620 (31 days \times 20 years) time

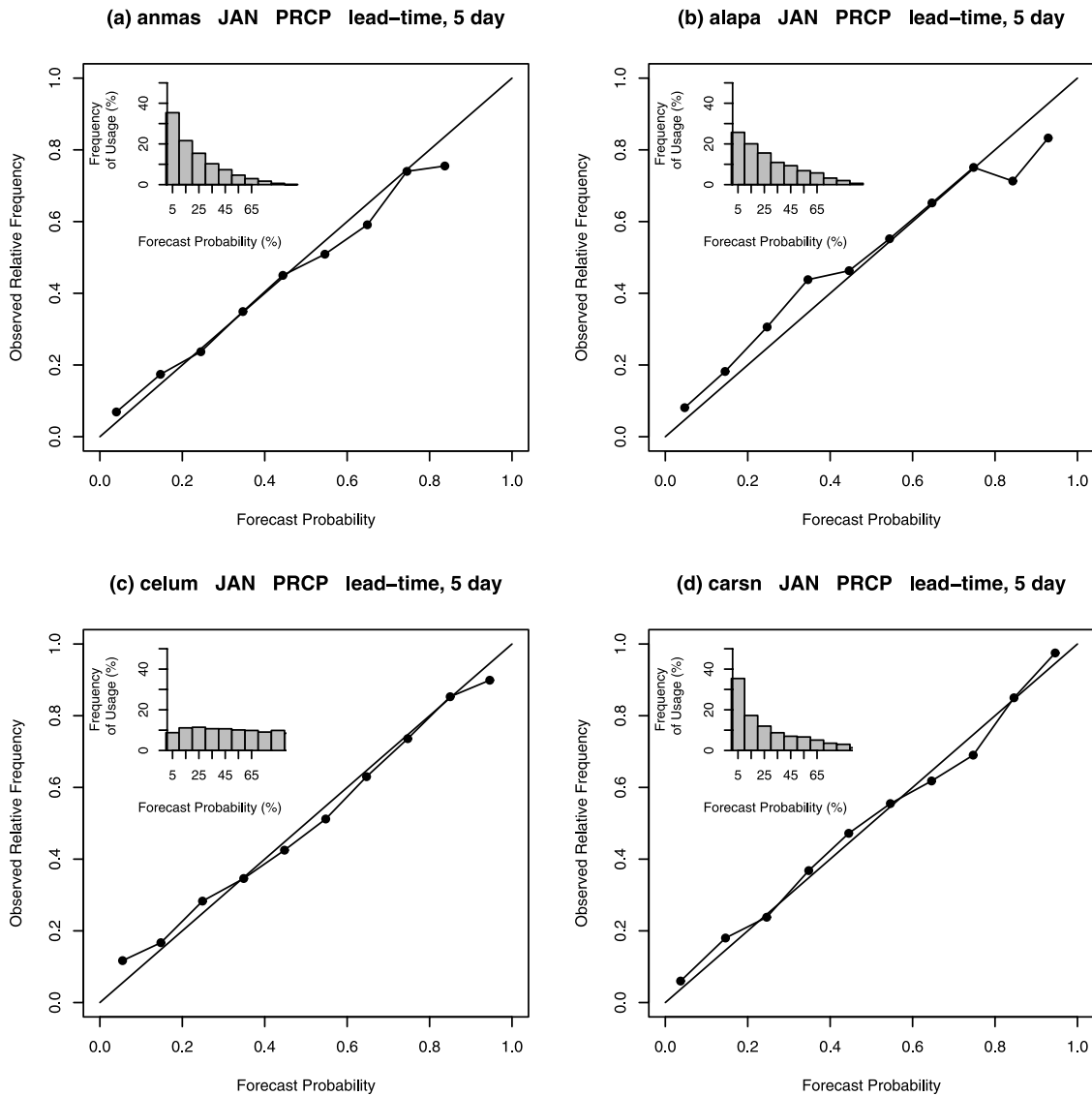


Figure 9. Basin reliability diagram for January precipitation in the four basins: (a) Animas, (b) Alapaha, (c) Cle Elum, and (d) East Carson, at 5-day forecast lead time. Inset histograms indicate frequency of use of the forecasts.

elements in this time series for a given forecast lead time. By tallying the ranks of the observed through this time series we can obtain a vector of ranks for the selected month (m) and lead time (R_m^l),

$$R_m^l = [r_1^l, r_2^l, \dots, r_N^l]^T, \tag{18}$$

where N is the length of the time series (or sample size, e.g., 620). The elements of R_m^l are then binned into the $(n + 1)$ possible categories for constructing the rank histogram. So the rank histogram constitutes the rank of the observed and the probability of the rank to fall in any one of the $(n + 1)$ categories.

[55] In case of precipitation when there are zero precipitation days in the observed and ensemble time series, a modified rule for rank assignment was implemented [Hamill and Colucci, 1998]. If, say, there are M members tied with the verification (i.e., M ensemble members with zero precipitation), a total of $(M + 1)$ uniform random deviates

[Press et al., 1992] are generated, corresponding to the M members, and one for the observed zero precipitation. Then the rank of the deviate corresponding to the observed in the pool of $(M + 1)$ deviates is determined. The rank histogram is then constructed in a manner similar to the one described for temperature.

[56] To interpret the rank histograms, it is assumed that the observations and the ensemble members are samples from the same probability distribution. In that case, counting the rank of the observation over several independent samples, an approximately uniform distribution should result across the possible ranks, i.e.,

$$E[P(x_{(i-1)} \leq V < x_{(i)})] = \frac{1}{n + 1}, \tag{19}$$

where, $E[]$ denotes the expected value and P is the probability. Hamill [2001] describes the interpretation of rank histograms and provides these guidelines. When the ensemble members are from a distribution with lack of

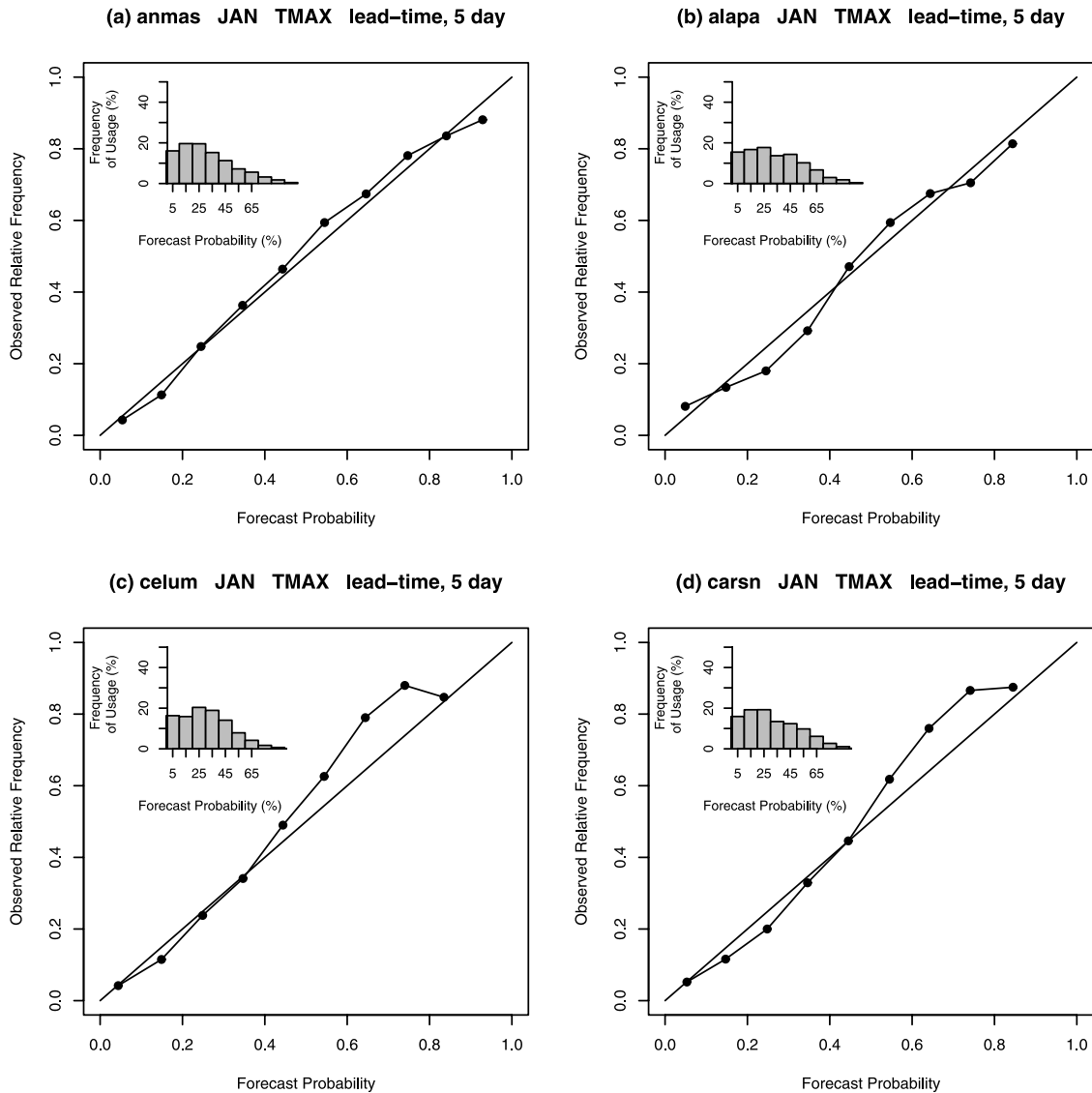


Figure 10. Same as Figure 9 but for temperature.

variability, a U-shaped rank histogram results. An excess of variability in the ensemble members, on the other hand, overpopulates the middle ranks, and ensemble bias (positive or negative) excessively populates the (left or right) extreme ranks.

[57] Figures 11 and 12 show the basin rank histograms for precipitation and maximum temperature, respectively. Basin rank histograms were constructed by pooling in ranks from all stations for a given basin. The basin rank histograms are shown for January at 5-day lead time. For precipitation (Figure 11), the rank histograms are relatively flat, demonstrating that the K -nn method produces realistic ensemble spread. The noise in the rank histograms simply reflects the noisy character of the precipitation time series. For temperature (Figure 12), the basin rank histograms are largely uniform in the middle ranks, except at the extremities where we observe some bias. We see that on average, nearly 2% of the time the observed temperature can be lower (greater) than the lowest (highest) ensemble member.

[58] In general, from all the cases (including summer) we see from the precipitation rank histograms that the

ensembles are relatively flat, and for temperature there is only a small fraction of cases ($\sim 2\%$) when the observed falls outside the ensemble range. We also constructed rank histograms for each of the individual stations used in the study (see Table 1) and overall found no unusual behavior in the structure of the rank histograms. The next question then is, Can we use the ensemble spread information to predict forecast skill? This topic is discussed in the next section.

4.7. Spread-Skill Relationships

[59] Ensemble forecasts provide an estimate of the forecast probability distribution. If the spread of this distribution varies from forecast to forecast, then the spread in the distribution may be related to the forecast skill [Kalnay and Dalcher, 1987; Whitaker and Loughe, 1998]. To analyze the spread-skill relationship, we first need to select appropriate measures to define the ensemble spread and ensemble skill. We used three measures of ensemble spread: (1) standard deviation of the ensembles, (2) interquartile range, and (3) the 95th minus the 5th quantiles. As skill

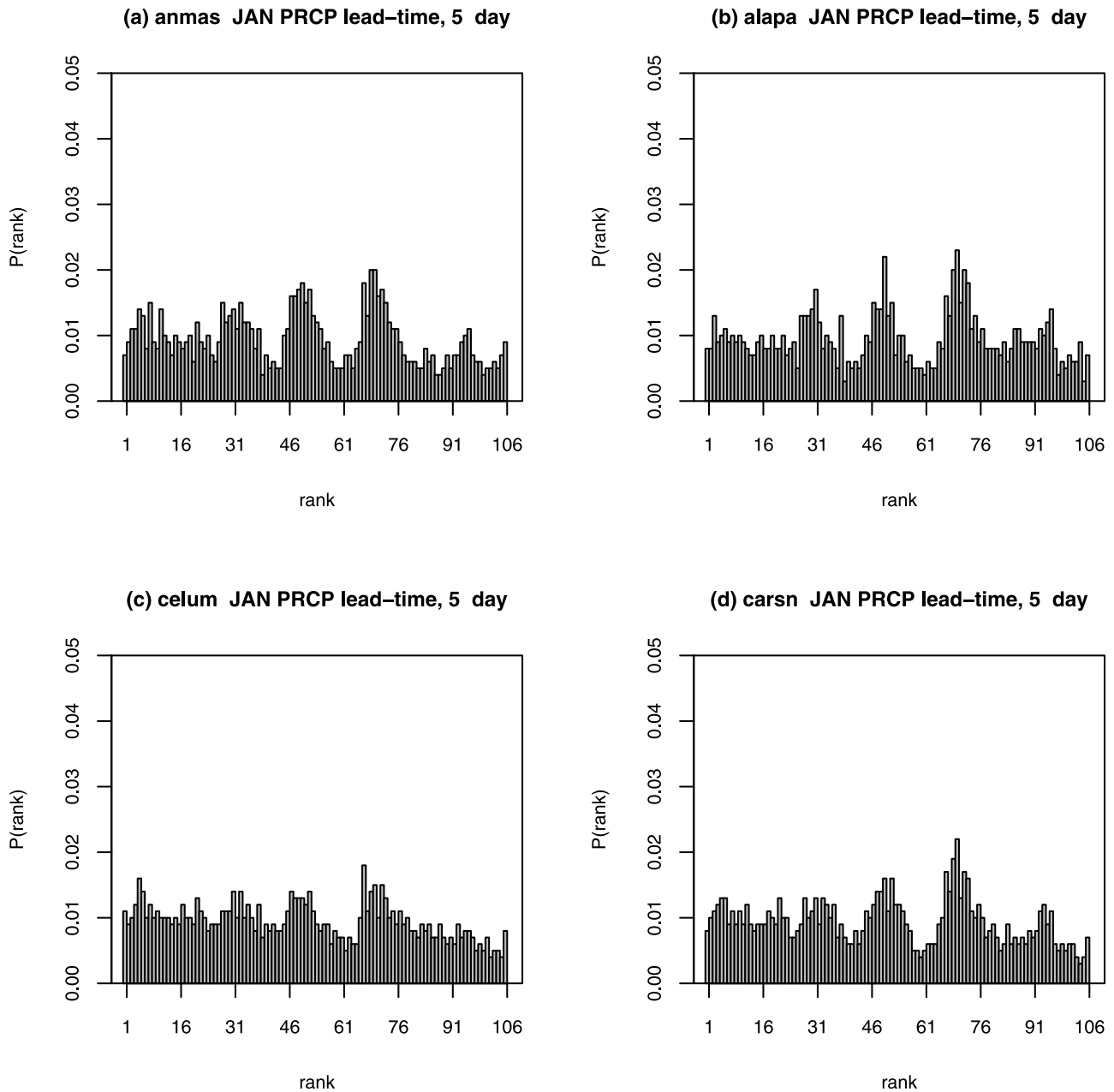


Figure 11. Rank histogram for January precipitation at 5-day forecast lead time with 105 members for the four basins.

measures we used (1) RPSS and (2) the absolute error of the ensemble mean (absolute difference between the observed and the ensemble mean). The utility of ensemble spread as a predictor of ensemble skill has traditionally been measured in terms of linear correlation, although *Whitaker and Loughé* [1998] suggest an analysis of the joint spread-skill probability distribution.

[60] Contingency table of spread (ensemble standard deviation) and skill (RPSS) for 5-day forecasts of January precipitation for example station WA0456.COOP is given in Table 2. Here we considered all days for which the observed precipitation was greater than 0.01 inch (0.3 mm). The entries Table 2 are the joint probability of obtaining the spread and skill values in the indicated quintiles. The columns are spread quintiles, and the rows are skill quin-

tiles. If there were no correlation between spread and skill, all entries in the table would be equal to 0.2. On the other hand, if there were a perfect linear relationship between spread and skill (correlation equal to one), all the diagonal elements would be one and the off-diagonals would be zero. Many of the entries in Table 2 are not very different from 0.2, except at the corners. For example, if the spread is in the lowest quintile, there is about 2.5 times higher probability of the skill to be in the lowest, rather than the highest quintile. This observation was consistent among all stations in the study.

[61] To summarize the contingency table for all stations in a basin, we constructed box plots showing the variation of the joint spread-skill probability for all spread and skill quintiles. Results are shown for January precipitation

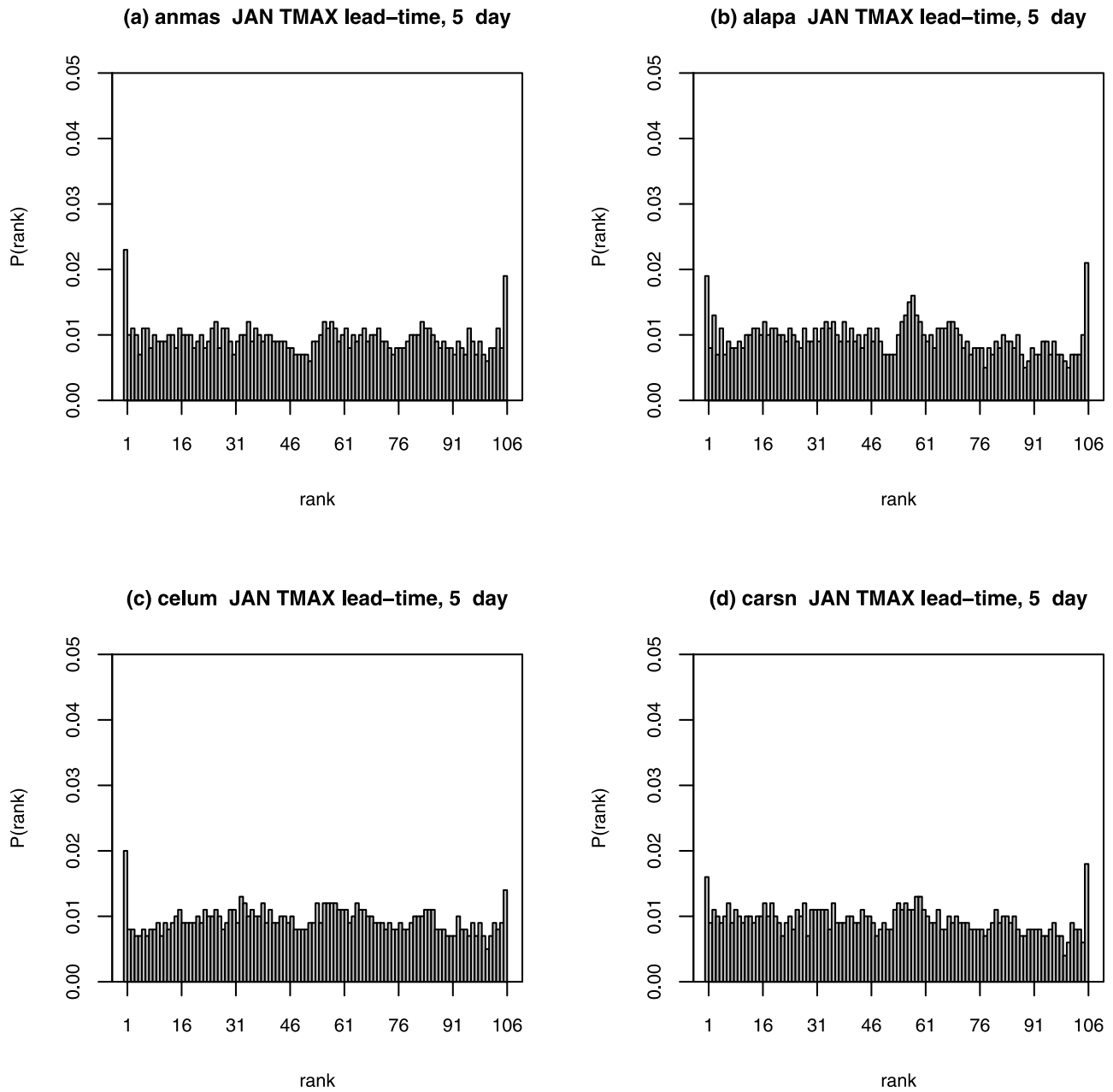


Figure 12. Same as Figure 11 but for temperature.

at 5-day forecast lead time for the Animas and Alapaha basins in Figures 13 and 14, respectively. In each of these figures we show three cases: (1) considering all days (left column); (2) days with precipitation within 0 mm and 0.3 mm, including the zero precipitation days (middle column); and (3) days with precipitation greater than 0.3 mm (right column). For each spread quintile, box plots are plotted showing the variation of the joint probability of spread-skill in all stations of the basin with the skill quintiles as the abscissa. The dashed horizontal line corresponds to a joint probability value of 0.2 when there is no spread-skill correlation.

[62] In both Figures 13 and 14 we see that when all days are considered, and also in the case where precipitation is within 0.3 mm (with zero precipitation days included), the spread-skill relationship is negatively correlated. That is, for

Table 2. Contingency Table of Spread (Ensemble Standard Deviation) and Skill (RPSS) for 5-Day Forecasts of January Precipitation for Station WA0456.COOP When the Observed Precipitation is Greater Than 0.3 mm^a

	0–20%	20–40%	40–60%	60–80%	80–100%
0–20%	0.47	0.14	0.14	0.22	0.03
20–40%	0.09	0.33	0.29	0.18	0.11
40–60%	0.14	0.24	0.20	0.26	0.16
60–80%	0.11	0.19	0.27	0.20	0.23
80–100%	0.19	0.10	0.10	0.14	0.47

^aThe entries are the joint probability of obtaining the spread and skill values in the indicated quintiles. The columns are spread quintiles, and the rows are skill quintiles.

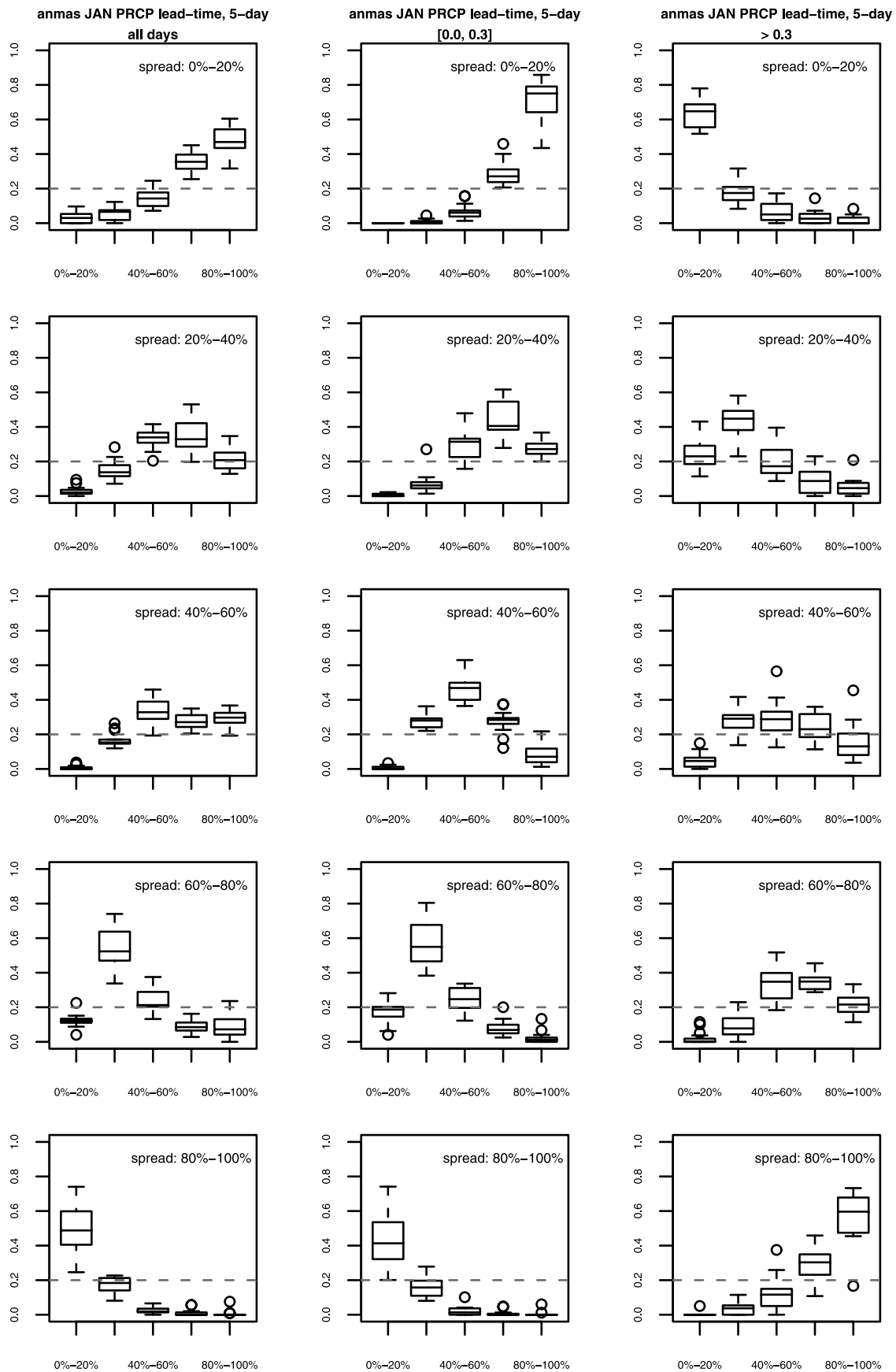


Figure 13

lower spread, there is a higher probability of greater skill. Here a large number of ensemble members with zero precipitation contribute to both a lower ensemble spread and higher skill for small precipitation amounts. Conversely, a small number of ensemble members with zero precipitation contribute to higher ensemble spread and lower skill for small precipitation amounts. Unfortunately, these results do not allow us to construct any meaningful spread-skill relationships in order to place time-variant confidence limits on precipitation forecasts.

[63] Similar box plots for maximum temperature (here data from all days were used) are shown in Figure 15 for the Animas (left column) and Alapaha (right column) basins. In all cases we see that the boxes are close to the dashed horizontal line (i.e., joint probability value of 0.2), which implies that there is no spread-skill correlation. Similar results for both precipitation and temperature were observed for July.

[64] All the results presented here used standard deviation and RPSS as the spread and skill measures, respectively. Analysis was also carried out using the other spread and skill measures, and the results were found to be robust, that is, the underlying spread-skill relationships do not change with the choice of different measures. Also, though no clear spread-skill relationships were apparent here, the K -nn method is theoretically capable of extracting the spread-skill relationship if it exists in the atmospheric model.

5. Summary and Conclusions

[65] A method for statistical downscaling using the K -nn algorithm in eigenspace was developed. A 20-year (1979–1998; 7305 days) data archive consisting of model outputs from the NCEP 1998 version of the operational medium-range forecast model from NOAA/CDC was used in this study. A total of 15 MRF runs (one control run plus 14 ensemble members) were available for analysis. Seven MRF model output variables going out to lead time of 14 days was used in the downscaling algorithm. Analogs to (7305×14) feature days using a 14-day temporal window were subsequently identified. All data were projected onto eigen space, and distance between a feature day and all candidate days were calculated using a weighted Euclidian norm. The weighting used considered the fractional variance explained by a given principal component. The distances were then sorted in ascending order, and weights were assigned to each using the bisquare weight function. On the basis of weights, and repeatedly generated uniform random numbers a set of seven ensemble members were created from each MRF run.

[66] Results were assessed over four river basins distributed across the contiguous United States. These were Animas (southwestern Colorado); Alapaha (southern Georgia); Cle Elum (central Washington); and east fork of

the Carson (California-Nevada border). The K -nn downscaling algorithm was repeated for the 15 MRF runs and for the four basins. Since from each MRF run seven ensemble members were generated, the 15 MRF runs yielded a total of 105 ensemble members for each basin. To obtain local estimates of precipitation and temperature, closest COOP stations (within a 100 km search radius) from the center of the basins were selected, and observed data corresponding to the downscaled dates were used to obtain these estimates. The precipitation and temperature estimates from these 105 ensemble members over 20 years and 14-day forecast lead times were used to evaluate the K -nn downscaling methodology.

[67] The statistics included seasonal cycles, bias, spatial correlations, and a suite of forecast verification statistics. The K -nn downscaling model in all cases largely captured the seasonal variation of precipitation and temperature. Precipitation biases were generally within 20%, but in many cases (mostly for the climatologically drier Animas basin at longer lead-times) exceeded 20%. This is consistent with the noisy character of precipitation time series. Temperature biases were small and within 0.5°C . Since we used data for all stations on a given day, the K -nn method intrinsically preserves the spatial autocorrelation structure and the consistency between variables. Furthermore, since this method relies solely on the atmospheric model output and does not incorporate any joint relationship between the atmospheric and surface variables, it does not fully preserve the lag-one correlation statistics (not shown). However, postprocessing using an ensemble reordering type method [Clark *et al.*, 2004] can be used to recover this serial correlation. Also, methods based on nonhomogeneous hidden Markov models (NHMM) [e.g., Hughes and Guttorp, 1994; Hughes *et al.*, 1999] to downscale synoptic atmospheric patterns to local scale precipitation have been shown to preserve temporal correlation but fail to preserve spatial correlation.

[68] Next we evaluated the skill, reliability, and time-variant spread-skill relationships in the downscaled forecast ensembles. The rank probability skill score (RPSS) was used to verify the forecast skills. For precipitation, the skills generally were higher in winter than in summer and valid at only short forecast lead times (2–3 days). Temperature RPSS scores were around 0.5, and valuable skill was present even up to lead times of 5 days in all seasons. Forecast reliability or conditional bias were evaluated using reliability diagrams, and we found that the observed relative frequencies of the event (days being in the upper tercile) matched well with forecasted probabilities, and there was very little conditional bias in the forecasts.

[69] Rank histograms showed that although precipitation ensembles are to an extent noisy, the ensemble spread is nevertheless meaningful. For temperature, the observed fell outside the ensemble range in about 2% cases. Next, we

Figure 13. Box plots of joint spread-skill probability for skill quintiles at given spread quintiles. The vertical axis is the joint probability of spread (ensemble standard deviation) and skill (RPSS), and the horizontal axis shows the skill quintiles. Results are shown for January precipitation at 5-day forecast lead time for the Animas basin. The box plots are constructed using data from all the stations in the basin. Three cases are shown: (left) using data from all days; (middle) using data for days when the observed precipitation is between 0 and 0.3 mm (both values inclusive); and (right) when the observed precipitation is greater than 0.3 mm. The dashed horizontal line in each plot corresponds to joint probability value of 0.2 when there is no spread-skill relationship. See color version of this figure in the HTML.

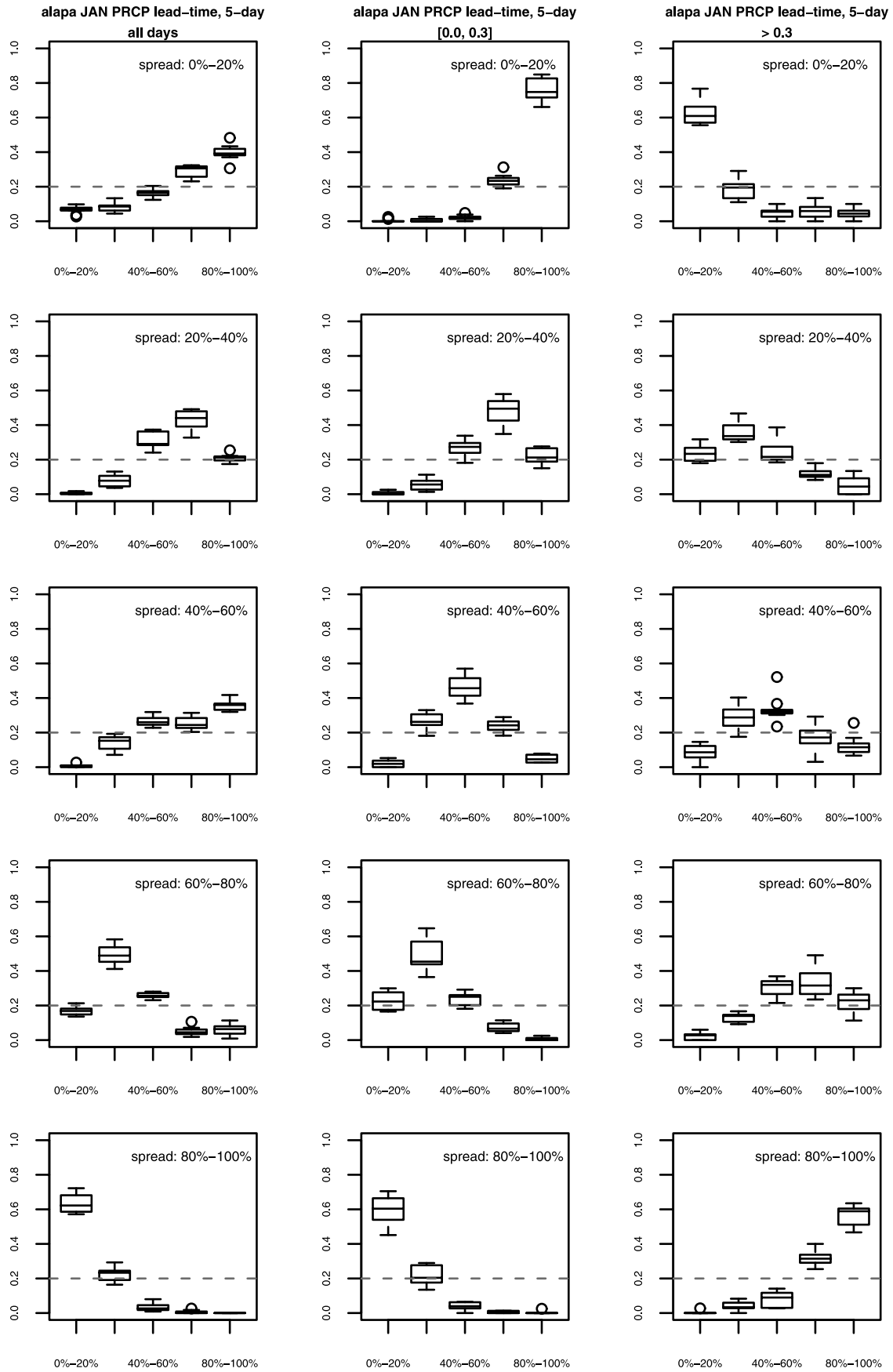


Figure 14. Same as Figure 13 but for the Alapaha basin. See color version of this figure in the HTML.

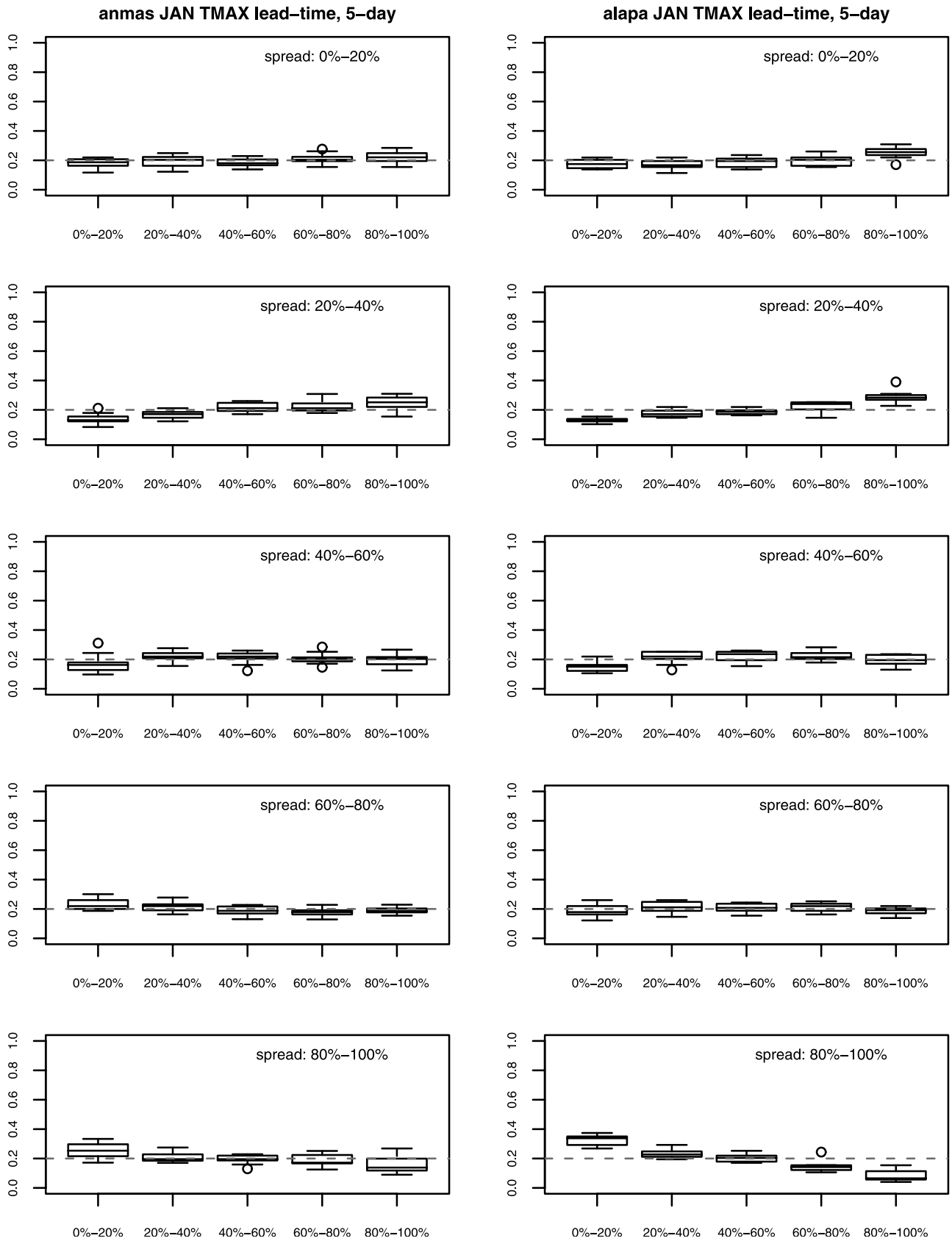


Figure 15

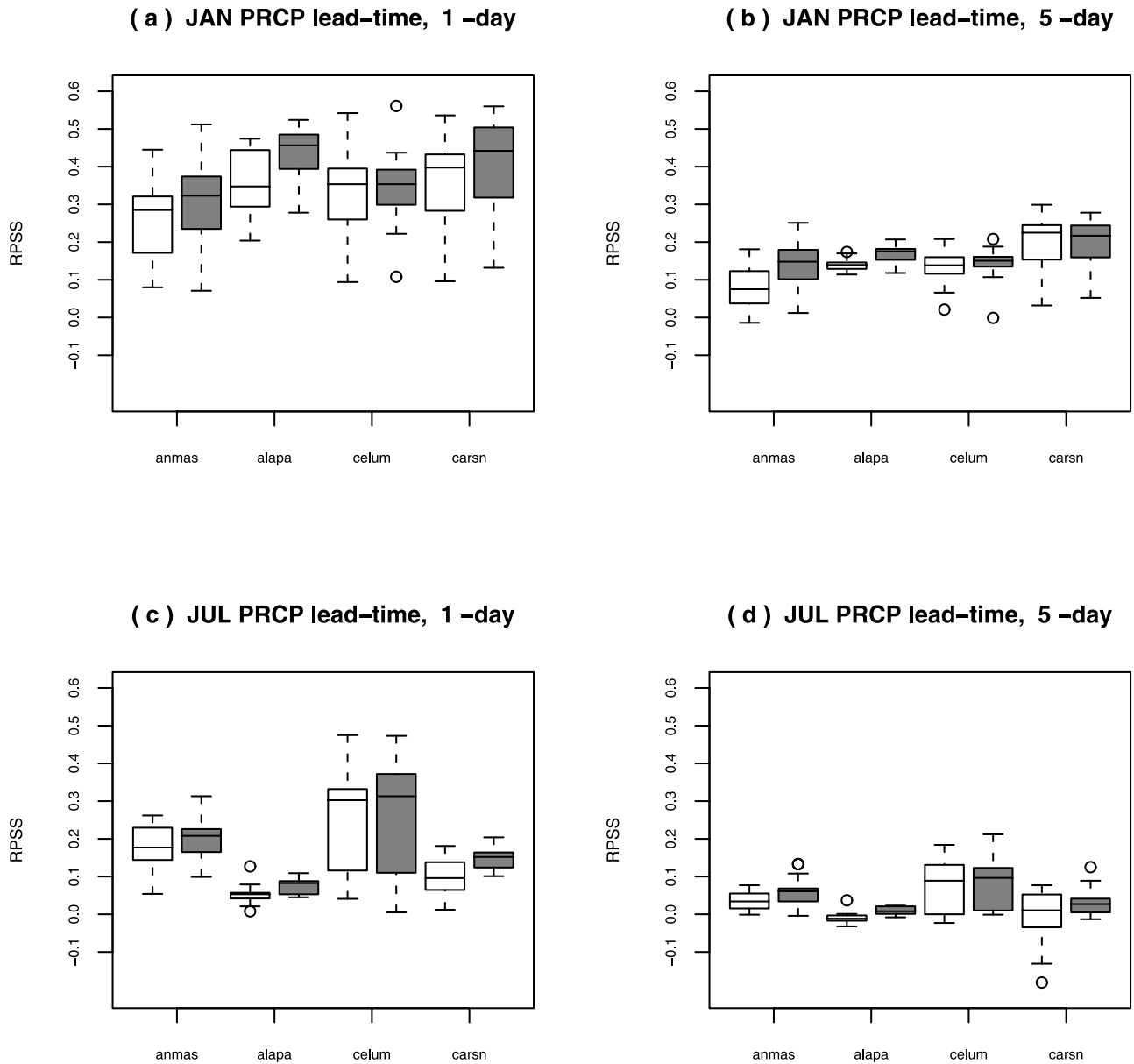


Figure 16. Box plots comparing skills (RPSS) in precipitation forecasts in the four study basins obtained from downscaling using K -nn (open), and MLR (shaded): (a) January precipitation for 1-day lead time; (b) January precipitation for 5-day lead time; (c) July precipitation for 1-day lead time; and (d) July precipitation for 5-day lead time.

analyzed possible spread-skill relationships. We did not find a meaningful relationship to forecast precipitation forecast skills. For temperature, results clearly showed that there is no relationship between the ensemble spread and skill.

[70] Though regression-based approaches are widely used to extract local-scale information from forecast models [e.g., Antolik, 2000], these methods are not data-driven, they need

variable transformations, they do not intrinsically preserve space-time autocorrelations and cross correlations of the downscaled variables, and they cannot be utilized to investigate spread-skill relationships. We did a comparison, however, to test the skill (RPSS) of the K -nn approach with a multiple linear regression (MLR) based downscaling method (see Clark and Hay [2004] and Clark et al.

Figure 15. Box plots of joint spread-skill probability for skill quintiles at given spread quintiles. The vertical axis is the joint probability of spread (ensemble standard deviation) and skill (RPSS), and the horizontal axis shows the skill quintiles. Results are shown for January maximum temperature at 5-day forecast lead-time for the (left) Animas and (right) Alapaha basins. The box plots are constructed using data from all the stations in the basins. The dashed horizontal line in each plot corresponds to joint probability value of 0.2 when there is no spread-skill relationship. See color version of this figure in the HTML.

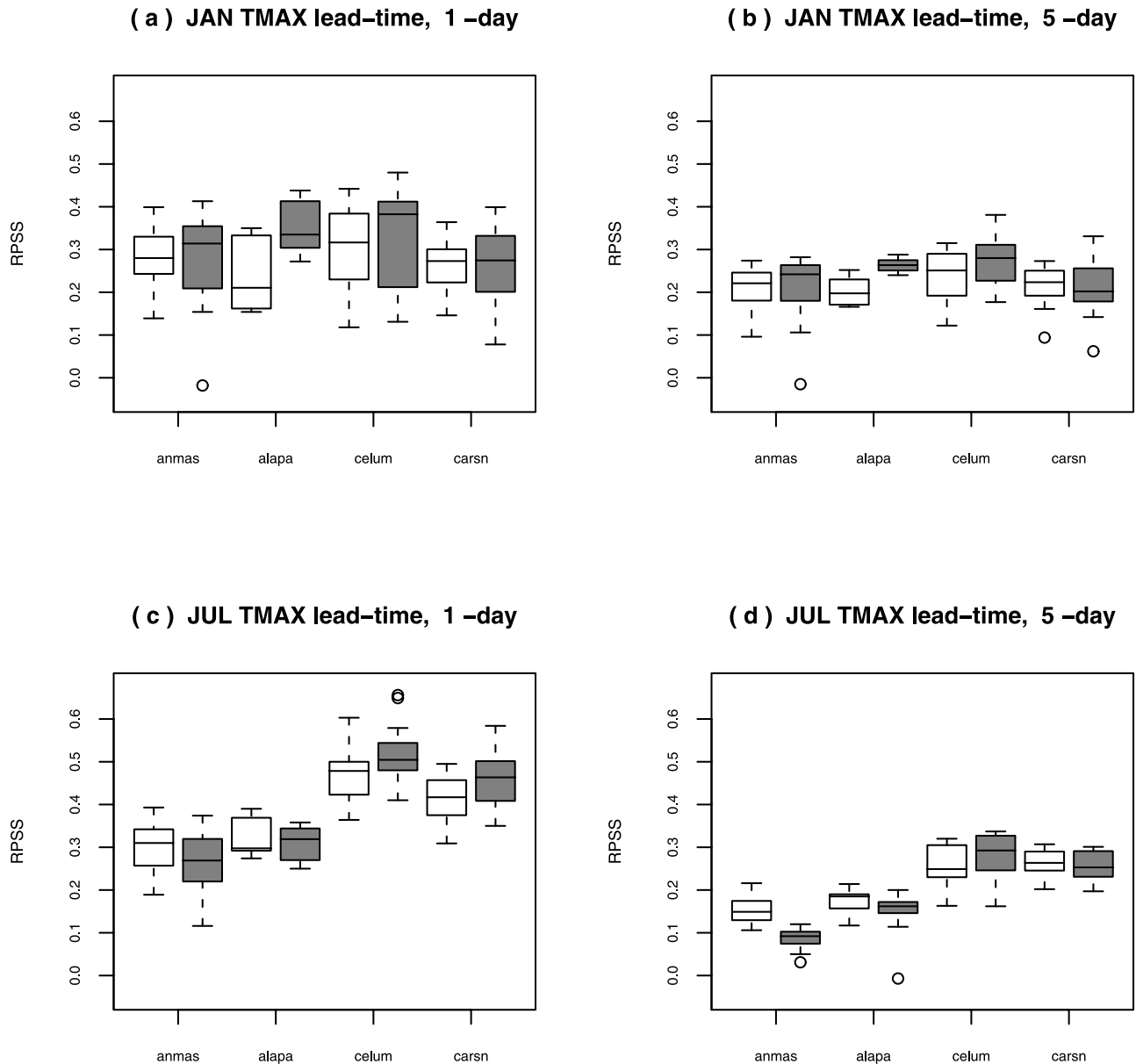


Figure 17. Same as Figure 16, but for temperature.

[2004] for descriptions of the MLR method). Results of this comparison are summarized in box plots shown in Figures 16 and 17 for precipitation and temperature, respectively. Given that the K -nn algorithm does not use the joint relationship between forecast model output and station data, the results are extremely impressive. The skill obtained from the K -nn method is competitive with the skill obtained using MLR. The MLR utilizes the joint relationship between surface and atmospheric variables and needs postprocessing to reconstruct the space-time variability between the ensembles (typically the downscaling is done for each station individually). The PCs in the K -nn method also provide a consistent spatial representation, whereas the variables in case of MLR typically change from one station to the other.

[71] The marginally higher skills that are seen in case of MLR are also because the 15-member ensemble mean from the MRFs are used as predictors. Furthermore, the sum of

squared errors between observed and downscaled values at each station is explicitly minimized in developing the MLR models. Finally, the K -nn method is computationally efficient and can be readily implemented. One shortcoming of K -nn type algorithms is that values not seen in the historical record are not reproduced. This could be a potential problem if the historical archive is short. *Rajagopalan and Lall* [1998] describe a modified resampling strategy with perturbations of the historical data to overcome this limitation. The results described here, however, demonstrate the strength of this algorithm and provide a viable alternative in providing skillful and reliable downscaled forecasts to other downscaling methods.

[72] **Acknowledgments.** This research was supported by the NOAA GEWEX Americas Prediction Program (GAPP) and the NOAA Regional Integrated Science and Assessment (RISA) Program under awards NA16GP1587 and NA17RJ1229. Thanks are due two anonymous reviewers for their comments, which improved the manuscript.

References

- Ang, A. H.-S., and W. H. Tang (1975), *Probability Concepts in Engineering Planning and Design: Basic Principles*, 409 pp., John Wiley, Hoboken, N. J.
- Antolik, M. S. (2000), An overview of the National Weather Service's centralized statistical quantitative precipitation forecasts, *J. Hydrol.*, 239, 306–337.
- 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(11), 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. Hydrometeorol.*, 5(1), 15–32.
- Clark, M. P., S. Gangopadhyay, L. E. Hay, B. Rajagopalan, and R. L. Wilby (2004), The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields, *J. Hydrometeorol.*, 5(1), 243–262.
- 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. Meteorol.*, 39, 1580–1591.
- Fukunaga, K. (1990), *Introduction to Statistical Pattern Recognition*, Elsevier, New York.
- Hamill, T. M. (1997), Reliability diagrams for multicategory probabilistic forecasts, *Weather Forecasting*, 12, 736–741.
- Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, 129, 550–560.
- Hamill, T. M., and S. J. Colucci (1998), Evaluation of the Eta-RSM ensemble probabilistic precipitation forecast, *Mon. Weather Rev.*, 126, 711–724.
- Hamill, T. M., J. S. Whitaker, and X. Wei (2004), Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts, *Mon. Weather Rev.*, 132, 1434–1447.
- Huber, P. J. (2003), *Robust Statistics*, 308 pp., Wiley-Interscience, Hoboken, N. J.
- Hughes, J. P., and P. Guttorp (1994), A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena, *Water Resour. Res.*, 30(5), 1535–1546.
- Hughes, J. P., P. Guttorp, and S. Charles (1999), A nonhomogeneous hidden Markov model for precipitation occurrence, *J. R. Stat. Soc.*, 48, 15–30.
- Kalnay, E., and A. Dalcher (1987), Forecasting forecast skill, *Mon. Weather Rev.*, 115, 349–356.
- Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, 77(3), 437–471.
- Lall, U., and A. Sharma (1996), A nearest neighbor bootstrap for time series resampling, *Water Resour. Res.*, 32(3), 679–693.
- Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1992), *Numerical Recipes in Fortran*, 2nd ed., 963 pp., Cambridge Univ. Press, New York.
- Rajagopalan, B., and U. Lall (1998), Interannual variability in western US precipitation, *J. Hydrol.*, 210, 51–67.
- Rajagopalan, B., and U. Lall (1999), A *K*-nearest neighbor simulator for daily precipitation and other variables, *Water Resour. Res.*, 35(10), 3089–3101.
- 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. Am. Meteorol. Soc.*, 73, 753–762.
- Salathé, E. P. Jr. (2003), Comparison of various precipitation downscaling methods for the simulation of streamflow in a rainshadow river basin, *Int. J. Climatol.*, 23, 887–901.
- Toth, Z., and E. Kalnay (1993), Ensemble forecasting at NMC: The generation of perturbations, *Bull. Am. Meteorol. Soc.*, 74, 2317–2330.
- Whitaker, J. S., and A. F. Loughe (1998), The relationship between ensemble spread and ensemble mean skill, *Mon. Weather Rev.*, 126, 3292–3302.
- Widmann, M., C. S. Bretherton, and E. P. Salathé, Jr. (2003), Statistical precipitation downscaling over the northwestern United States using numerically simulated precipitation as a predictor, *J. Clim.*, 16, 799–816.
- Wilks, D. S. (1995), *Statistical Methods in the Atmospheric Sciences: An Introduction*, 467 pp., Elsevier, New York.
- Yates, D., S. Gangopadhyay, B. Rajagopalan, and K. Strzepek (2003), A technique for generating regional climate scenarios using a nearest neighbor algorithm, *Water Resour. Research*, 39(7), 1199, doi:10.1029/2002WR001769.
- Zorita, E., and H. von Storch (1998), The analog method as a simple statistical downscaling technique: Comparison with more complicated methods, *J. Clim.*, 12, 2474–2489.

M. Clark, Center for Science and Technology Policy Research, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Campus Box 488, Boulder, CO 80309, USA.

S. Gangopadhyay, Hydrosphere Resource Consultants, 1002 Walnut, Suite 200, Boulder, CO 80302, USA. (gangopad@colorado.edu)

B. Rajagopalan, Department of Civil, Environmental and Architectural Engineering, University of Colorado, Campus Box 428, Boulder, CO 80309, USA.