A Bootstrap Technique for Testing the Relationship between Local-Scale Radar Observations of Cloud Occurrence and Large-Scale Atmospheric Fields

ROGER MARCHAND, NATHANIEL BEAGLEY, SANDRA E. THOMPSON, AND THOMAS P. ACKERMAN

Pacific Northwest National Laboratory, Richland, Washington

DAVID M. SCHULTZ

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, and NOAA/National Severe Storms Laboratory, Norman, Oklahoma

(Manuscript received 9 March 2005, in final form 7 October 2005)

ABSTRACT

A classification scheme is created to map the synoptic-scale (large scale) atmospheric state to distributions of local-scale cloud properties. This mapping is accomplished by a neural network that classifies 17 months of synoptic-scale initial conditions from the rapid update cycle forecast model into 25 different states. The corresponding data from a vertically pointing millimeter-wavelength cloud radar (from the Atmospheric Radiation Measurement Program Southern Great Plains site at Lamont, Oklahoma) are sorted into these 25 states, producing vertical profiles of cloud occurrence. The temporal stability and distinctiveness of these 25 profiles are analyzed using a bootstrap resampling technique.

A stable-state-based mapping from synoptic-scale model fields to local-scale cloud properties could be useful in three ways. First, such a mapping may improve the understanding of differences in cloud properties between output from global climate models and observations by providing a physical context. Second, this mapping could be used to identify the cause of errors in the modeled distribution of clouds—whether the cause is a difference in state occurrence (the type of synoptic activity) or the misrepresentation of clouds for a particular state. Third, robust mappings could form the basis of a new statistical cloud parameterization.

1. Introduction

Clouds are of tremendous importance to climate because of their direct effect on the earth radiation budget and because of their important role in global energy and water cycles. Limitations in the ability of global climate models (GCMs) to predict clouds create significant uncertainties in predicting and understanding climate. Clouds are among the largest source of uncertainty in GCM simulations (e.g., Cess et al. 1990; Potter and Cess 2004). Predicting clouds in GCMs is difficult for a variety of reasons, many of which arise because GCMs have horizontal grid spacings of many tens to hundreds of kilometers, whereas many of the processes responsible for the formation and dissipation of clouds occur on much smaller scales. These subgrid-scale processes cannot be explicitly resolved, meaning that cloud occurrence and cloud microphysical properties (e.g., hydrometeor type and size) must be parameterized or predicted from the larger-scale fields that the GCM can resolve. This parameterization is usually accomplished using semiempirical relationships, which are difficult to evaluate.

At present, most comparisons of model output and observational data average or aggregate the observations to put them on the same large spatial scale as the model. For ground-based measurements, this process generally requires making approximations that are difficult to quantify (e.g., assuming that temporal averaging of local-scale time series observations is equivalent to spatial averaging over the model grid cell). Moreover, whereas a numerical weather prediction (NWP) model predicts specific weather events, GCMs predict climate. Thus, whether a GCM accurately predicts the cloud field on 10 August over Ohio cannot be addressed directly. Rather, observations must be aggregated over some period of time and over some hori-

Corresponding author address: Dr. Roger Marchand, Pacific Northwest National Laboratory, 902 Battelle Blvd., Richland, WA 99352.
E-mail: roj@pnl.gov

zontal distance to determine the degree to which the predicted clouds match the observed clouds. If a difference exists, it is difficult to determine the source of the problem (i.e., what physical processes or situations are not sufficiently represented by the parameterization) or to determine a corrective action (i.e., how to alter the parameterization).

In this paper we investigate using a classification scheme based on fields resolved by GCMs and NWP models as a means to map the large-scale (synoptic scale) atmospheric state to distributions of local-scale cloud properties. The idea of weather typing or dividing observed weather into states or weather regimes is not new, but has been used extensively in meteorology [e.g., see discussion in Zivkovic and Louis (1992) and Michelangeli et al. (1995)]. Weather typing has been used as a tool to evaluate GCMs and NWP models (e.g., Hewitson and Crane 1992, 1996; Tennant 2003), including cloud properties (e.g., Norris and Weaver 2001; Jakob and Tselioudis 2003; Jakob et al. 2005). In particular, Jakob et al. (2005) exposed shortcomings in the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) data in the Tropics by examining model data as a function of cloud regime. These model shortcomings were not apparent from annually averaged values.

In this analysis, we aggregate local-scale properties of clouds according to the synoptic-scale state. Specifically, we examine vertical mean profiles of cloud occurrence obtained from a vertically pointing millimeter-wavelength cloud radar operated by the U.S. Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) Program at its primary Southern Great Plains site near Lamont, Oklahoma. The mean profile of cloud occurrence is the relative frequency that clouds (or precipitation) are detected by the radar over a fixed period of time and at a given altitude above ground level. We analyze these vertical profiles to assess whether these states are temporally stable and distinct in a statistically meaningful way.

A stable state or class-based map could be of great utility in the analysis of GCM predicted cloud properties. By aggregating and comparing model output with observations according to the atmospheric state, a physical context is provided from which to understand any differences between the model output and observations, as well as to separate differences (in the total distribution) that are caused by having different weather regimes (or synoptic-scale activity) rather than problems in the representation of clouds for a particular regime. Furthermore, if stable mappings can be established, it could form the basis of future model parameterizations, a point to be addressed further at the end of this article.

Section 2 of this paper describes the classification scheme developed in this study, including a brief description of the NWP and cloud-radar datasets. Section 3 provides an overview of the meteorology of the atmospheric states produced by the objective classification scheme and comments on the profiles of cloud occurrence derived from the radar data. Section 4 presents a detailed description of a moving-blocks bootstrap difference test for the similarity of profiles of cloud occurrence with height. Section 5 evaluates both the stability and the distinctiveness of the radar profiles for each atmospheric state using the bootstrap difference test. Although this paper concentrates on testing the similarity of the mean profiles of cloud occurrence with height, the bootstrap technique described here could be expanded to include other statistics and other kinds of atmospheric data. Section 6 discusses further testing and expansion of the technique we hope to undertake in the near future.

2. Classification technique and dataset description

Several approaches have been developed to create atmospheric classification schemes. These approaches include principal component analysis (PCA; e.g., Hewitson and Crane 1992; Hewitson 1994), clustering of PCA components (e.g., Kalkstein et al. 1987, 1990; Zivkovic and Louis 1992; Ye et al. 1995; Michelangeli et al. 1995; Romero et al. 1998), discriminant function analysis (e.g., Kalkstein et al. 1996, 1998), fuzzy logic (e.g., Bardossy et al. 1995; Ozelkan et al. 1998), and neural networks (e.g., Hewitson and Crane 1996).

In this investigation, we use a simple competitive (or self-organizing) neural network (e.g., Kohonen 1995; Haykin 1998) to objectively identify patterns in 17 months of analysis data from an NWP model starting from December 1996. Analysis data are the inputs used to initialize NWP-model forecasts and are obtained through a data-assimilation process that combines model first-guess fields with observations. In this study, we used analysis data from the rapid update cycle (RUC) model, which is run operationally at the National Centers for Environmental Prediction (Benjamin et al. 1991, 1996, 2004a,b). The RUC data were stored in a convenient form over an approximately 600-km by 600-km region (14×12 grid points) centered over the U.S. DOE ARM site in Lamont, Oklahoma, and was readily available through the ARM archive (www.arm.gov). Over this 17-month period, the archive is reasonably complete with RUC analysis data missing for approximately 5% of the total time (in small chunks

ranging from 3 h to as much as a couple of days). In the future, we hope to examine more than 17 months of RUC data, as well as other NWP datasets.

The neural network used in this analysis classifies the atmosphere as belonging to one of 25 possible states, given 6888 input variables (i.e., geopotential height, relative humidity, temperature, and wind at eight predetermined pressure levels, as well as the surface pressure, at all 12×14 grid points). The output of the neural-network classifier consists of a set of 25 synoptic-scale patterns (state definitions) and a state number (from 1 to 25) that indicates the state each input pattern most closely resembles. Generating the 25 synoptic-scale patterns each time the neural-network classifier is used is generally not necessary, rather, once a set is established, it can be used to classify data that were not used in the original training set. Although the number 25 was chosen based on intuition, it appears to be more states than are actually needed, as we will discuss in section 5.

To assess whether our identified atmospheric states might be used to map local-scale cloud observations to the synoptic scale, data from the cloud radar (for the same 17-month period) were used to determine the 3-h probability of occurrence of clouds as a function of height for each of the 25 atmospheric states. The choice of a 3-h period was guided by the RUC analysis dataset, which was available once every 3 h. Our analysis was accomplished using a personal computer and, to reduce computational time, we used 35 vertical levels from the radar data, ranging from 285 to 15 585 m. The mean vertical profile of cloud occurrence for each state is shown in Fig. 1.

3. Meteorology of the 25 atmospheric states

In this section, the 25 synoptic patterns obtained from the neural-network classifier are briefly discussed. The synoptic patterns define the atmospheric state. For example, state 10 (Fig. 2) is characterized by northwesterly winds and cold temperatures at 1000 hPa (upper right panel), southwesterly flow with 30%–60% relative humidity (RH) at 500 hPa (middle left and lower left panels), and dry ($RH < 50\%$) upper levels (lower right panels). In this section, we introduce the 25 states, describing them both by their characteristic synoptic-scale patterns (e.g., Fig. 2) and their vertical profiles of probability of cloudiness (Fig. 1). The synoptic patterns of the 25 states are not shown in this paper for brevity, but a short description of each state is provided in Table 1, and is described further in this section. To facilitate their description, we discuss the states either by the surface-wind direction over the domain (i.e., south, southeast, east, northwest, north, calm) or by a characteristic mesoscale feature within the domain (i.e., cold or stationary front, dryline).

a. South and southeast

States 1, 2, 3, and 14 have surface southerlies over the domain (principally Kansas, Oklahoma, and northern Texas). States 1 and 3 occur in the warm season, with among the hottest surface temperatures of all 25 states. These states typically feature the advection of warm and moist air from the Gulf of Mexico northward through the Plains. State 1 is associated with strong surface southerlies and relatively zonal 500-hPa flow, whereas state 3 is associated with weaker surface southerlies and a weaker 500-hPa flow. States 2 and 14, on the other hand, occur during the cool season. Both are associated with 500-hPa west-northwesterly flow. Of the two, state 14 has cooler surface temperatures and is associated with a cyclone to the west of the domain. All four states with surface southerlies have a typical gradient in moisture that is moist to the south and dry to the north, except state 3, which has more of an east–west gradient (moist to the east).

States 6, 18, and 20 possess surface southeasterlies and bear remarkable similarity to each other in the surface and 500-hPa height and moisture fields. The 500-hPa flow is typically from the west. The characteristic that distinguishes among these three states is the near-surface temperature (Table 1). State 6 typically occurs during the summer with 1000-hPa temperatures about 24°C, state 18 typically occurs during the spring with 1000-hPa temperatures about 18°C, and state 20 typically occurs during the cool season with 1000-hPa temperatures 6°–16°C.

Six of these states (e.g., 1, 2, 3, 14, 18, and 20) exhibit similar vertical profiles of probability of cloudiness, possessing a maximum near 10 km (Fig. 1). Such a maximum around tropopause height probably indicates the presence of jet stream cirrus clouds and anvils associated with nearby thunderstorms, which can occur with the moist southerly winds and westerly flow aloft. State 6 has a near-constant probability of clouds at all-tropospheric levels (Fig. 1), suggesting low- and midlevel clouds also tend to be present.

b. East

States 4, 5, 17, and 24 have surface easterlies. State 4 is a hot and humid summer state underneath a 500-hPa ridge. State 4 has more moisture at upper and mid levels relative to the other dominant summertime states 1 and 3 (which is also reflected in the cloud occurrence profiles in Fig. 1). States 5 and 17 both possess south-

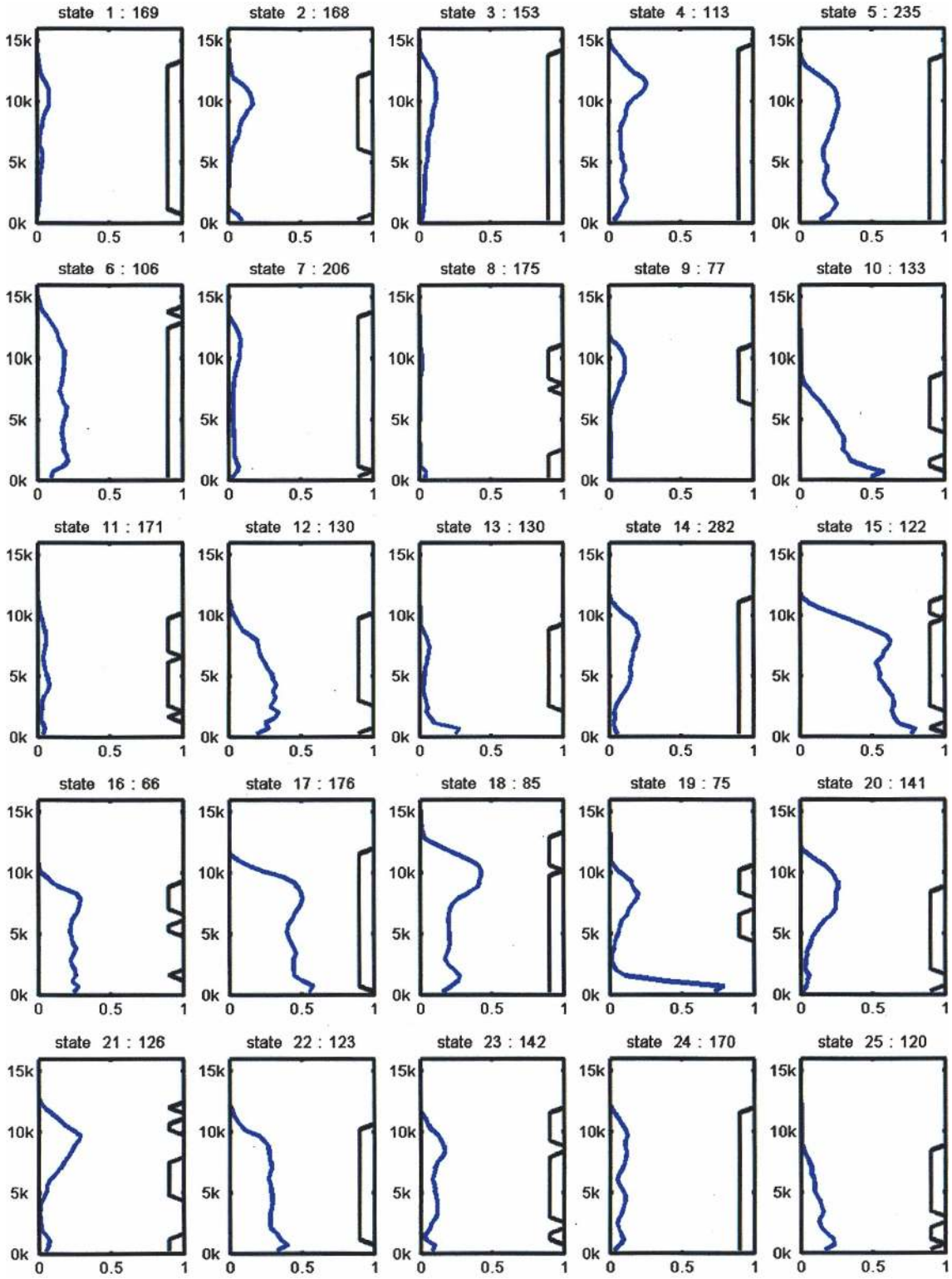

FIG. 1. Vertical profile of the fraction of time a cloud was observed at each radar level (blue solid line) for each of the 25 states. The radar level is on the vertical axis and is given in meters above ground level. The top of each subplot lists the state number and the number of 3-h time blocks in which the atmosphere was classified as being in that state. The solid black line on the right side of each panel signifies the radar levels that can be used in a comparison for that state (i.e., passes the sample size tests; see section 4c).

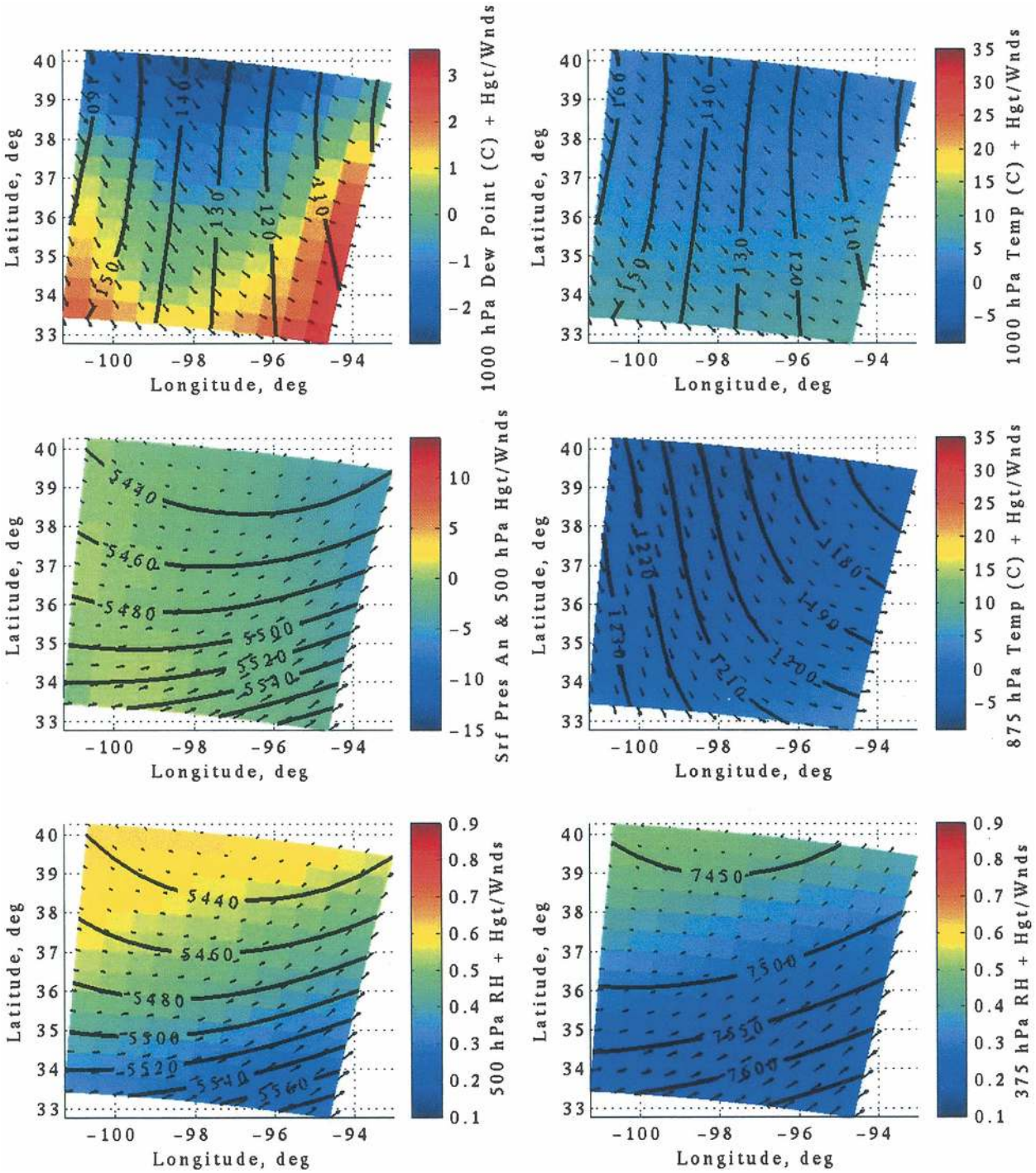

FIG. 2. Synoptic fields for atmospheric state 10. (upper left) The 1000-hPa dewpoint (°C, colored) and 1000-hPa geopotential height (m, black lines). Arrows indicate wind direction and speed. (middle left) Surface pressure anomaly (hPa, colored) with 500-hPa geopotential height (m, black lines) and wind. The surface pressure anomaly is the difference from the 17-month mean value. (lower left) 500-hPa fractional relative humidity (colored) with 500-hPa geopotential height (m, black lines) and wind; (upper right) 1000-hPa temperature (°C, colored) with 1000-hPa geopotential height (m, black lines) and wind; (middle right) 875-hPa temperature (°C, colored) with 875-hPa geopotential height (m, black lines) and wind; (lower right) 375-hPa fractional relative humidity (colored) with 375-hPa geopotential height (m, black lines) and wind.

TABLE 1. Summary of the 25 atmospheric states: state number, months in which they occur, number of 3-h time blocks in dataset, 1000-hPa winds (temperature/dewpoint, both in °C), 500-hPa winds and relative humidity, and synoptic features (e.g., fronts, drylines) and relationship to clouds from Fig. 1. Ranges are given when large gradients exist across the domain.

State	Months	No.	1000 hPa	500 hPa	Notes on synoptic features and clouds
1	May–October	169	South (32/20)	West, 40%	
2	October–May	168	South (15–21/6)	West-northwest, 30%	
3	June–October	153	South (33/22)	Weak, 35%	
4	July–September	113	East (30/19–24)	West, 30%–60%	South side of jet
5	May–October	235	East (23/14–23)	Southwest, 30%–50%	Return flow of moisture
6	June–September	106	Southeast (24/22)	West, 30%–50%	
7	May–October	206	North (23/17)	Northwest, 20%	Cool front, clear skies
8	October–June	175	Calm (15/4–10)	Northwest, 15%	Clear skies
9	November–May	77	Calm (5–12/–2)	West-northwest, 25%	
10	October–April	133	Northwest (5/–3 to 3)	Southwest, confluent 20%–60%	Shallow postfrontal clouds
11	October–May	171	Northwest (2/–3 to −10)	West, 30%	Clear skies
12	October–April	130	North-northeast (2/0 to −10)	West-southwest, 55%	
13	October–April	130	Northwest (5/0 to −5)	North-northwest, 25%	
14	November–May	282	South west–west (5–15/–2 to 3)	Northwest, 30%–50%	Lee cyclone
15	October–April	122	North, confluent (10/2–9)	Southeast, 60%–80%	Cutoff low, surface cyclone, abundant clouds within troposphere
16	November–April	66	North (10/2–9)	West-southwest, 50%	
17	October–April	176	East (5–15/2–12)	Southwest, 60%	Warm front, abundant clouds within troposphere
18	March–June	85	Southeast (18/12–18)	West, 40%–60%	Lee cyclone southwest
19	December–March	75	North-northeast (2–12/–5 to 10)	Southwest, confluent, 30%	Shallow postfrontal clouds
20	October–May	141	Southeast (6–16/–3 to 5)	West-southwest, 30%–50%	Lee trough southwest, high clouds
21	October–May	126	South, confluent (13–23/7–14)	Southwest, 20%–45%	Dryline, high clouds
22	October–May	123	South, confluent (13–22/9–16)	South-southwest, 35%–65%	Dryline
23	October–May	142	South–northwest confluent (8–23/2–11)	West-southwest, 30%–50%	Southwest–northeast-oriented cold/stationary front
24	October–May	170	East (12–22/3–10)	West-northwest, 35%	
25	October–May	120	Northwest (2–11/–2)	Northwest, 20%–60%	

westerlies at 500 hPa, although state 5 is much warmer than 17 (23°C versus 5°–15°C at 1000 hPa, respectively). State 17 is relatively moist within the troposphere and exhibits near-constant probability of clouds with height (Fig. 1), suggesting deep ascending air associated with a warm front, consistent with the near-surface easterlies. State 24 exhibits a 1000-hPa temperature intermediate to the other states (~17°C), has 500-hPa westerlies, and is quite dry at all levels. The dryness is consistent with the small probability of clouds with height (Fig. 1).

c. Northwest and north

States 10, 11, 13, and 25 exhibit surface northwesterly flow, and states 7, 12, 15, 16, and 19 exhibit surface northerly or north-northeasterly flow. Such flow patterns typically occur after the passage of an equatorward-moving cold or arctic front (e.g., Mecikalski and Tilley 1992; Colle and Mass 1995; Schultz et al. 1998; Schultz 2004) or an eastward-moving cold front originating from the Pacific (e.g., Hobbs et al. 1996; Neiman and Wakimoto 1999).

The northwesterly states all occur during the cold season with the coldest 1000-hPa temperatures around or less than 5°C. The primary distinguishing factor between the four northwesterly states is the 500-hPa flow: confluent southwesterly (state 10), westerly (state 11), north-northwesterly (state 13), or northwesterly (state 25).

On the other hand, the northerly states are distinguished primarily by near-surface temperature. State 7 occurs during the warm season when equatorward-moving fronts occasionally bring warm, dry air to relieve the Southern Plains' summer heat and humidity (often colloquially referred to as cool fronts). States 12 and 19 exhibit among the coldest 1000-hPa temperatures (as low as 2°C), whereas states 15 and 16 exhibit 1000-hPa temperatures around 10°C (Table 1). The principal distinction between states 12 and 19 is that

state 12 is much more moist than state 19 at 500 hPa (RH is about 55% versus 30%, respectively; Table 1). The difference between states 15 and 16 is that the 500-hPa flow has a closed circulation and is relatively moist (RH 60%–80%) for state 15. In contrast, state 16 exhibits a relatively dry (RH about 50%), zonal 500-hPa flow.

States 10, 12, 13, 15, 19, and 25 are similar in that they have a maximum of cloud fraction in the lower troposphere (Fig. 1), all indicative of low-level postfrontal cloudiness. In contrast, the westerly 500-hPa flows of states 11 and 16 are associated with a relatively dry troposphere, consistent with the small frequencies of clouds observed for those states (Fig. 1).

d. Calm winds

States 8 and 9 are characterized by relatively weak surface winds across the domain. Both states are similar, with dry 500-hPa northwesterly flow and no appreciable cloudiness (Fig. 1). The principal difference appears to be the 1000-hPa temperature (about 15°C for state 8 versus 5°–12°C for state 9).

e. Cold or stationary front

State 23 exhibits a surface cold or stationary front, oriented southwest–northeast across the domain. Air is more moist and warm to the southeast and more dry and cool to the northwest, consistent with its tendency to occur during cool season. See Schultz (2004, especially his Figs. 4b, d) for more discussion of state 23.

f. Dryline

States 21 and 22 are characteristic of periods where a strong dryline is present. Both feature strong low-level moisture gradients across northern Texas and Oklahoma and southwesterly–westerly flow at 500 hPa. The principal difference between states 21 and 22 appears to be the amount of 500-hPa moisture; state 21 is dry (RH 20%–45%), whereas state 22 is more moist (RH 35%–65%). The relatively dry air in state 21 inhibits cloud development in the mid and lower troposphere relative to that of state 22 (Fig. 1).

4. A bootstrap test for the similarity of profiles of cloud occurrence with height

Given these 25 synoptic states, we would like to know if the cloud-radar profiles of probability of cloudiness associated with each state are temporally stable and distinct in a statistically meaningful way. In this section, we develop techniques to answer this question.

In the statistical analysis of atmospheric properties, comparing datasets from different sources, either from different times or locations, or comparing observational data with model output is often desirable. Although differences in some cases can be obvious and readily understood based on physical reasoning, in other cases, whether or not differences are statistically significant may not be clear. Many of the standard statistical tests of differences (e.g., the Student's t test) assume that data points are independent of each other and that the underlying distribution of the sample is known. As with the t test, these standard statistical tests are frequently built upon asymptotic behavior and the Central Limit Theorem to identify distributional properties of the statistic. These assumptions limit the applicability of such tests to many remote sensing observations from the atmosphere (such as the radar data examined in this paper) where both strong spatial and temporal correlations exist, in addition to unknown (and likely non Gaussian) underlying distributions. We account for the spatial and temporal correlation in this analysis using a moving-blocks bootstrap resampling approach (e.g., Efron and Tibshirani 1993; Wilks 1997).

The underlying principle of the bootstrap resampling approach is that, given a sample **X** from a parent population with an unknown distribution, statistics of interest for the true parent population **F** can be calculated by taking a series of resamples from **X**. The key is that the resampling is done from the original sample with replacement, thus imitating as closely as possible the act of sampling from the parent population. For the bootstrap approach to work, the original sample set **X** must have a sufficient number of independent samples to represent the true population **F.**

In the following section, we describe a bootstrap difference test that compares the probability of cloud occurrence at a single radar altitude (section 4a), extend this basic technique to handle vectors (or profiles) of radar observations (section 4b), and discuss further the minimum number of independent samples needed to undertake a comparison (section 4c).

a. Single-variable bootstrap difference test

Assume two vectors, **Y** of length n_y and **Z** of length n_z. We want to know if the distributions of the data in **Y** and **Z** are the same (or more formally stated, if they come from a common parent population). We want to formulate a hypothesis test to determine if a significant difference exists between **Y** and **Z**.

In the bootstrap approach, given a sample of n data values $\mathbf{X} = \{x_1, x_2, x_3, \ldots x_n\}$ from an unknown parent population **F**, a bootstrap resample $\mathbf{X}^*$ of **X** is created by randomly choosing n elements from **X** with replace-

ment. With replacement means that after an element is selected, it remains in **X** and can be chosen again. Essentially, the sample **X** is treated as if it were the true population **F**, and **X*** is then a new sample (or new realization) drawn from it. A large number of bootstrap resamples **X*** are taken, and a statistic of interest is calculated for each resample. The distribution of a statistic from the many resamples can then be used to create confidence intervals (e.g., the 95% confidence interval) for that statistic.

In our application, $\mathbf{x}_i$ is not a single value, but a vector of cloud fraction (where each element in the vector is the cloud fraction at one radar layer) at each 3-h time step (Fig. 3). By keeping all the elements of $\mathbf{x}_i$ together, the bootstrap technique maintains all spatial (i.e., vertical) correlations, but not the temporal correlation in the data. (Section 4b discusses how to apply a scalar bootstrap technique to the vector of the radar data.) If the time series of **X** is resampled to create many time series of **X***, we ignore any temporal correlation between $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, effectively treating them as independent samples.

To account for the temporal correlation, we use a moving-blocks bootstrap approach. This approach is similar to the original bootstrap in that it creates a set of new samples **X*** from the original sample set **X**, but instead of sampling elements from **X** with replacement, the moving-blocks bootstrap approach samples blocks of consecutive elements from **X** with replacement. Figure 3 illustrates this process. In this example, a series of bootstrap samples with a fixed block size of 4 is created from an original set of length 12. By choosing blocks of elements, much of the temporal correlation in the time series is preserved.

The bootstrap approach provides a good framework for hypothesis testing. Taking a large number of bootstrap resamples yields an approximate distribution of what a test statistic should look like for the true population, and we can reject the null hypothesis if the test statistic for the original sample is, for example, outside of the 95% confidence interval. In our application, we take the null hypothesis to be that two different sets (of cloud occurrence with height data) are from the same underlying distribution (i.e., the two sets can be thought of as two realizations from the same unknown population). The remainder of this subsection describes a test of this null hypothesis for a single radar layer, whereas section 4b describes a modified test to account for all of the radar layers.

Although we want to know whether the data in **Y** and **Z** come from a common parent population, no statistical test on a finite dataset can definitely prove this

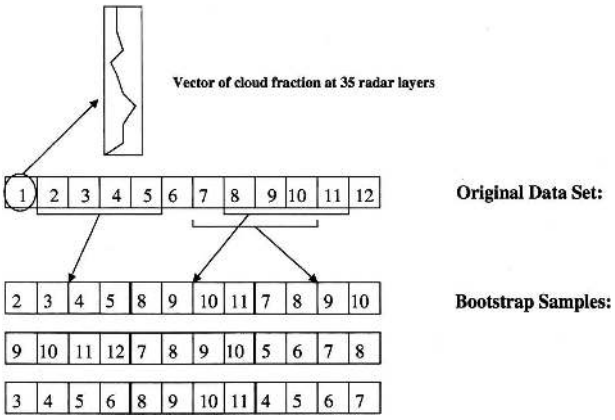

FIG. 3. Diagram of moving-blocks bootstrap with replacement [adapted from Wilks (1997)].

statement. Rather, we assume they are the same, and determine if this statement appears unlikely. Any two finite samples (**Y** and **Z**) are unlikely to have precisely the same mean values ($\langle\mathbf{Y}\rangle$ and $\langle\mathbf{Z}\rangle$) even when they derive from a common parent population. Consequently, we examine if the difference in the mean ($\langle\mathbf{Y}\rangle - \langle\mathbf{Z}\rangle$) is unlikely to be a result of having a finite sample size. If the difference in the mean is unlikely, **Y** and **Z** are probably not from the same parent population, and we reject the null hypothesis. (That two different atmospheric states will have the same mean probability of cloud occurrence at some given altitude and yet come from different parent populations is possible, however, we will ultimately apply the test at many levels and it is unlikely that two distinct states would have the same mean cloud fraction at all levels.)

The test described here for a single radar layer follows Efron and Tibshirani (1993). The two sets, **Y** and **Z**, are combined into a single set **X**, where $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$, and **X** has length $n = n_y + n_z$. Next, a large number n_b of bootstrap resamples **Y*** of length n_y are taken from **X**, and n_b resamples **Z*** of length n_z are also taken from **X**. That is, each **Y*** has the same length as **Y**, but from the total population **X**, and similarly for **Z***. Typically, n_b is several thousand. To perform this resampling, we randomly select an element x_i from **Z** and take a block of elements of some length L containing x_i as an element of the new sample. Leaving those elements in the original sample (since we are resampling with replacement), we repeat the process until the new sample is the same length as the original sample of **Y**.

For each of the n_b resample pair, the following statistic (standard difference) is calculated

$$d^* = \frac{\bar{z} - \bar{y}}{\bar{\sigma}\sqrt{1/n_z + 1/n_y}}, \tag{1}$$

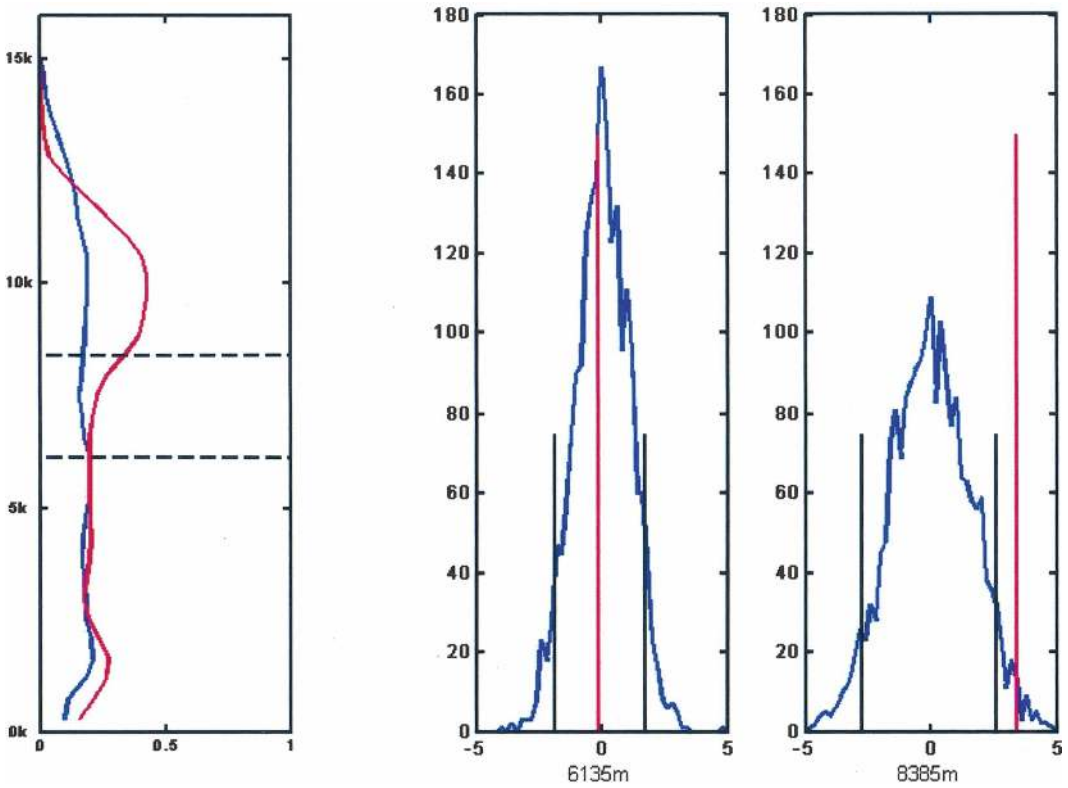

FIG. 4. (left) The mean radar profile of cloud occurrence for state 6 (blue line) and state 18 (red line). The two states have similar mean cloud occurrence below about 6 km, but state 18 has more cloud between about 6 and 12 km. Is the difference in the mean cloud occurrence statistically significant? (right) The d statistic (red line) in relation to the d^* distribution (blue line) for two radar layers (the altitude of these layers is indicated by the dashed black lines in the left panel). The short black lines mark the boundaries of the 95% confidence interval for the d^* distributions; (middle) (radar altitude = 6135 m) the difference in the cloud occurrence is not significant, whereas (rightmost) (radar altitude = 8385 m) the difference test shows that the two states contain a different local-scale cloud occurrence.

where $\bar{z} - \bar{y}$ is the difference of means between $\mathbf{Y}^*$ and $\mathbf{Z}^*$ and

$$\bar{\sigma} = \left(\frac{\sum_{i=1}^{n_z} (z_i - \bar{z})^2 + \sum_{j=1}^{n_y} (y_j - \bar{y})^2}{n_z + n_y - 2} \right)^{1/2},$$

which is the estimated standard deviation.

This process produces n_b values of d^*, forming a distribution of the test statistic from the bootstrap resamples. This same difference statistic for the original sample sets $\mathbf{Y}$ and $\mathbf{Z}$, which we call d, can be calculated, similarly. We can determine if d is an unlikely or extreme value by comparing the value of d to the distribution of d^* and seeing where it falls in relation to the 95% confidence interval. If d falls outside of the 95% interval, we reject the null hypothesis that the two sets come from the same distribution. Figure 4 shows two examples of comparing the distribution of d with d^*: one where d falls outside the 95% confidence interval, so that we reject the null hypothesis and state that the radar observations (at this one altitude) are likely from different populations (Fig. 4c), and one where d falls within the 95% confidence interval so we do not reject the null hypothesis and claim the radar observations may be from the same population (Fig. 4b).

A remaining issue is selecting the size of the block length L. The block length needs to be large enough to capture the temporal correlation structure of the time series data. If too small a block length is used, the bootstrap blocks are treated as being independently sampled when in fact they are not, leading to the bootstrap underestimating the variability in the distribution of the d^* statistic. In such a case, the test is permissive, meaning the null hypothesis is rejected more often than is warranted. The choice of block length is, in general, an outstanding problem for most moving-block bootstrap applications. Wilks (1997), for example, discusses

several methods for determining an optimal block length. These methods, however, all involve fitting a first- or second-order autoregressive model to the time series, neither of which fit our radar data well. We also recognize that a single common block length will likely not fit all the atmospheric data. The goal in dividing the original radar data into a finite set of classes was to map this data according to its synoptic regime. Different regimes are likely to contain different scales of temporal correlation leading to different nominal block lengths for each state. Therefore, instead of choosing a fixed block length, we use a block-selection method that adjusts to the correlation scale of the current data.

In our application, the original time series of radar observations is broken into 25 disjoint time series, prior to application of the bootstrap difference test (section 2). When the neural network classifies the RUC data into different states, a set of contiguous elements in one state represents observations, which may come from a single synoptic event, so the data may well be correlated. We expect most of the temporal correlation occurs entirely within that pattern. Therefore, in the bootstrap test, after randomly selecting $\mathbf{x}_i$ from the combined dataset $\mathbf{X}$, we choose as our block length the largest possible set of elements $\mathbf{x}$ such that all the elements are in the same class and are contiguous in time. So, when making a bootstrap resample for state 21, for example, we first randomly pick a point in our time series of atmospheric states in which the atmosphere was determined (by the classifier) to be most like state 21. We then see how many consecutive elements in time (adjacent to the randomly chosen one) are also in state 21. All of these elements are then placed into the resample. This is a conservative choice in the sense that we are likely erring on the side of keeping more elements together than may be needed. This choice reduces the ability of the test to differentiate between our radar profiles, but should also minimize false detections. In our analysis, we found the median contiguous block length ranged from 1.5 to 6 (4.5 to 18 h) depending on the state, with the summer states generally having the shorter block lengths.

b. *Extension of the bootstrap technique to a vector of observations*

The bootstrap test described thus far can be applied to the radar observations at any single altitude. However, we want a test that considers the entire vertical cloud profile (all the radar layers) when deciding whether or not to reject the null hypothesis. A simple test, in principle, is to run the single-layer test, calculating a d and d^* distribution for each radar layer. If d is unlikely to have occurred in any layer, we could then reject the null hypothesis and claim that $\mathbf{Y}$ and $\mathbf{Z}$ are not the same because they are not the same for every vector element. In practice, the difficulty with this approach is that there may be many vector elements in the dataset (e.g., 35 levels in this case) so that a false positive (which we would expect to occur 5% of the time using a 95% confidence interval) becomes likely in at least a few of the elements almost all of the time. The data are also spatially correlated, so we cannot treat the results of the individual test as independent. Rather, we need to consider all the layers simultaneously. We have examined doing this in two ways.

The first approach follows from Wilks (1997) and uses a method of summing the test statistics across all the radar layers. As shown before, the d statistic (for the original sample) and d^* statistic (for each of the n_b bootstrap resamples) are calculated for each radar layer. From the layer values, one then calculates a global statistic k,

$$k = \sum_i |\, d_i \,| \text{ (original data)}$$

$$k_j^* = \sum_i |\, d_{ij}^* \,| \text{ (resampled data)}, \tag{2}$$

where i ranges over the 35 radar levels, and j ranges over the n_b bootstrap resamples. We then compare k to the distribution of k^* and reject the null hypothesis if k is an extreme value falling outside the 95% confidence interval of the distribution. We refer to this test as the sum-of-absolute-values multivariate test.

The second test we evaluated creates a similar global statistic, but based on counting the number of single layer tests where d is unlikely at the 5% level. That is,

$$k = \text{number of radar layers where } d \text{ is 5\% or less likely,}$$

and

$$k_j^* = \text{number of radar layers where } d_j^* \text{ is 5\% or less likely,} \tag{3}$$

where j ranges over the n_b bootstrap resamples. Similar to the first test, we compare k to the distribution of all k_j^* and reject the null hypothesis if k falls outside the 95% confidence interval. We will refer to this as the *number-of-unlikely-radar-layers* multivariate test.

For both of the above tests, we also calculate a significance level, or p value, for each comparison by comparing k with k^*. The p value is the percent of bootstrap resamples that are as extreme as, or more extreme than, the original sample. We reject the null hypothesis if the calculated p value is less than 0.05.

To evaluate these two multivariate tests, we first examined what happens when we compare a given set of

radar profiles against themselves. For this basic test, we randomly divided the radar data in each state into two disjoint subsets of as close to equal size as possible and ran the comparison, testing the two halves against each other. Since the two sets we are testing are drawn from a single parent population, we expect to be unable to reject the null hypothesis that the sets are distinct. For both multivariate tests, the p values for each of the 25 states comparisons are above the 0.05 threshold (Fig. 5). Thus, our two tests are unable to distinguish between the two halves of each state. [We ran this test several times (not shown), randomly dividing the data in each state into two halves each time, and found that, as one might expect, we did occasionally generate one or two false detections (states where we reject the hypothesis at the 5% level, even though we know the data come from a common parent population).]

While the results are similar between the two tests, the p values are quite different. For the sum-of-absolute-values test (Fig. 5a), the p values are fairly evenly distributed throughout the range from about 0.1 to 1. Few of the tests have a p value close to one, even though the two profiles come from the same parent population. In the number-of-unlikely-radar-layers multivariate test (Fig. 5b); on the other hand, almost all the p values have a value of 1.0, with a handful of points having low values. This occurs because, in the number-of-unlikely-radar-layers test, most of the k^* resamples have zero layers, which appear unlikely at the 5% level, a few resamples (typically 5%–20%) have 1 layer unlikely at the 5% level, and fewer yet with two or more unlikely layers, such that the distribution of $k^* =$ *number-of-unlikely-radar-layers* is very sharply peaked at zero. Thus, whenever we find $k = 0$ (not having any layers unlikely at the 5% level in the original samples) this leads to a p value of 1. When it happens that $k = 1$ or $k = 2$ (by random chance in this case), the p value is far from 1 (Fig. 5b).

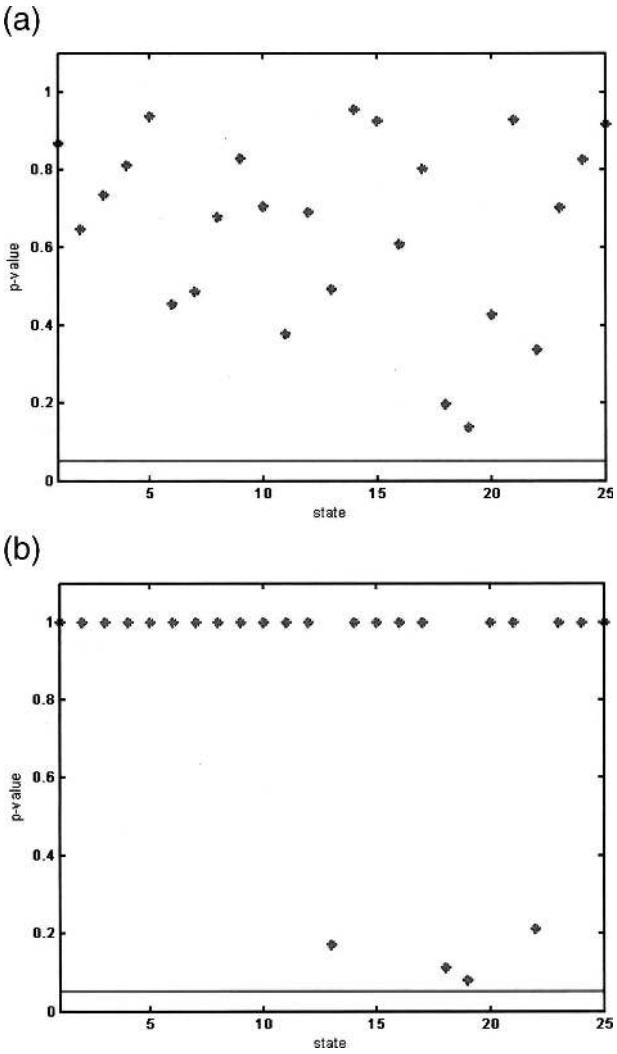

FIG. 5. Test of each state vs itself (a) using $k =$ *sum-of-absolute-values* multivariate test; (b) using $k =$ *number-of-unlikely-radar-layers* multivariate test. All the p values are above the 0.05 threshold, so we are unable to claim that any of the states are divided into distinct halves.

c. Sample-size constraint

In general, a requirement of the bootstrap resampling approach is that the initial sample must represent the underlying population. If the initial sample is too small, the parent population is not likely to be represented accurately by the resampling. Such undersampling can result in the bootstrap either underestimating or overestimating the variability in the distribution of the d^* statistic. Underestimating leads to a test that is permissive (i.e., rejecting the null hypothesis when it should not) and overestimating the variability leads to a test that is too stringent (i.e., not rejecting the null when it should). Hypothesis tests used on a small sample size tend to be unable to detect many real differences because the variance due to sampling will tend to be large. In general, if the null hypothesis cannot be rejected we must make sure to identify those cases where real differences are unlikely to be detected based on the sample size before we claim the data are from the same or similar parent populations.

It is not the actual number of samples that needs to be considered, but rather the number of independent data points, sometimes called the effective sample size. Using a difference test similar to the one described in this paper, Wilks (1997) suggests a minimum effective sample size of 10, where the effective sample size, n_e, is estimated using

$$n_e \cong n \frac{1 - \rho_1}{1 + \rho_1}, \quad (4)$$

and ρ_1 is the lag-1 autocorrelation coefficient of the time series.

In our application of this methodology, we found that this constraint was not completely sufficient because we found cases where a set consisting primarily of data points of value zero and only a few cloud detections (nonzero values) will pass this sample-size test. The variance of the data is not sufficiently well estimated in this circumstance to use the hypothesis test. Thus, we also required the dataset to have more than 10 nonzero elements.

We apply these constraints to the combined dataset **X**, rather than individually to the two sets being compared (**Y** and **Z**). This approach permits comparing situations where a given layer may be void of clouds in set **Y**, but has more than a sufficient number of samples for that layer in set **Z**. If the minimum sample-size constraint were applied to the individual sets first, then any region with few clouds in either set would not be compared, even though the lack of cloud in one set compared with a high cloud fraction in another set may itself be very meaningful.

We stress that the estimated value of n_e is not being used in the difference test, but only as a measure of when it is safe to apply the bootstrap difference test. In the results presented in this paper, we opted to use an effective sample size threshold of 10 as suggested by Wilks. In general, we found that comparisons with data with an effective sample size below 10 have little resolving power, meaning that data with seemingly large differences in the mean value do not cause the hypothesis to be rejected (because with such a small number of samples large deviations between the means are likely). We will discuss one such example in the next section.

5. Stability and distinctiveness of the atmospheric states

In order for the 25 synoptic-scale states to serve as a map to local-scale cloud properties, the distribution of the local-scale variables must be statistically stationary. Thus, for a given state, the local-scale cloud properties should be the same regardless of the year in which that state occurs. To evaluate the temporal stability of cloud occurrence for our identified states, we compared the radar profiles in two wintertime periods using the bootstrap test described in section 4. For each state, we took all the data elements from winter 1996–97 (1 December 1996 to 1 March 1997) and compared them to the elements from winter 1997–98 (1 December 1997 to 1 March 1998). Figure 6 shows 25 plots (one for each state) of the probability of cloud occurrence versus height during winter of 1996–97 (solid line) and winter 1997–98 (dashed line). The legend for each atmospheric state shows the percentage of each winter that was occupied by that state. For example, state 14 occurred 7.14% of the time in winter 1996–97 but 17.88% of the time in winter 1997–98 (Fig. 6). Some states do not occur in winter (i.e., states 3, 4, 5, 6, 7, and 18). Interannual variability can be significant. Specifically, a La Niña occurred during winter 1996–97, whereas an El Niño occurred during winter 1997–98. Such large-scale circulation changes are known to affect the local weather, for example, favoring more cold-frontal passages in the south-central United States (e.g., Schultz et al. 1998). The considerable interannual variability observed in many cloud properties hampers comparing GCM climate predictions with observations and is the motivation behind this research.

Many of the states show similar profiles of cloudiness between the two winters (e.g., states 11, 15, 17, and 25), while other appear quite different (e.g., states 10, 12, 16, 19, and 20). Most of the states showing poor agreement are composed of a small number of members. For example, state 12 is only occupied 0.73% of the time during winter 1997–98 and state 19 is only occupied 1.04% of the time during winter 1996–97.

Using the moving-blocks bootstrap difference test for the means, we can examine the similarity of the cloud profiles for these two seasons. Our null hypothesis is that the two profiles are the same (or rather from a common distribution) and we test the mean value with height to determine if this is unlikely. If the cloud occurrence profiles are not different, this suggests that we may have obtained a useful map from the large-scale resolved variables to the small-scale cloud properties. Subsequently, this state would be useful to compare GCM output and observational data.

Figure 6 shows that, at the 5% level of significance, only in state 10 can we reject the hypothesis that the profiles are the same. The p values shown in Fig. 6 are based on the sum-of-absolute-values multivariate test. We also found that only in state 10 can we reject the hypothesis using the *number-of-unlikely-radar-layers* multivariate test, although the p value differs as one would expect (not shown).

Of the states that are occupied during the winter, several (states 8, 12, 16, 19, and 24) have insufficient data to make a comparison at any radar altitudes. That is, these states do not pass the minimum sample-size criteria discussed in the previous section at any altitude. Several states (9, 20, 21, and 22) have sufficient data to make a good comparison at only at a limited number of

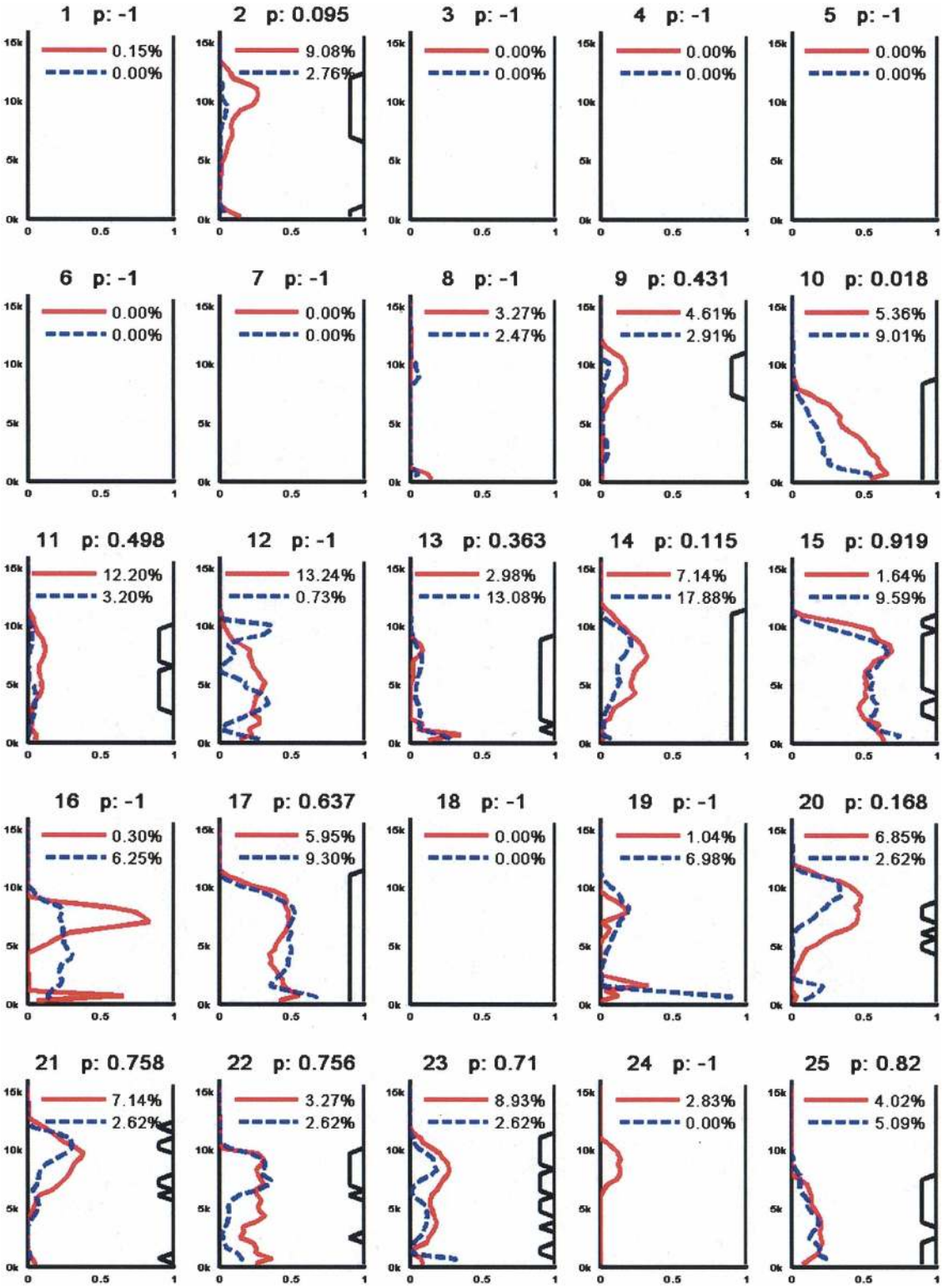

FIG. 6. Vertical profile of the fraction of time a cloud was observed at each radar level for each of the 25 states during winter 1996–97 (red line) and winter 1997–98 (blue dashed line). The legend shows the percentage of each winter occupied by that state. The *p* values for the comparisons between the two winters are also listed. The solid black line on the right side of each panel signifies the radar levels that can be used in a comparison for that state (i.e., passes the sample size tests in section 4c).

altitudes (as denoted by the solid black line on the right side of each plot). Specifically, for state 22, winter 1996–97 has greater cloud occurrence below 5 km than winter 1997–98. However, no comparisons are done at radar altitudes below 5 km (for the purpose of calculating the p value shown in Fig. 6) because the number of independent samples in this region is estimated to be less than 10. Using Eq. (4) these altitude bins are estimated to have between 5 and 9 independent samples. However, lowering our threshold for the minimum number of independent samples does not change the result significantly, with the p value changing from 0.756 to 0.5. As discussed in section 4, comparisons with a small number of independent samples have little ability to determine if a difference is significant.

Of the states that are well occupied (10, 11, 13, 14, 15, 17, and 23), only in state 10 does the difference in the radar profiles between the two winters appear significant. We stress that finding well populated and stable mappings from the large-scale synoptic patterns to the local-scale cloud properties will be useful for comparisons of model output with observed data, even if less than 100% of all atmospheric conditions can be mapped successfully.

For all of the well-occupied states, we compared the average atmospheric state variables (e.g., geopotential heights, winds, and temperatures) for the two winters against the state definitions created by the neural-network classifier. We found the average values for both winters to be similar to the state definition for all of the states, except state 10. For state 10, the averages for winter 1997–98 (not shown) were very close to those of the state definition (Fig. 2), but the averages for winter 1996–97 (Fig. 7) were quite different from the state definition (Fig. 2). Although the fields of wind and pressure for winter 1996–97 were similar to the state definition, winter 1996–97 featured higher 500-hPa relative humidity (lower left panel), higher surface pressure (middle left panel), and lower 1000-hPa dewpoint temperatures (upper left panel) than the state definition. The higher 500-hPa relative humidity in winter 1996–97 was associated with more midlevel cloudiness (Fig. 6).

This example shows an important limitation of using an objective classifier to create the state definitions: a given state may be too broad and encompass too wide a range of conditions. On the other hand, this example also shows the value of examining the local-scale data (not used in the classification process) to identify when this occurs, a point to be discussed further in section 6.

To examine how well the neural network separated the atmosphere into distinct states with distinct cloud profiles, we compared the profiles of cloud occurrence for each state (Fig. 1) to that of every other state, again using the bootstrap test. The set of p values for all these comparisons are summarized in Fig. 8 for the two multivariate tests. For a majority of the state-to-state profile comparisons, the p value falls below the 0.05 significance level (green boxes). Thus, we can reject the null hypothesis and claim the states are pairwise distinct.

For each multivariate test, about 40–45 comparisons had large p values (red boxes), so we are unable to reject the suggested null for these comparisons, suggesting these states may not be distinct from one another. In about 10–15 more comparisons the p values fall slightly above the 0.05 cutoff (yellow boxes). Not surprisingly, many of the state-to-state comparisons that suggest the states are not distinct involve those states with the fewest points. These comparisons suggest that, although the neural network did a reasonable job in identifying states containing distinct atmospheric conditions, some states may be too narrowly defined, at least in the context of the 17 months of data analyzed.

6. Conclusions and future plans

In this paper, an evaluation of objectively identified atmospheric states obtained from NWP large-scale atmospheric fields was presented. The evaluation was based on aggregating profiles of cloud occurrence, obtained from the U.S. DOE ARM program cloud radar in Lamont, Oklahoma, as a function of an objectively determined set of atmospheric states. These profiles were examined for stability and distinctiveness using a bootstrap resampling comparison test.

The bootstrap comparison technique was drawn largely from Wilks (1997). The most significant departures from Wilks are that 1) we follow Efron and Tibshirani (1993) in combining the initial datasets into a single combined set from which to create the bootstrap resamples, 2) we apply two minimum sample-size constraints to the data on a vector-element by vector-element basis, 3) we use a conservative block-selection method that adapts to the temporal correlation structure of the dataset, and 4) we examine a number-of-unlikely-layers multivariate test. Combining the datasets prior to resampling effectively maximizes the number of available samples, which was an important consideration in our application because many of the sample sets contained few points. Problems could occur with this approach if the two datasets are in fact drawn from a different population, if one of the initial datasets is much larger than the other and the overrepresented population is substantially more or less variable than the underrepresented population. In general, it may be

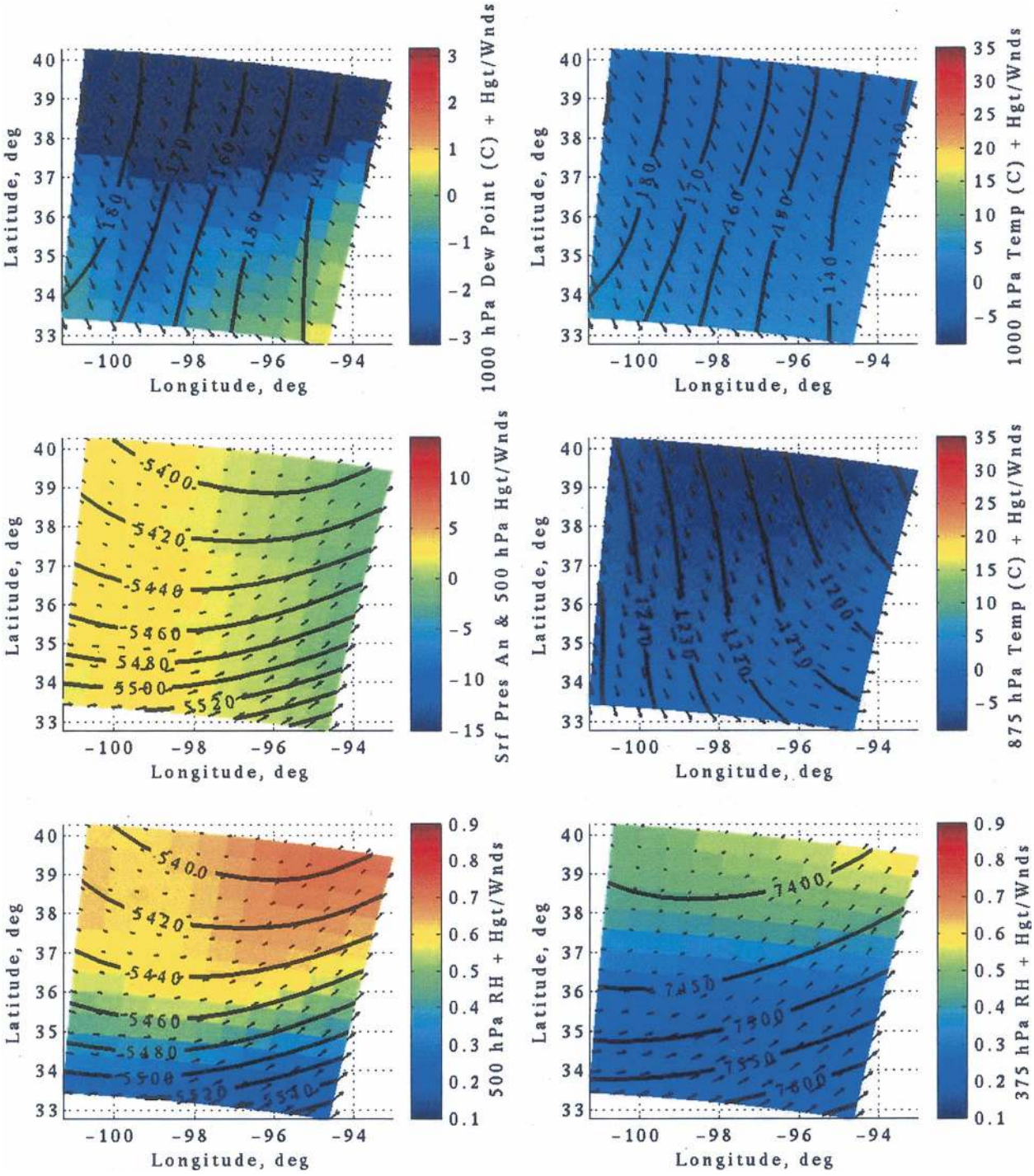

FIG. 7. Same as Fig. 2 but only for winter 1996–97. Winter 1996–97 shows similar wind and pressure patterns to those in Fig. 2, but features (lower left) higher 500-hPa relative humidity, (middle left) higher surface pressure anomaly, and (upper left) lower 1000-hPa dewpoint temperature.

worth testing the individual sets and combining them only when the number of effective independent samples appears low. For the data in this paper, we found no significant differences existed between the combined and noncombined approaches when both sets did have a sufficient number of samples, but where one set had many more samples than the other.

The stability of the atmospheric states was examined

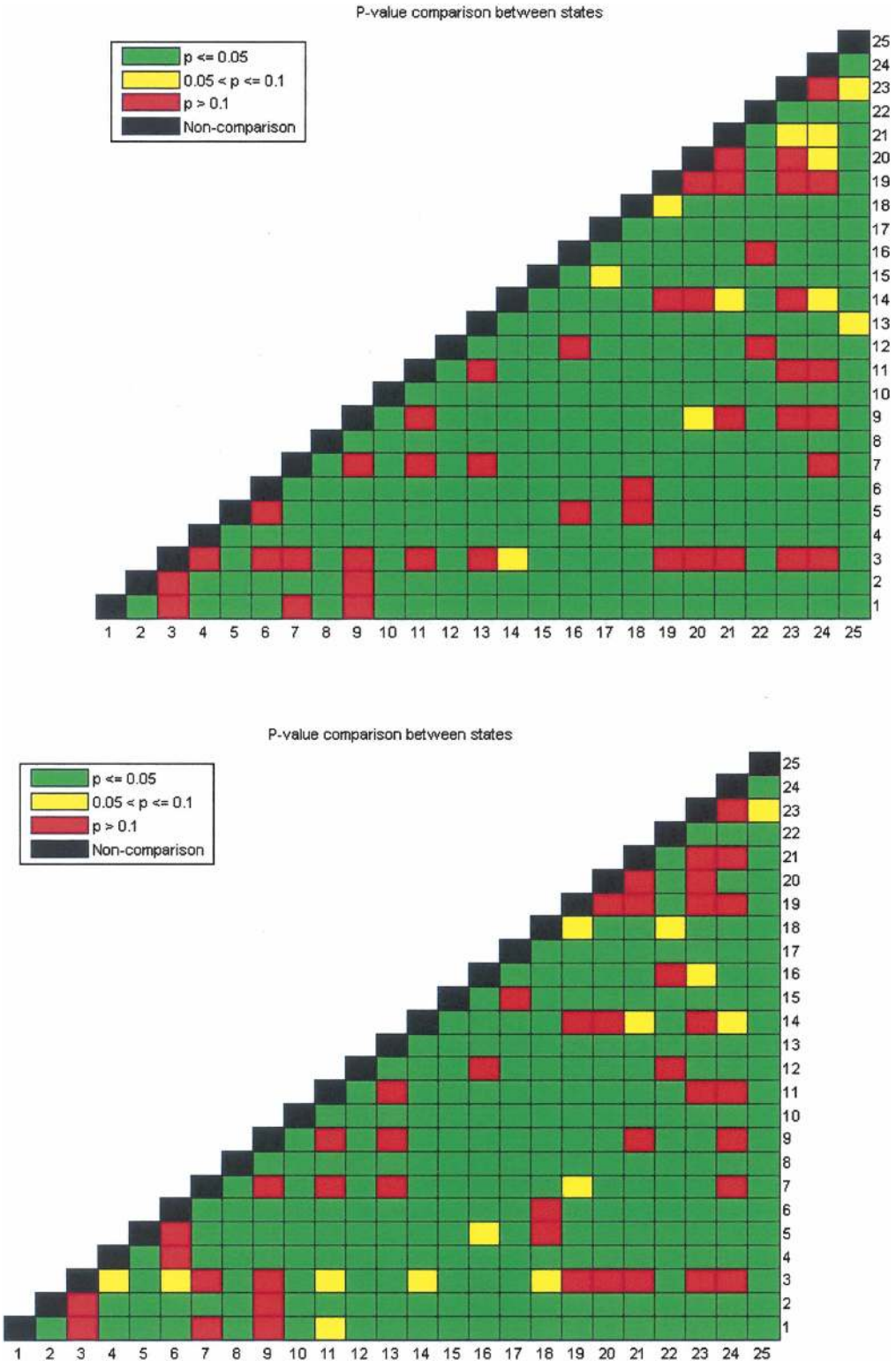

FIG. 8. A summary of the results of each state-to-state comparison. The null hypothesis is rejected for p values less than 0.05, implying that the radar profiles are likely to be distinct, (top) using k = *sum-of-absolute-values* multivariate test and (bottom) using k = *number-of-unlikely-radar-layers* multivariate test.

by comparing the radar profiles obtained during winter of 1996–97 with those from 1997–98. For these two winters, of states that had sufficient data, only one state was shown to have a statistically significant change in its profile of cloud occurrence. This result suggests that large-scale atmospheric fields of the type produced by NWP models and GCMs can be mapped in a stable and statistically meaningful way to distributions of local-scale observations of cloud properties, at least much of the time. Such a mapping could be of great utility in the analysis of GCM-predicted cloud properties. Although the results shown are extremely encouraging, more months of data than the 17 analyzed here need to be evaluated. Such an evaluation could also benefit significantly from analyzing other local-scale observables (e.g., cloud water content or particle size).

One of the difficulties faced by all objective classification schemes is how to determine the optimal number of classes. We are currently studying how best to combine or divide (e.g., in the case of state 10) the states based on the bootstrap test. In a more general sense, we are proposing an approach whereby one intentionally chooses too many classes, and then allows the local-scale data (which is not used in the classification) to drive the determination of what is or is not a useful class. In a similar manner, the bootstrap comparison of the local-scale data could be used as a mechanism to compare the various classification techniques, as well as to examine the influence of different inputs on the neural-network classifier.

In this paper, we initially chose 25 states. After comparing all the state profiles to each other (Fig. 8), some of the profiles were not statistically distinct. In retrospect, a longer time series was probably needed. Based on our initial results, we speculate that 10 to 15 states are required to represent the cool season adequately. This dataset did not permit us to test the temporal stability of the warm season states, and hope to test such in the near future. We expect that summer states will be stable but also recognize that these will be fewer in number and will have more inherent variability (at least in terms of cloud occurrence).

Finally, we speculate that a large-scale classification and associated map to local-scale cloud properties could form the basis for future GCM cloud parameterizations, if a sufficient "library" of cloud properties were available. We envision this process would work by having a custom large-scale classification associated with every GCM grid cell. When a given synoptic-scale pattern was encountered, the library or distribution of cloud properties profiles (and heating rates) for that location would be consulted and a possible cloud realization selected. In the near future, we hope to apply this classification technique to many years of model output from a novel GCM that contains an embedded high-resolution cloud-resolving model (e.g., Khairoutdinov and Randall 2001; Randall et al. 2003). This GCM may prove capable of providing both the large-scale atmospheric fields and library of cloud properties needed to create a statistical parameterization.

REFERENCES

Bardossy, A., L. Duckstein, and I. Bogardi, 1995: Fuzzy rule-based classifications of atmospheric circulation patterns. *Int. J. Climatol.,* **15,** 1087–1097.

Benjamin, S. G., K. Brewster, R. Brummer, B. F. Jewett, T. W. Schlatter, T. L. Smith, and P. A. Stamus, 1991: An isentropic three-hourly data assimilation system using ACARS aircraft observations. *Mon. Wea. Rev.,* **119,** 888–906.

——, J. M. Brown, K. J. Brundage, D. Devenyi, B. E. Schwartz, T. G. Smirnova, T. L. Smith, and F.-J. Wang, 1996: Recent upgrades to MAPS/RUC. FSL Forum, 33 pp. [Available from NOAA/Forecast System Laboratory, 325 Broadway, Boulder, CO 80303-3328.]

——, G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004a: Mesoscale weather prediction with the RUC hybrid isentropic–terrain-following coordinate model. *Mon. Wea. Rev.,* **132,** 473–494.

——, and Coauthors, 2004b: An hourly assimilation–forecast cycle: The RUC. *Mon. Wea. Rev.,* **132,** 495–518.

Cess, R. D., and Coauthors, 1990: Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *J. Geophys. Res.,* **95,** 16 601–16 615.

Colle, B. A., and C. F. Mass, 1995: The structure and evolution of cold surges east of the Rocky Mountains. *Mon. Wea. Rev.,* **123,** 2577–2610.

Efron, B., and R. J. Tibshirani, 1993: *An Introduction to the Bootstrap.* Chapman and Hall, 436 pp.

Haykin, S., 1998: *Neural Networks: A Comprehensive Foundation*. 2d ed. Prentice Hall, 700 pp.

Hewitson, B. C., 1994: Regional climates in the GISS general circulation model: Surface air temperature. *J. Climate,* **7,** 283–303.

——, and R. G. Crane, 1992: Regional climates in the GISS global circulation model: Synoptic-scale circulation. *J. Climate,* **5,** 1002–1011.

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

Hobbs, P. V., J. D. Locatelli, and J. E. Martin, 1996: A new conceptual model for cyclones generated in the lee of the Rocky Mountains. *Bull. Amer. Meteor. Soc.,* **77,** 1169–1178.

Jakob, C., and G. Tselioudis, 2003: Objective identification of cloud regimes in the tropical western Pacific. *Geophys. Res. Lett.,* **30,** 2082, doi:10.1029/2003GL018367.

——, ——, and T. Hume, 2005: The radiative, cloud, and thermodynamic properties of the major tropical western Pacific cloud regimes. *J. Climate,* **18,** 1203–1215.

Kalkstein, L. S., G. Tan, and J. A. Skindlov, 1987: An evaluation of three clustering procedures for use in synoptic climatological classification. *J. Climate Appl. Meteor.,* **26,** 717–730.

——, P. C. Dunne, and R. S. Vose, 1990: Detection of climate change in the western North American Arctic using a synoptic climatological approach. *J. Climate,* **3,** 1153–1167.

——, C. D. Barthel, J. S. Greene, and M. C. Nichols, 1996: A new

spatial synoptic classification: Application to air mass analysis. *Int. J. Climatol.,* **16,** 983–1004.

——, S. C. Sheridan, and D. Y. Graybeal, 1998: A determination of character and frequency changes in air masses using a spatial synoptic classification. *Int. J. Climatol.,* **18,** 1223–1236.

Khairoutdinov, M. F., and D. A. Randall, 2001: A cloud resolving model as a cloud parameterization in the NCAR Community Climate Model: Preliminary results. *Geophys. Res. Lett.,* **28,** 3617–3620.

Kohonen, T., 1995: *Self-Organizing Maps.* Springer, 362 pp.

Mecikalski, J. R., and J. S. Tilley, 1992: Cold surges along the front range of the Rocky Mountains: Development of a classification scheme. *Meteor. Atmos. Phys.,* **48,** 249–271.

Michelangeli, P.-A., R. Vautard, and B. Legras, 1995: Weather regimes: Recurrence and quasi stationarity. *J. Atmos. Sci.,* **52,** 1237–1256.

Neiman, P. J., and R. M. Wakimoto, 1999: The interaction of a Pacific cold front with shallow air masses east of the Rocky Mountains. *Mon. Wea. Rev.,* **127,** 2102–2127.

Norris, J., and C. Weaver, 2001: Improved techniques for evaluating GCM cloudiness applied to the NCAR CCM3. *J. Climate,* **14,** 2540–2550.

Ozelkan, E. C., A. Galambosi, L. Duckstein, and A. Bardossy, 1998: A multi-objective fuzzy classification of large scale atmospheric circulation patterns for precipitation modeling. *Appl. Math. Comput.,* **91,** 127–142.

Potter, G. L., and R. D. Cess, 2004: Testing the impact of clouds on the radiation budgets of 19 atmospheric general circulation models. *J. Geophys. Res.,* **109,** D02106, doi:10.1029/2003JD004018.

Randall, D. A., M. F. Khairoutdinov, A. Arakawa, and W. Grabowski, 2003: Breaking the cloud parameterization deadlock. *Bull. Amer. Meteor. Soc.,* **84,** 1547–1564.

Romero, R., G. Sumner, C. Ramis, and A. Genoves, 1998: A classification of the atmospheric circulation patterns producing significant daily rainfall in the Spanish Mediterranean area. *J. Climatol.,* **19,** 765–785.

Schultz, D. M., 2004: Cold fronts with and without prefrontal wind shifts in the central United States. *Mon. Wea. Rev.,* **132,** 2040–2053.

——, W. E. Bracken, and L. F. Bosart, 1998: Planetary- and synoptic-scale signals associated with Central American cold surges. *Mon. Wea. Rev.,* **126,** 5–27.

Tennant, W., 2003: An assessment of intraseasonal variability from 13-yr GCM simulations. *Mon. Wea. Rev.,* **131,** 975–990.

Wilks, D. S., 1997: Resampling hypothesis tests for autocorrelated fields. *J. Climate,* **10,** 65–82.

Ye, H., L. S. Kalkstein, and J. S. Green, 1995: The detection of climate change in the Arctic: An updated report. *Atmos. Res.,* **37,** 153–173.

Zivkovic, M., and J.-F. Louis, 1992: A new method for developing cloud specification schemes in general circulation models. *Mon. Wea. Rev.,* **120,** 2928–2941.