Spectrum of Near-Storm Environments for Significant Severe Right-Moving Supercells in the Continental United States

ROBERT A. WARREN,a,b HARALD RICHTER, c AND RICHARD L. THOMPSONd

a School of Earth, Atmosphere and Environment, Monash University, Clayton, Victoria, Australia
b Australian Research Council Centre of Excellence for Climate Extremes, Monash University, Clayton, Victoria, Australia
c Bureau of Meteorology, Melbourne, Victoria, Australia
d Storm Prediction Center, Norman, Oklahoma

(Manuscript received 12 January 2021, in final form 5 July 2021)

ABSTRACT: Proximity soundings have long been used to explore how the vertical structure of temperature, humidity, and winds influence convective storms and their associated hazards. In severe thunderstorm research and forecasting, convective parameters are often used to summarize certain characteristics of the sounding. While extremely useful, these parameters are unable to describe the richness that is readily apparent in hodographs and skew $T$–log $p$ diagrams. Motivated by a desire to retain more of these details, the present study uses self-organizing maps (SOMs) to group soundings based on their full vertical structure. The analysis makes use of a sample of more than 10,000 model proximity soundings for right-moving supercells associated with tornadoes and significant severe hail and straight-line winds in the contiguous United States (CONUS). Separate SOMs are developed for the wind and thermodynamic profiles, each with 3 × 3 nodes, resulting in a set of nine hodographs and nine skew $T$–log $p$ diagrams that broadly represent the spectrum of near-storm environments for significant severe right-moving supercells in the CONUS. Both SOMs are shown to provide a good representation of the variability in key convective parameters, although, for the thermodynamic SOM, variations in LCL heights and midlevel lapse rates are somewhat limited. Based on the soundings assigned to them, the SOM nodes are characterized in terms of their associated hazards, their relationship with storm mode and mesocyclone strength, and their spatial and temporal variability. Potential applications of the SOMs in severe weather forecasting and idealized numerical simulations are also highlighted.

KEYWORDS: Supercells; Storm environments; Soundings; Clustering; Severe storms

1. Introduction

It has long been recognized that the characteristics of convective storms are strongly governed by the vertical structure of the atmosphere in their immediate environment. Much of our understanding of the relationship between the near-storm environment (NSE) and the occurrence of severe convective weather (tornadoes, large hail, and damaging straight-line winds) stems from the study of proximity soundings, which dates back to the mid-twentieth century (Showalter and Fults 1943; Fawbush and Miller 1952, 1954; Beebe 1955, 1958). This work has also led to the development of many of the parameters and indices that are routinely used in severe thunderstorm forecasting, not just in the United States (where the vast majority were developed), but all around the world. Examples range from simple single-level or layer-based indices (e.g., lifted index, 700–500-hPa lapse rates), through vertical integrals, such as convective available potential energy (CAPE), convective inhibition (CIN), and storm-relative helicity (SRH), to so-called composite parameters, such as the supercell composite parameter (SCP) and significant tornado parameter (STP). The value of these diagnostics has been repeatedly demonstrated, both in peer-reviewed research (e.g., Rasmussen and Blanchard 1998; Craven and Brooks 2004; Taszarek et al. 2017) and through many years of operational use. However, it is important to recognize that these convective parameters represent a form of data reduction, with an associated loss of potentially important information. Doswell and Schultz (2006) note the difficulty of defining a diagnostic variable “that distills the relatively rich complexity of a complete sounding into a single number” and emphasize the value of “simply looking at the soundings.” The present study is in part motivated by this philosophy.

The concept of a representative proximity sounding also underlies many of the numerical modeling studies that have informed our modern understanding of convective storm dynamics. Idealized simulations typically employ time-invariant and horizontally homogeneous base-state conditions, based on either observed soundings or analytical functions designed to represent typical NSEs for particular storm types. Analytical soundings are particularly useful since they permit a controlled investigation of storm sensitivities to specific environmental variables by varying a small number of parameters. Prominent examples include the Weisman and Klemp (1982, hereinafter WK82) thermodynamic sounding and the quarter-circle hodograph from Weisman and Rotunno (2000, hereinafter WR00), both of which continue to be widely used in idealized
supercell simulations\(^1\) (e.g., Dennis and Kumjian 2017; Brown and Nowotarski 2019; Boyer and Dahl 2020). While these soundings capture the basic features of right-moving supercell NSEs, they also have some notable deficiencies. In particular, the WK82 thermodynamic sounding is considerably moister than the majority of observed supercell environments (e.g., Warren et al. 2017, their Fig. 3), while the WR00 hodograph imposes low-level veering over a deeper layer than typically observed (2 km vs. ~1 km in observed supercell NSEs; see, for example, Markowski et al. 2003, their Fig. 12) and neglects wind shear above 6 km. These issues naturally bring into question the applicability of many idealized modeling results to real supercells.

One simple way to create a more realistic representation of the NSE is to average over a large number of observed soundings. Such composite profiles have long been used to characterize the environments in which specific storm types or severe phenomena occur (e.g., Fawbush and Miller 1954; Maddox 1976; Bluestein and Jain 1985; Schaefer and Livingston 1988; Rasmussen and Straka 1998; Doswell and Evans 2003; Markowski et al. 2003; Ramsay and Doswell 2005; Bunkers et al. 2006; Parker 2014; Wade et al. 2018; Coniglio and Parker 2020). Examples include the Type I, II, and III thermodynamic composites for tornadoes presented by Fawbush and Miller (1954), the short-, moderate-, and long-lived supercell composite hodographs presented by Bunkers et al. (2006), and the tornadic and nontornadic composite supercell soundings from the second Verification of the Origins of Rotation in Tornados Experiment (VORTEX2; Wurman et al. 2012) presented by Parker (2014). The latter have been applied in a series of idealized modeling studies investigating the process of tornadogenesis and its sensitivity to the NSE (Coffer and Parker 2017; Coffer et al. 2017; Coffer and Parker 2018; Flourney et al. 2020).

The present study seeks to create composite supercell soundings, similar to those mentioned above, but based on a much larger sample of cases. Rather than grouping soundings according to their associated hazards or predefined storm types, we use the properties of the soundings themselves; that is, the vertical structure of the temperature, humidity and wind profiles. This is similar to the subjective feature-based composing method proposed by Brown (1993), but here grouping is performed objectively using self-organizing maps (SOMs; Kohonen 2013). The SOM algorithm is a type of artificial neural network, which maps input data to a set of representative clusters or “nodes” through an iterative learning process. The nodes are arranged in a (usually two-dimensional, rectangular) lattice, with similar nodes being adjacent to each other and dissimilar nodes being widely separated. This topological ordering is the defining characteristic of SOMs and it provides a key advantage over other clustering methods (such as K-means): namely that the input data are treated as a continuum rather than discrete realizations of a set of distinct patterns (e.g., Hewitson and Crane 2002; Reusch et al. 2007).

Although originally developed in the 1980s (Kohonen 1982), SOMs have only been widely used in meteorological and climate sciences for the last 20 years (Liu and Weisberg 2011). Common applications include synoptic climatology (e.g., Hewitson and Crane 2002; Alexander et al. 2010; Jiang et al. 2015), satellite cloud classification (e.g., Ambroise et al. 2000; McDonald et al. 2016), and the characterization of weather and climate extremes (e.g., Cavazos 2000; Cassano et al. 2015; Zhuang et al. 2020). In the context of severe thunderstorm research, applications of SOMs have so far been very limited. Most relevant to the present work are the studies of Nowotarski and Jensen (2013, hereinafter NJ13) and Nowotarski and Jones (2018, hereinafter NJ18). Both applied SOMs to proximity soundings, with the aim to develop classifications that could distinguish between supercell and nonsupercell NSEs (NJ13) and/or between nontornadic, weakly tornadic, and significantly tornadic supercell NSEs (NJ13; NJ18). A key finding was that SOMs based on wind-related variables, or a combination of wind and thermodynamic variables, provided better discrimination of storm types than those based only on thermodynamic variables. SOMs have also been used to characterize tornadic storm environments by Anderson-Frey et al. (2017, 2018) and Anderson-Frey and Brooks (2019); however, in this case the input fields were two-dimensional grids of STP rather than individual proximity soundings.

Two SOMs are developed in the present study: one based on wind profiles (visualized as hodographs) and one based on thermodynamic profiles (visualized as skew T-logp diagrams). These are trained on a sample of 11 483 model proximity soundings associated with right-moving supercells in the contiguous United States (CONUS). Each sounding is associated with a severe weather report, with corresponding information on the location, timing, and magnitude (tornado intensity, hail size, or wind speed) of the event, as well as radar-based storm attributes (convective mode and mesocyclone strength). This allows us to explore variations in storm and hazard characteristics across the SOM nodes, together with the spatial and temporal distributions of the associated environments. In contrast to NJ13 and NJ18, our objective is not to optimally distinguish between different storm/hazard types, but rather to create a set of temperature, humidity, and wind profiles that broadly encompass the spectrum of right-moving supercell NSEs. We suggest that these profiles could be used in future idealized modeling studies, as an alternative to observed or analytical soundings. They may also have applications in severe weather forecasting.

The remainder of this paper is organized as follows. Section 2 introduces the sounding database and outlines basic preprocessing that was used to prepare the data for input to the

\(^1\) These profiles are implemented together as the default supercell environment in both the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2019) and Cloud Model 1 (CM1; Bryan and Fritsch 2002; Bryan 2017), which no doubt contributes to their continued widespread use. For the WK82 thermodynamic sounding, another factor may be the difficulty in simulating sustained convection in environments lacking deep moisture (e.g., McCaul and Cohen 2004; Naylor and Gilmore 2012; Nowotarski et al. 2020).
SOM algorithm. Section 3 then details the construction of the wind and thermodynamic SOMs, including the choice of input variables and algorithm settings. Following Huva et al. (2015), a filtering procedure was used to reject soundings that were poorly represented by the SOM nodes. This is described in section 4, which also highlights the error characteristics of the SOMs. Analysis of the wind and thermodynamic SOMs is performed in sections 5 and 6, respectively, and section 7 examines the correspondence between the two. In section 8, conclusions are drawn and potential applications of the SOMs are highlighted.

2. Data

This study makes use of a large proximity sounding database developed over more than a decade by forecasters at the Storm Prediction Center (SPC). The original version of this database, presented by Smith et al. (2012) and Thompson et al. (2012), covered the period 2003–11 but was later extended to include 2012 (Smith et al. 2014). It comprised storm classifications, model proximity soundings, and derived convective parameters associated with reports of tornadoes, significant severe hail [sighail; ≥2 in. (1 in. = 2.54 cm)], and significant severe straight-line winds (sigwind; ≥65 kt; 1 kt = 0.51 m s\(^{-1}\)) in the CONUS. For each hazard type, reports were filtered for the largest magnitude report per hour on the 40-km grid used in the SPC mesoanalysis system (Bothwell et al. 2002). Based on archived radar observations, storms associated with each report were manually classified into three convective mode classes: supercell, quasi-linear convective system (QLCS), and disorganized. QLCSs were then further divided into lines and bow echoes, while supercell and disorganized storms were subclassified as discrete cells, cells in a cluster, or cells in a line (see Smith et al. 2012 for details). Supercells were also separated into right-moving (RM) and left-moving storms, and their mesocyclones were classified as weak, moderate, or strong using range-dependent azimuthal shear nomograms (Andra 1997; Stumpf et al. 1998). For each report, proximity soundings and associated environmental parameters were extracted from the SPC mesoanalysis, which used the Rapid Update Cycle (RUC; Benjamin et al. 2004) prior to May 2012 and the Rapid Refresh (RAP; Benjamin et al. 2016) thereafter. In this system, pressure-level data from the hourly RUC/RAP analysis are combined with a surface objective analysis (SFCOA), with the latter produced by applying a two-pass Barnes analysis (Barnes 1973) to available surface observations.

Since 2012, the SPC database has been updated several times as part of a series of projects exploring the relationship between low-level rotational velocities, the NSE, and tornado intensity (Smith et al. 2015; Thompson et al. 2017; Smith et al. 2020a,b). The current version spans the years 2003–19. Because the focus on tornadoes, hail and wind events were not included in 2013 or for the most recent four years (2016–19). However, motivated by the need for a null event sample, all tornado, severe hail (≥1 in.), and severe wind (≥50 kt) events associated with RM supercells and QLCSs were analyzed for the years 2014 and 2015. Since the convective mode was not known a priori, reports for 2014 were only considered if the effective bulk wind difference (EBWD; Thompson et al. 2007) at the nearest mesoanalysis grid point exceeded 20 kt. Following analysis of the 2014 data, this criterion was changed to a minimum EBWD or 0–6-km bulk wind difference (BWD06) of 40 kt for data collection in 2015. This maintained the majority of cyclonically rotating storms (63% when applied retroactively to the 2014 data) while significantly reducing the number of cases to be examined. For additional details see Thompson et al. (2017).

For the present study, all tornado, sighail, and sigwind reports associated with RM supercells during the years 2005–12 and 2014–15 were extracted from the SPC database. Data prior to 2005 were excluded due to the coarse vertical resolution of the model profiles (50 hPa vs 25 hPa for later years). To avoid biasing our results toward tornadic supercells, we also excluded all years where only tornado reports were analyzed. While the minimum BWD thresholds used to filter reports in 2014 and 2015 likely lead to the exclusion of more marginal supercell environments (application of the 2014 and 2015 thresholds retroactively to 2005–12 eliminated 3% and 23% of soundings, respectively), we opted to retain these years to achieve a larger sample size. However, nonsignificant severe hail and wind reports for 2014 and 2015 were excluded from our analysis since these were not available in other years.

During the 10-yr study period there were a total of 13 896 reports, each with a corresponding RUC/RAP sounding. Filtering was applied to remove soundings associated with tropical cyclones (409) and marginal supercells (704), those characterized by most unstable CAPE < 100 J kg\(^{-1}\) and/or most unstable CIN > 250 J kg\(^{-1}\) (134),\(^2\) and those with missing data (22). Of the remaining 12 627 soundings, 11 444 were found to be duplicates, either associated with multiple reports of the same hazard (330 soundings) or reports of multiple hazards (814 soundings) occurring in the same RUC/RAP grid box during the same hour. These duplicated soundings were removed. In the case of multiple reports of a single hazard, only the highest severity report was retained. The multihazard reports consist of 369 tornado–sighail events, 266 tornado–sigwind events, 117 sighail–sigwind events, and 31 tornado–sighail–sigwind events. In each case, all reports were retained but assigned to a single sounding. The final dataset contains 11 483 unique soundings associated with a total of 12 297 reports. The latter consist of 6590 tornado reports, of which 1256 (19%) are significant tornado (sigtor; ≥EF2 on the enhanced Fujita scale) reports, 3815 sighail reports, and 1892 sigwind reports. The spatial distribution of these reports is shown in Fig. 1.

The raw proximity soundings consist of profiles of pressure \(p\), temperature \(T\), dewpoint temperature \(T_d\), horizontal wind speed \(V\) and direction \(\phi\), and height \(z\) above mean sea level (MSL) at regular pressure levels up to 100 hPa, together with corresponding values at the surface. For processing, winds were first converted to zonal \(u\) and meridional \(v\) components.

\(^2\)These CAPE and CIN thresholds are equivalent to those used to define the effective inflow layer (Thompson et al. 2007).
Profiles were then linearly interpolated to a regular height grid with 100 m spacing, extending from the surface to 15 km above ground level (AGL). In elevated regions (surface height above ~1500 m MSL) the 100-hPa level is located below 15 km AGL and data had to be extrapolated. To do this we simply assumed that $T$, $T_d$, $u$ and $v$ remained fixed at their 100-hPa values and then computed $p$ using hydrostatic balance. A total of 258 soundings (2% of the dataset) were extrapolated, over an average distance of 268 m. The largest extrapolation distance was 1208 m for a sounding in Colorado.

From the interpolated data, convective parameters were recomputed to maintain consistency with the profiles. These include CAPE and CIN, together with the lifted parcel level (LPL), lifted condensation level (LCL), level of free convection (LFC), and level of neutral buoyancy (LNB), for surface-based (SB), most-unstable (MU), and 500 m mixed-layer (ML) parcel ascents$^3$ (calculated using the virtual temperature correction; Doswell and Rasmussen 1994), BWD06 (computed as the magnitude of the vector difference between the 0–500-m and 5.5–6-km mean winds), and SRH (computed using storm motion estimated following Bunkers et al. 2000) for the 0–1- and 0–3-km layers (SRH01 and SRH03, respectively).

Before discussing the construction of our SOMs it is worth noting some limitations of the SPC database. First, the reports used to identify events are subject to a number of known issues including: 1) spatial and temporal biases (e.g., Schaefer and Galway 1982; Trapp et al. 2006; Allen and Tippett 2015); 2) errors in the estimation of tornado intensity, hail size, and wind speed (e.g., Doswell and Burgess 1988; Schaefer et al. 2004; Edwards et al. 2018); and 3) secular (nonmeteorological) trends associated with changes in reporting and forecast verification practices (e.g., Weiss et al. 2002; McCarthy and Schaefer 2004; Doswell et al. 2005). Another reporting issue, not discussed in previous studies, is the relative dearth of multihazard events. While one might intuitively expect many tornadic supercells to also produce significant severe hail and, to a lesser extent, straight-line winds, this is not reflected in the SPC database: only 10% of the tornado reports considered in the present study were accompanied by reports of hail and/or wind. It is very possible that large hail and straight-line winds are less likely to be reported when a tornado is present. In addition, straight-line wind damage may be incorrectly attributed to the tornado. Whatever the cause, we must be cognizant that the three hazards are not mutually exclusive. This is particularly relevant when considering the relative proportion of tornado, hail, and wind reports assigned to each SOM node (sections 5b and 6b).

As with any study involving proximity soundings, it is also important to consider issues of data accuracy and representativity (e.g., Brooks et al. 1994; Potvin et al. 2010). Mesoscale models such as the RUC and RAP offer obvious advantages over rawinsonde observations in terms of spatial and temporal sampling; however, they can also introduce errors. For the RUC analysis, Thompson et al. (2003) found cool and dry biases at the surface (leading to a negative bias in SBCAPE) and a positive wind speed bias in the lower troposphere when compared with observed soundings (also noted by Coniglio 2012). For the RAP analysis, Wade et al. (2018) identified warm and dry biases relative to soundings from the Mesoscale Predictability Experiment (MPEX; Trapp et al. 2016). Use of the SFCOA is expected to mitigate biases at the surface, but does not redress errors farther aloft. Thompson et al. (2012) cite the 2007 Greensburg, Kansas, tornado as an example of an event in which errors above the surface (in this case, a dry bias) limited the representativity of

Figure 1. Map showing the distribution of filtered tornado, sighail, and sigwind reports used in this study.

---

$^3$Parcel ascents were performed using a modified version of the CAPE code from CM1 (https://www2.mmm.ucar.edu/people/bryan/Code/getcape.F).
the mesoanalysis proximity sounding. In addition, the Barnes analysis procedure and relatively coarse horizontal resolution will tend to smooth out features such as surface boundaries that may locally enhance the storm environment. A key assumption of the present work is that the mesoanalysis provides an accurate representation of the NSE in the majority of cases and that the very large sample size will limit the effect of outliers.

3. Constructing the SOMs

Two SOMs were developed in this study: one for wind profiles between the surface and 10 km AGL and one for thermodynamic profiles between the surface and 15 km AGL. Observations (e.g., Markowski et al. 2003; Rasmussen 2003; Thompson et al. 2007; Bunkers et al. 2014) and modeling studies (e.g., Weisman and Klemp 1982, 1984; McCaul and Weisman 2001; Adlerman and Droegemeier 2005) suggest that supercell dynamics are primarily governed by winds in the lower troposphere (below 6 km), although winds at higher altitudes have been found to influence storm motion and morphology (Rasmussen and Straka 1998; Ramsay and Doswell 2005; Warren et al. 2017). The maximum altitude for the wind SOM (W-SOM) was set at 10 km to capture some variation in upper-level winds without having these dominate the SOM nodes. For the thermodynamic SOM (T-SOM), a higher maximum altitude of 15 km was used in order to represent soundings with deep layers of positive buoyancy (i.e., a high LNB).

For the W-SOM, profiles of $u$ and $v$ were transformed by first subtracting the 0–500-m mean wind and then rotating each hodograph so that the 0–6-km shear vector (computed using the 0–500-m and 5.5–6-km mean winds) aligns with the positive $x$ direction. Expressed mathematically:

$$u'(z) = V^y(z) \cos[\phi^y(z) - \phi^y_0], \quad \text{and}$$
$$v'(z) = V^y(z) \sin[\phi^y(z) - \phi^y_0],$$

where

$$V^y(z) = \sqrt{[u(z) - u_0]^2 + [v(z) - v_0]^2},$$

$$\phi^y(z) = \arctan \left( \frac{v(z) - v_0}{u(z) - u_0} \right), \quad \text{and}$$

$$\phi^y_0 = \arctan \left( \frac{v_0 - v_0}{u_0 - u_0} \right).$$

Here, $(u', v')$ are the transformed wind components, $V^y$ and $\phi^y$ are, respectively, the magnitude and direction of the shifted wind, $\phi^y_0$ is the direction of the 0–6-km shear, and $(u_0, v_0)$ and $(u_0, v_0)$ are the wind components averaged over the 0–500-m and 5.5–6-km layers, respectively. Assuming supercell motion is approximately Galilean invariant (e.g., Bunkers et al. 2000), this transformation means that soundings are effectively grouped based on the storm-relative wind profile, which is a primary control on storm dynamics (e.g., Rotunno and Klemp 1982, 1985; Weisman and Klemp 1982, 1984). Ground-relative winds were used in the SOMs of NJ18 as they appear to be more relevant to the formation and intensity of tornadoes (Markowski et al. 2003; NJ13).

For the T-SOM, raw profiles of $T$ and $T_d$ were initially used as inputs, together with surface pressure $p_{sfc}$ (which is needed to reconstruct the full pressure profile). However, the resulting map was dominated by variations in column-mean temperature and did a poor job of capturing variations in the vertical thermodynamic structure (not shown). After further testing, profiles of buoyancy $B$ and relative humidity (RH), together with $p_{sfc}$ and surface temperature $T_{sfc}$, were selected as inputs to the T-SOM. Here, buoyancy is defined as

$$B = g \left( \frac{T_{v,p} - T_{v,e}}{T_{v,e}} \right),$$

where $T_{v,p}$ is the virtual temperature of a parcel lifted dry adiabatically from the surface to the LCL and pseudoadiabatically thereafter, $T_{v,e}$ is the environmental virtual temperature, and $g$ is the acceleration due to gravity. Note that $B$ at the surface is always zero by definition and is therefore not included as an input to the SOM. However, $T_{sfc}$ is needed to reconstruct the profiles of $T$ and $T_d$ so that the SOM nodes can be represented on a skew $T$–$\log p$ diagram.

For input to the SOM algorithm, profiles were collected into a single array with dimensions of $N_x \times N_y$, where $N_x$ is the number of input samples (soundings) and $N_y$ is the number of features in each sample. Here a feature corresponds to a single variable on a single height level; thus $N_y = 202$ for the W-SOM and $N_y = 303$ for the T-SOM. Each feature is then normalized by subtracting the mean and dividing by the standard deviation (computed across all input samples). In theory, this procedure ensures that all features carry equal weight in the training of the SOM; however, it relies on the assumption that features are normally distributed. This turns out to be a reasonable approximation for $u', v', B$, and $T_d$, but is less valid for RH and $p_s$. In particular, the distribution for RH varies significantly with altitude, with negative skewness in the lower troposphere, positive skewness in the stratosphere, and a bimodal structure in the mid- to upper troposphere (not shown). To try to address this issue, other normalization procedures (e.g., histogram normalization) were tested; however, these were found to degrade SOM quality (see below). As such, the standard normalization was retained for all variables.

Training of the SOM was performed using the SOM Toolbox for MATLAB, developed by Helsinki University of Technology (Vesanto et al. 2000). Details of the training procedure and relevant parameter choices are provided in the appendix. Consistent with previous meteorological applications of SOMs, a rectangular arrangement of nodes was used. For the dimensions of the SOM, numerous configurations were tested, with sizes ranging from 2 to 30 nodes and aspect ratios between 1 and 2. Two standard measures of SOM quality were used to compare these configurations: quantization error (QE) and topological error (TE). QE is the average Euclidean distance between each feature vector and its closest SOM node, referred to as the best matching unit (BMU). TE is the proportion of feature vectors for which the first and second BMUs are not adjacent nodes. Lower QE is associated with more compact data groupings, while lower TE indicates better topological ordering. In addition to these metrics we also examined the
ability of each SOM to capture the variability in key environmental parameters (e.g., BWD06 for the W-SOM and MLCAPE for the T-SOM). A $3 \times 3$ SOM was ultimately selected as the best compromise between quality and interpretability.

The output of the SOM algorithm is an array of dimensions $N_c \times N_f$, where $N_c$ is the number of SOM nodes (clusters). Every input sample is assigned to its BMU based on the Euclidean distance between the normalized feature vectors and the SOM nodes. The features composing each node are then denormalized (by multiplying by the previously derived standard deviation and adding the mean) and separated into the component variables. The profiles of $u'$ and $v'$ for each of the W-SOM nodes can be directly plotted as hodographs. To plot the T-SOM nodes as skew $T$–log$p$ diagrams, profiles of $T$, $T_d$, and $p$ must be derived from $B$, $R$H, $T_{dc}$, and $p_{dc}$. This is achieved by performing a surface-based parcel ascent to obtain $T_{rd}$, solving Eq. (5) for $T_{sh}$, and combining this with RH in an iterative procedure to retrieve $T$ and $T_d$. The pressure profile is then obtained using the hypsometric equation.

4. SOM errors and filtering

Like any data-reduction technique, SOMs cannot capture the full range of variability represented in the input data. While every sounding is assigned to a SOM node, there can be significant differences between the two. This is illustrated in Fig. 2, which plots the Euclidean distance $e$ between every normalized feature vector and its BMU in increasing order for both the W- and T-SOMs. For ease of comparison between the two SOMs, $e$ is normalized by $e_{50}$, the median Euclidean distance across all soundings. The distribution for both SOMs is highly skewed. Most values fall close to the median but a few are considerably larger: up to five times $e_{50}$ for the T-SOM and seven times $e_{50}$ for the W-SOM. The longer tail for the W-SOM appears to be associated with large variability in upper-level winds (not shown).

Following Huva et al. (2015), we apply filtering to separate these poorly represented soundings into their own “rejected” category. Specifically, we reject those soundings with $e > e_{50}$, the 95th percentile of the distribution (dashed line in Fig. 2). Thus, for both SOMs, 574 soundings are rejected. It should be emphasized that the set of rejected soundings differs between the W- and T-SOMs (the degree of overlap between them is assessed in section 7). The SOMs were not retrained after filtering, as inclusion of the rejected soundings was not found to strongly influence the nodes. Nevertheless, we consider these soundings separately in our analysis in order to highlight the characteristics of more “atypical” supercell NSEs.

Table 1 lists the number and percentage of soundings assigned to each node of the W- and T-SOMs. Even after filtering, there is a very robust sample size for every node. Over 1000 soundings are assigned to each of the W-SOM nodes and all but one of the T-SOM nodes, with a fairly even spread across the nodes. Node 9 of the T-SOM is the exception, with only 608 soundings (5.3% of the total). As we will show in section 6, this is the only node to feature a pronounced surface stable layer and an elevated MU parcel, which highlights the rarity of severe supercells in environments lacking significant surface-based instability.

To assess the representativeness of the SOM nodes, Fig. 3 plots the distribution of the difference between each sounding and its assigned node as a function of height for four variables: $u'$ and $v'$ for the W-SOM and $T$ and $T_d$ for the T-SOM. For $u'$ and $v'$, the distribution is roughly symmetric and the median error is close to zero, indicating that the W-SOM is essentially unbiased. The reduction in spread for both variables just above the surface, and for $v'$ just below 6 km, is a consequence of the wind transformation, which constrains $u'$ and $v'$ to be near zero in the 0–500-m layer and $v'$ to be near zero in the 5.5–6-km layer. Below 6 km, the interquartile range (IQR) of the distribution is characterized by errors of $<2.5$ m s$^{-1}$. Above this, error magnitudes increase, particularly for $u'$, with the IQR reaching $\pm 5$ m s$^{-1}$ at 10 km. For $T$ and $T_d$, the IQR remains within $\pm 5^\circ$C at all levels and is mostly within $\pm 2.5^\circ$C for temperature. However, for dewpoint temperature there is a positive bias of around 1–2°C above 3 km and the error distribution is much wider and positively skewed in the middle troposphere. These features likely reflect the non-Gaussian distribution of RH and its nonlinear relationship with $T_d$. Overall, however, it appears that both SOMs are broadly representative of the soundings assigned to them.

5. Wind SOM

The following sections examine the W-SOM and the characteristics of the soundings and associated reports assigned to each of its nodes.

a. SOM nodes and parameters

Figure 4 shows hodographs corresponding to each of the W-SOM nodes. Very similar profiles are obtained if we average $u'$ and $v'$ across the soundings assigned to each node (thin colored lines). Consistent with composite supercell hodographs presented in previous studies (Markowski et al. 2003; Ramsay and Doswell 2005; Bunkers et al. 2006; Parker 2014; Coniglio and Parker 2020), all nine profiles feature veering winds in the lowest 1 km and predominantly unidirectional shear above this. However, there are clear variations in hodograph length and structure across the SOM. From top to bottom, we see a pronounced decrease in hodograph length,
while the left-to-right variation is characterized by both a decrease in hodograph length and a trend toward backing midlevel shear. The net result is a dramatic reduction in both BWD06 and SRH from the top-left to bottom-right corners of the SOM.

It is clear that strong veering (approximately 90° of directional shear) within the lowest 1 km AGL is a fundamental characteristic of the NSE for significant severe RM supercells in the CONUS. This is not too surprising given the dynamical link between hodograph turning and supercell propagation (Rotunno and Klemp 1985; Davies-Jones 2002). However, it is interesting to note that this veering is imposed over a substantially deeper layer (the lowest 2 km AGL) in all of the W-SOM nodes. Observations presented by Rasmussen and Straka (1998) suggest that anvil-level storm-relative flow (which is a function of wind shear over the full depth of the troposphere) may influence supercell morphology, with weaker and stronger flow favoring high-precipitation and low-precipitation storms, respectively. This hypothesis was not supported by the more recent idealized modeling study of Warren et al. (2017); however, the authors did find significant changes in storm characteristics (including updraft width, rainfall intensity, and cold pool temperature) with increasing upper-level shear. While further work is needed to reconcile these observational and modeling results, it is apparent that wind shear above 6 km is a feature of supercell NSEs and should not be neglected.

A final characteristic of the W-SOM worth noting is the presence of pronounced backing (i.e., counterclockwise curvature) in the middle troposphere for some of the nodes (most notably Nodes 2, 3, and 6). This feature is also present in several nodes of NJ13’s 0–6 km ground-relative wind SOM (their Fig. 6) as well as a number of other composite hodographs (e.g., Maddox 1976, his Fig. 9; Coniglio and Parker 2020, their Fig. 10). Parker (2017) noted a common perception among forecasters and storm chasers that such wind profiles are unfavorable for long-lived discrete supercells, but found limited support for this hypothesis from idealized simulations. On the other hand, composite hodographs presented by Bunkers et al. (2006, his Fig. 5) showed more pronounced midlevel backing for short-lived supercells (average lifetimes of $\leq 2 h$) relative to moderate- and long-lived supercells (average lifetimes of 2–4 h and $\geq 4 h$, respectively), although this was accompanied by weaker deep-layer shear and SRH. While information on storm longevity was not available for the present study, it is clear that midlevel backing is not, on its own, an impediment to the formation of severe RM supercells.

To further explore variations between and within the W-SOM nodes, Fig. 5 plots distributions of six kinematic parameters for soundings assigned to each node. For comparison, the distributions across all soundings and those rejected from the W-SOM are also shown. The parameters are BWD06, SRH03, SRH01, and mean storm-relative winds (SRW) in the 0–1-, 0–3-, and 9–10-km layers (SRW01, SRW03, and SRW910, respectively). SRW01 and SRW03 are shown, in addition to the corresponding SRH parameters, as low-level storm-relative flow has been found to strongly modulate supercell updraft width and velocities (Warren et al. 2017; Peters et al. 2019, 2020). SRW910 represents the anvil-level storm-relative flow, which, as noted above, has been hypothesized to influence supercell morphology (Rasmussen and Straka 1998).

The distributions of BWD06 and SRW01 are quite narrow and mostly overlap to only a limited degree, indicating that the nodes are both compact and well separated in terms of deep-layer shear and low-level storm-relative flow. There is greater spread and more overlap between nodes for the other...
parameters, but still with clear differences across the SOM. Node 1 is characterized by the highest values of BWD06 (32.7 m s\(^{-1}\)), SRH03 (492 m\(^2\) s\(^{-2}\)), and SRH01 (364 m\(^2\) s\(^{-2}\)), while node 9 shows the lowest values of these parameters (15.8 m s\(^{-1}\), 134 m\(^2\) s\(^{-2}\), and 81 m\(^2\) s\(^{-2}\), respectively). Variations in SRW tend to mirror those of SRH, consistent with the results of Peters et al. (2020). A notable exception is node 2, which features higher SRH03 but near-average SRW03 relative to the all-soundings distribution. This characteristic appears to be associated with strong midlevel backing in the hodograph (Fig. 4), which reduces SRW in the 1–3-km layer, while maintaining large SRH01. SRW910 shows relatively smaller variations across the SOM, with node values just spanning the IQR of the all-soundings distribution. Node 3 has the highest SRW910 (22.8 m s\(^{-1}\)) and node 8 has the lowest (12.1 m s\(^{-1}\)).

In terms of their parameter values, the nodes appear to be broadly representative of the soundings assigned to them. This can be seen by comparing the values for each node (red crosses) with the mean of the corresponding distribution (black dots). Nodes with stronger low-level SRH and SRW tend to show a slight negative bias in these parameters (particularly pronounced for node 1). A subtle negative bias is also seen in SRW910 for node 3. Overall, however, the W-SOM does an excellent job of capturing the variability in these kinematic parameters, particularly BWD06, for which the nodes span the 10th to 90th percentiles of the all-soundings distribution.

b. Hazards and storm characteristics

Figure 6 shows how tornado, sigtor, sighail, and sigwind reports are distributed among the W-SOM nodes. The colored bars in each panel show the percentage of reports associated with soundings assigned to each node. For comparison, the overall percentage of soundings assigned to each node is shown in the gray bars (cf. Table 1). We can assess whether the number of reports associated with each node is significantly different from what we would expect for an equivalent random sample using the hypergeometric distribution.\(^4\) The probability

\(^4\)The hypergeometric distribution is used, rather than the binomial distribution, because sampling is performed without replacement.
that the number of reports \( N_j \) of hazard \( i \) associated with node \( j \) exceeds some value \( x \) is given by

\[
Pr(N_j \geq x) = \sum_{k=x}^{N_j} \binom{N_j}{k} \left( \frac{N_S - N_j}{N_j - k} \right). \tag{6}
\]

where \( N_j \) is the total number of reports of hazard \( i \), \( N_j \) is the total number of soundings assigned to node \( j \), and

\[
\binom{a}{b}
\]

is a binomial coefficient. If \( Pr < 0.01 \) then we say that there is a disproportionately high number of reports of hazard \( i \) for node \( j \). Conversely, if \( Pr > 0.99 \) then we say that there is a disproportionately low number of reports of hazard \( i \) for node \( j \). These two cases are indicated in Fig. 6 by upward and downward triangles, respectively.

Some care must be taken when interpreting Fig. 6. Given that the vast majority (93\%) of soundings are associated with a single hazard, nodes with a high proportion of reports of one hazard must automatically have a low proportion of reports of at least one of the other two hazards. This effect is most pronounced for tornado and sigsail reports, since these make up nearly 85\% of the total. Thus, nodes that are particularly favorable for tornadoes tend to appear less favorable for hail and vice versa. As discussed in section 2, it is quite possible that, in reality, these hazards occur together more often, as severe hail may go unreported when a tornado is present.

We see that nodes 1 and 2 are associated with a disproportionately high number of tornado and, in particular, sigsail reports, while the opposite is true for nodes 6, 8, and 9 (Figs. 6a,b). Nodes 1 and 2 collectively represent less than 20\% of all soundings but over 40\% of sigsail reports. Conversely, while over a third of soundings are assigned to nodes 6, 8, and 9, these are associated with just 16\% of sigsail reports. These results are consistent with the differences in BWD and SRH seen in Fig. 5. For example, nodes 1 and 2 are characterized by SRH01 values in excess of 250 m^2 s^{-2} [above the 75th percentile for sigsail soundings in Thompson et al. (2003, their Fig. 11)], while, for nodes 6, 8, and 9, SRH01 is only around 100 m^2 s^{-2} [close to the median for non-tornadic soundings in Thompson et al. (2003, their Fig. 11)]. Variations in the proportion of sigsail and sigwind reports across the SOM are less pronounced (Figs. 6c,d). Nodes 4 and 6 are associated with a high proportion of sigsail reports, while sigwind reports are disproportionately associated with node 6. Conversely, node 2 is associated with a low proportion of sigsail and sigwind reports.
reports; however, this may be an artifact of the high proportion of tornado reports associated with this node.

Another way of examining the distribution of hazards across the SOM is presented Fig. 7a. Here storms are classified as nontornadic (NT) if they are associated with only hail or wind reports, weakly tornadic (WT) if they are associated with an EF0 or EF1 tornado report, and significantly tornadic (ST) if they are associated with an EF2+ tornado report. The relative frequency of each category is then plotted for all soundings, those rejected from the W-SOM, and those assigned to each of the W-SOM nodes. Consistent with Fig. 6, we find the highest frequency of ST storms and the lowest frequency of NT storms for nodes 1 and 2. Conversely, the lowest frequency of ST storms occurs for nodes 6, 8, and 9. Interestingly, while the percentage of all tornadic (WT and ST) storms is highest for node 2 (72% vs 57% for all soundings), node 1 has the highest percentage of ST storms (25% vs 11% for all soundings). In other words, tornadoes associated with node 1 are more likely to be significant than those associated with node 2.

Figure 7 also reveals how mesocyclone strength and storm mode vary across the W-SOM nodes. We see large variations in mesocyclone strength across the SOM nodes (Fig. 7b), consistent with the variations in BWD06 and SRH seen in Fig. 5. Almost all of this variation occurs between the weak and strong categories, with the proportion of moderate mesocyclones varying by less than 10 percentage points. Around two-thirds (66%) of node-1 mesocyclones are classified as strong (cf. 37% for all soundings), whereas, for node 9, less than 16% are classified as strong and nearly 60% are classified as weak (cf. 35% of all soundings). In contrast, variations in storm mode across the SOM are much less pronounced (Fig. 7c). Nodes 3, 8, and 2 have the highest proportion of discrete, cell-in-cluster, and cell-in-line modes, respectively; however, in each case the difference with respect to all soundings is less than 10 percentage points. This finding does not contradict the well-established link between wind shear and convective organization, since all of the storms considered were supercells and the vast majority occurred in environments with at least 15 m s\(^{-1}\) of BWD06 (Fig. 5a). However, it does suggest that wind shear has limited impact on whether supercells occur as discrete storms or as part of larger convective systems.

### c. Spatial and temporal distributions

In this section we consider the spatial and temporal (seasonal and diurnal) distributions of the supercell soundings assigned to each W-SOM node. Our analysis procedure broadly follows that of Brooks et al. (2003), Doswell et al. (2005), and Krocak and Brooks (2018), who applied spatial and temporal smoothing to report data to examine tornado and severe thunderstorm probabilities in the CONUS. In particular, we use the same smoothing parameters and spatial grid resolution as these studies. For the temporal analysis, a two-dimensional array was created, containing the number of soundings for every day of the year and hour of the day, summed over all years. Hours in UTC were converted to local solar time (LST) by adding \(\lambda/15\) (rounded to the nearest integer), where \(\lambda\) is the sounding longitude. For nonleap years, artificial counts were created for 29 February by averaging the counts for 28 February and 1 March. The resulting \(366 \times 24\) array was then wrapped along both axes and smoothed using a two-dimensional Gaussian filter, with standard deviations of 15 days and 2 h for the seasonal (day of the year) and diurnal (hour of the day) axes, respectively. For the spatial analysis, sounding counts were evaluated on an 80-km grid and smoothed using a Gaussian filter with a standard deviation of 120 km. Note that this procedure inevitably leads to underestimation of values close to the coast and the borders with Canada and Mexico. Separate temporal and spatial counts were created for each of the SOM nodes, as well as for the rejected
Figure 7. Relative frequency of (a) nontornadic (NT), weakly tornadic (WT), and significantly tornadic (ST) supercells; (b) weak (W), moderate (M), and strong (S) mesocyclones; and (c) discrete (D), cell-in-cluster (C), and cell-in-line (L) supercell modes. Results are shown for all soundings (A), soundings rejected from the W-SOM (R), and soundings assigned to each of the W-SOM nodes (1–9).

Figure 8 presents the seasonal and diurnal distribution of all soundings, rejected soundings, and soundings assigned to each W-SOM node, both as joint-frequency heat maps and univariate normalized histograms. The all-soundings distribution shows high frequencies extending from April to June (with the maximum in late May) and from 1500 to 1800 LST (with the maximum at 1600 LST), with low frequencies between October and February and from 0100 to 0900 LST. A broadly similar pattern is seen for each of the SOM nodes, but with some noteworthy variations. Nodes 1–3 all show an earlier seasonal peak in mid- to late April, consistent with their large deep-layer shear (Fig. 4), which reflects strong baroclinicity at the start of the warm season. Conversely, nodes 6, 8, and 9, which have the weakest deep-layer shear (Fig. 4), show a later seasonal peak in late May or early June, when baroclinicity is much reduced. Diurnal variations between the nodes are more subtle, with slightly later peaks in the top row of the SOM (1700 LST) relative to the bottom row (1600 LST), and a slightly broader distribution in the left column relative to the right. Overall, it appears that seasonal variations in the NSE (often associated with different geographical regions; see below) dominate the variability in hodograph structure represented by the W-SOM.

Figure 9 shows the spatial distributions of soundings assigned to each node, as well as the distributions for all soundings and those rejected from the W-SOM. The all-soundings distribution shows two distinct maxima: one in the Great Plains (extending from northern Texas into central Nebraska), and a secondary weaker maximum in the Southeast (extending from central Mississippi into northern Alabama and southern Tennessee). This basic pattern exists in many of the nodes, but with significant variations in the precise location and relative magnitude of the two maxima, as well as additional maxima in some cases. For example, for node 1, the Great Plains maximum is shifted to the north and is weaker than the Southeast maximum, which is also broader in its zonal extent. Frequencies in both maxima are increased relative to the all-soundings case, with a corresponding reduction in frequencies around the Great Lakes and in the Northeast. In contrast, nodes 6, 8, and 9 all lack the Southeast maximum but feature higher frequencies in the Midwest (node 8) or Northeast (nodes 6 and 9).

The preceding analysis shows that similar shear profiles often occur in geographically distinct regions. To further explore this idea, we identify the subset of node 1 soundings located in the Great Plains (GP) and Southeast (SE) regions, defined for this purpose by simple latitude and longitude bounds (red-outlined boxes in Fig. 9). Of the 1039 soundings assigned to node 1, 216 (21%) fall in the GP region and 384 (37%) fall in the SE region. Figure 10 shows composite hodographs for the two regions, produced by averaging the raw (i.e., untransformed) wind components across these soundings. The hodographs have similar shapes (as we would expect, given that they are assigned to the same SOM node); however, the SE hodograph is shifted in the positive u and v directions, and rotated slightly clockwise, with respect to the GP hodograph. As a result, surface winds are from the south, rather than the east, upper-level winds are stronger, and estimated storm motion is faster. In addition, near-surface backing and weaker 0–1-km speed shear results in notably smaller SRH01 for the GP hodograph (304 m$^2$ s$^{-2}$) relative to the SE hodograph (460 m$^2$ s$^{-2}$).

Figure 11 shows corresponding temporal distributions for the two regions. We see that the rather broad seasonal and diurnal frequency maximum for node 1 in Fig. 8 reflects the superposition of quite distinct distributions associated with the GP and SE regions. Specifically, the GP region is characterized by a seasonal maximum centered in late May and a diurnal maximum centered at 1800 LST. In contrast, the SE region has earlier seasonal and diurnal frequency maxima, around late April and 1600 LST, respectively. It also features a pronounced secondary seasonal peak in early March. Further analysis reveals that the earlier peak is associated with a spatial maximum in southern Mississippi and Alabama, while the later peak is associated with a zonally elongated maximum (extending from eastern Oklahoma to the Carolinas) centered on the Mississippi–Alabama–Tennessee border (not shown). The later diurnal peak in the Great Plains may reflect the influence of strong capping inversions associated with elevated mixed layers, which are much more common in this region than in areas farther east (Ribeiro and Bosart 2018).

d. Rejected soundings

We now briefly consider the characteristics of the 574 soundings rejected from the W-SOM. Figure 5 reveals that...
these soundings are very diverse in terms of their associated parameter values. For example, the whiskers (5th–95th percentiles) of the BWD06 distribution extend from around 10 m s$^{-1}$ (below the 5th percentile for all nodes except node 9) to almost 45 m s$^{-1}$ (above the 95th percentile for all nodes). Relative to the all-soundings distribution, mean and median values are higher for the rejected soundings across all parameters, particularly for BWD06 and SRW910. Profiles with strong deep-layer shear are more likely to be rejected from the SOM since they are characterized by larger $u'$ and $v'$ at upper-levels, giving the potential for larger $e$ (section 4). The proportion of tornado, sigtor, and sighail reports associated with rejected soundings is consistent with the sounding proportion (i.e., close to 5%); however, there is a significantly higher proportion of sigwind reports (Fig. 6). Furthermore, Fig. 7 shows that rejected soundings have the second highest proportion of linear supercell modes (24% vs 15% for all soundings). In terms of their temporal distribution, the rejected soundings show an earlier and narrower seasonal peak and a broader diurnal peak (with higher frequencies during the overnight period) relative to the all-soundings distribution (Fig. 8). The spatial distribution, meanwhile, shows the highest
frequencies in the central Great Plains, with secondary maxima in the Mississippi and Ohio valleys (Fig. 9).

6. Thermodynamic SOM

We now consider the characteristics of the T-SOM. The reader is reminded that this was trained independently to the W-SOM (section 3) and thus the set of soundings assigned to each node is different. The correspondence between W- and T-SOM nodes is examined in section 7.

a. SOM nodes and parameters

Figure 12 shows skew $T$–log$p$ diagrams corresponding to each of the T-SOM nodes. Significant variations in stability and humidity are apparent across the SOM. From left to right, there is a pronounced decrease in surface temperature, dewpoint temperature, and MUCAPE, with the latter associated with a reduction in both LNB height and mean parcel buoyancy. From top to bottom, we observe a dramatic increase in humidity throughout the column, together with an increase in surface pressure. Node 9 is the only node to feature a significantly elevated MU parcel (at 1100 m AGL), associated with a stable boundary layer. As noted in section 4, this node accounts for a markedly smaller percentage of soundings, highlighting the rarity of severe supercells in environments lacking significant surface-based instability.

We note that, for all nodes, the low-level temperature and dewpoint profiles are somewhat atypical, both characterized by

---

Fig. 9. Heat maps showing the smoothed spatial distribution of all soundings (A), soundings rejected from the W-SOM (R), and soundings assigned to each of the W-SOM nodes (1–9). The red-outlined boxes for node 1 show the Great Plains and Southeast regions analyzed in Figs. 10 and 11.
increasing lapse rates with height, before a sudden decrease immediately above the LCL. This structure is an artifact of using buoyancy to define the thermodynamic profile, which implies a discontinuity where the parcel lapse rate transitions from dry adiabatic to pseudoadiabatic. More realistic-looking profiles are obtained if we average $T$ and $T_d$ across the soundings assigned to each node (thin colored lines in Fig. 12). By definition, this procedure also eliminates the positive dewpoint temperature bias seen in Fig. 3, resulting in a pronounced reduction in $T_d$ above 700 hPa in all but the very moist nodes. There is also a slight decrease in temperature in some nodes, consistent with a subtle positive bias in $T$ visible in Fig. 3. Though not indicated in Fig. 12, these changes also influence the parcel ascent, with the only noteworthy difference being an upward shift in the LPL for node 8, from 100 to 700 m AGL.

Similar to Fig. 5, Fig. 13 plots the distributions of several parameters—in this case MLCAPE, MLCIN, MLLCL height, 700–500-hPa lapse rate (LR75) and mean relative humidity (RH75), and precipitable water (PWAT)—for the soundings assigned to each node, together with the corresponding node values. For each parameter, the distribution for all soundings is also included for comparison. We see that the SOM captures reasonably well the variability in MLCAPE (Fig. 13a), with values ranging from less than 100 J kg$^{-1}$ for node 9 to over 2300 J kg$^{-1}$ for node 1. For both MLCAPE and MLCIN, node values tend to be biased slightly low, with the notable exception of MLCIN for node 9, which shows a large positive bias. There is limited bias in MLLCL height; however, the SOM underestimates variability in this parameter, with the node values encompassing less than one-third of the all-soundings distribution (a similar bias is seen in 0–1-km mean RH; not shown). This is clearly a limitation of the T-SOM since LCL height is an important predictor of tornado severity (e.g., Thompson et al. 2003; Craven and Brooks 2004; Thompson et al. 2012).

Turning to the remaining parameters, we see that the SOM does a reasonable job of capturing the variability in RH75 and PWAT, with node values spanning the interquartile range for both (Figs. 13e,f). However, it does a poor job for LR75, with node values encompassing only around a quarter of the all-soundings distribution (Fig. 13d). In particular, the SOM fails to capture very steep lapse rates, with the maximum value being 7.1°C km$^{-1}$ for node 1. For comparison, the 75th and 95th percentiles of the all-sounding distribution are 7.5 and 8.3°C km$^{-1}$, respectively. Lower-tropospheric lapse rates are an important control on supercell characteristics, particularly in low-CAPE environments (McCaul and Weisman 2001; Sherburn and Parker 2014) so this represents another limitation of the T-SOM. The difficulty in capturing variability in
FIG. 12. Skew T–log p diagrams for the nine T-SOM nodes. Solid red and blue lines show the environmental temperature $T$ and dewpoint temperature $T_d$ profiles, respectively, with darker-colored thin lines showing the corresponding composite profiles (created by averaging the soundings assigned to each node). The dotted red line shows the environmental virtual temperature $T_v$ profile, and the black line shows the virtual temperature of the most unstable parcel $T_{vp}$. Areas of positive and negative buoyancy are shaded light gray and dark gray, respectively. The corresponding CAPE and CIN values are given in the top-right corner together with the pressure at the LPL, LCL, LFC, and LNB.
both MLLCL and LR75 likely reflects the fact that the associated layers are relatively shallow in comparison with the full profiles used to train the SOM. When testing alternative input parameters, we found that using lapse rate, rather than buoyancy, in combination with RH unsurprisingly led to a much better representation of LR75 variability, but also a much poorer representation of variability in CAPE (not shown).

b. Hazards and storm characteristics

The distribution of tornado, sigtor, sighail, and sigwind reports among the T-SOM nodes is shown in Fig. 14. As in Fig. 6, markers at the top of each panel show where the percentage of reports is significantly higher or lower than what would be expected for an equivalent random sample. We see that nodes 1, 2, and 4 are associated with a disproportionately low number of tornado and sigtor reports, while the opposite is true for nodes 6 and 8. In comparing with Figs. 12 and 13, it is seen that the most striking difference between nodes that are most and least favorable for tornadoes is in the moisture profile, with the more tornadic nodes being characterized by much higher humidity throughout the column. This finding is somewhat surprising in light of previous studies, which generally show a strong decrease in RH above the boundary layer in tornado environments (e.g., Fawbush and Miller 1952; Doswell and Evans 2003). While it may in part be an artifact of variations in sighail across the SOM (see below), it could also reflect the tendency for more humid environments to feature low LCLs. Indeed, over one-half of the soundings assigned to nodes 1, 2, and 4 have MLLCLs < 1000 m, whereas the majority of those assigned to nodes 6 and 8 have MLLCLs > 1000 m (Fig. 13).

Variations in sighail across the T-SOM show almost the exact opposite pattern to those for tornadoes, with a high proportion of reports for nodes 1, 2, and 4 and a low proportion for Nodes 5–8 (Fig. 14). Nodes 1 and 2 collectively account for around one-quarter (26%) of all soundings but over a third (35%) of sighail reports. Conversely, node 8 accounts for 12% of all soundings but just 6% of sighail reports. These variations are notably larger than for the W-SOM (Fig. 6), suggesting that the thermodynamic environment is a stronger control on large hail formation than the wind profile [although recent modeling studies have demonstrated a relationship between deep-layer vertical wind shear and maximum hail size (Dennis and Kumjian 2017; Kumjian and Lombardo 2020)]. Comparison with Fig. 13 suggests that steep midlevel lapse rates are favorable for large hail, consistent with previous studies from the United States and Europe (Johnson and Sugden 2014; Púčik et al. 2015; Taszarek et al. 2017). Turning to the sigwind reports, we observe a high proportion for nodes 7 and 9 and a low proportion for node 3 and, in particular, node 1 (although this may again be an artifact of the high proportion of hail reports for node 1).

Figure 15 shows how storm severity, mesocyclone strength, and storm mode vary across the T-SOM. Consistent with the previous figure we find the highest relative frequency of tornadic and significantly tornadic storms associated with node 8, while the frequency of nontornadic storms is highest for nodes 1, 2, 4, and 9. Relative to the W-SOM (Fig. 7), we see much smaller variations in mesocyclone strength, but...
more significant variations in storm mode. Node 7 has the highest frequency of strong mesocyclones (46% vs 37% for all soundings), while node 3 has the highest frequency of weak mesocyclones (40% vs 35% for all soundings). In agreement with previous studies (e.g., Doswell and Evans 2003; Bunkers 2010), discrete supercells occur most frequently in dry environments (nodes 1–3) and least frequently in moist environments (nodes 7–9). The cell-in-cluster mode is most common for node 7 and least common for node 3, while the cell-in-line mode is most common for node 9 and least common for node 1.

c. Spatial and temporal distributions

The seasonal and diurnal distributions of soundings assigned to each of the T-SOM nodes are shown in Fig. 16. The basic pattern of a late afternoon to early evening and spring to early summer peak is maintained across all nodes, but there is significant variability in the shape of the distributions. There is a general trend from narrow to broad diurnal distributions from the top-left to bottom-right corners of the SOM, as well as a shift toward an earlier seasonal peak but later diurnal peak. Node 1 shows peak frequencies in May and June between 1500 and 1800 LST, whereas node 9 shows peak frequencies in March and April from 1700 to 2100 LST. The latter frequency maximum would mostly occur close to or after sunset, consistent with the presence of a surface stable layer and an elevated MU parcel (Fig. 12).

The spatial distributions for the T-SOM also show significant variations (Fig. 17). Nodes 1–4 all show a dominant maximum in the Great Plains, with generally lower frequencies in the Southeast relative to the all-soundings distribution. This is consistent with their relatively dry midlevels and steep lapse rates (Fig. 13), which may reflect the presence of elevated mixed layers (Ribeiro and Bosart 2018), as well as lower surface pressures (Fig. 12). Conversely, for node 8, which features the largest RH75 and smallest LR75 of all nodes, high frequencies are largely confined to the Southeast. Although not shown, we again find that distinct spatial maxima are associated with different temporal distributions (cf. Fig. 11).

d. Rejected soundings

Consistent with results for the W-SOM (section 6d), the 574 soundings rejected from the T-SOM encompass a diverse set of NSEs, as indicated by the wide parameter distributions in

---

**Fig. 14.** As in Fig. 6, but for the T-SOM.

**Fig. 15.** As in Fig. 7, but for the T-SOM.
Fig. 13. One noteworthy characteristic is the tendency for large LR75 (a mean value of 7.4°C km$^{-1}$) as compared with the nodes. This further highlights the inability of the SOM to capture steep midlevel lapse rates. The rejected soundings are associated with a disproportionately low number of tornado and sigtor reports, and a disproportionately high number of sighail and sigwind reports (Fig. 14). Consistent with this, the relative frequency of weak mesocyclones is higher than for any of the SOM nodes (Fig. 15). Relative to the full dataset, the seasonal distribution for rejected soundings shows a later peak (in June), while the diurnal distribution is more disperse (Fig. 16). The spatial distribution shows a maximum in the central Great Plains, with lower frequencies in the Southeast but higher frequencies in the Northern Plains and Upper Midwest relative to the all-soundings distribution (Fig. 17).

7. Joint frequencies

In this section, we briefly examine the association between the W-SOM and T-SOM. Figure 18 shows the number and percentage of soundings associated with every combination of W-SOM and T-SOM nodes, as well as combinations involving the rejected soundings for each SOM. To assess whether these values are significantly higher or lower than what would be expected for independent samples, we again make use of the hypergeometric distribution. The probability that the number of soundings $N_{ij}$ associated with W-SOM node $i$ and T-SOM node $j$ is:

$$P(N_{ij} | n, N, m) = \frac{\binom{N_{ij}}{n} \binom{N - N_{ij}}{m-n}}{\binom{N}{n}}$$
node $j$ exceeds some value $x$ is given by Eq. (6). In this case, $N_i$ is the total number of soundings assigned to W-SOM node $i$ and $N_j$ is the total number of soundings assigned to T-SOM node $j$. If $P_r < 0.01$ then we say that there is a disproportionately high number of soundings for this node combination. Conversely, if $P_r > 0.99$ then we say that the number of soundings is disproportionately low. These two cases are indicated in Fig. 18 by bold and italic font, respectively. Herein, we use the shorthand $W_jT_i$ to refer to the combination of W-SOM node $i$ and T-SOM node $j$.

The four most common node combinations, each comprising at least 2% of all soundings, are $W_6T_1$, $W_8T_1$, $W_9T_1$, and $W_9T_4$. Referring back to previous figures, we see that these environments occur predominantly in the Great Plains and are characterized by moderate shear and large instability. The least common node combination is $W_9T_9$, which is associated with moderate shear and marginal instability. We note that a disproportionately high number of soundings are rejected from both the W- and T-SOMs, indicating a statistically significant overlap between “atypical” dynamic and thermodynamic environments. Furthermore, of the 81 node combinations, 59 (73%) occur with a frequency that is significantly higher or lower than the expectation for independent samples. The strong association between the SOMs suggests there may be value in developing a combined thermodynamic and wind SOM (cf. NJ18); however, this is left as a topic for future work.

8. Summary and discussion

This study used a sample of 11,483 model proximity soundings from the SPC’s storm mode and environment database.

**Fig. 17.** As in Fig. 9, but for the T-SOM.
(Smith et al. 2012; Thompson et al. 2012) to explore the variability in near-storm environments (NSEs) associated with significant severe right-moving supercells in the CONUS. Soundings were clustered using self-organizing maps (SOMs), which have seen widespread use in meteorological and climate sciences, but have only recently been applied in severe thunderstorm research. Two SOMs, each with 3×3 nodes, were developed: a wind SOM (W-SOM), trained on profiles of wind velocity, transformed to align their 0–6-km shear vectors, and a thermodynamic SOM (T-SOM), trained on profiles of surface-based parcel buoyancy and relative humidity, together with the surface temperature and pressure. The result was a set of nine hodographs and nine skew T–logp diagrams that collectively represent the spectrum of significant severe right-moving supercell environments in the CONUS.

The W-SOM was found to provide an excellent representation of the variability in key kinematic parameters such as 0–6-km bulk wind difference and storm-relative helicity, although it was less able to capture very strong upper-level (9–10 km) storm-relative winds. The T-SOM similarly captured variability in mixed-layer CAPE, midlevel relative humidity, and precipitable water, but underestimated the variability in mixed-layer LCL height. It was also unable to represent very steep midlevel lapse rates, which are an important control on updraft intensity, particularly in low-CAPE environments (McCaul and Weisman 2001; Sherburn and Parker 2014).

Early exploratory analysis, it was found that the latter issue could be mitigated by training the T-SOM using lapse rates in place of buoyancy, but at the expense of a much worse representation of CAPE variability. Future work could explore the use of other thermodynamic variables as SOM inputs, to try and achieve a better all-round representation of key parameters.

The rich nature of the SPC dataset permitted exploration of a variety of aspects of the SOM nodes, including their spatial and temporal variability, their relationship with storm mode and mesocyclone strength, and their associated hazards (tornadoes, large hail, and straight-line winds). Key findings include the following:

- For the W-SOM just two nodes account for over 40% of significant tornado reports. Both of these nodes are characterized by large 0–6-km BWD and 0–1-km SRH, emphasizing the role of both deep-layer shear and low-level streamwise vorticity in supercell tornadogenesis.
- The proportion of significant hail reports varies more markedly across the T-SOM than the W-SOM, suggesting that the thermodynamic environment is a stronger control on large hail formation than the vertical shear profile.
- Mesocyclone strength varies strongly across the W-SOM and less dramatically across the T-SOM, highlighting the secondary role of thermodynamics (relative to dynamics) in updraft rotation.
- The relative frequency of discrete, cluster, and linear storm modes varies more across the T-SOM than the W-SOM, suggesting that the thermodynamic environment is more relevant to large-scale organization of supercells than the wind profile.
- Many nodes show multiple discrete spatial maxima, with corresponding temporal maxima, indicating that broadly similar environments occur in different parts of the country at different points in the seasonal and diurnal cycles.

Overall, the results presented here serve to highlight the diverse range of dynamic and thermodynamic environments that support right-moving supercells in the CONUS. However, it is important to emphasize that the SOMs do not represent the full spectrum of right-moving supercell NSEs encompassed by the SPC database. This can easily be seen when we compare environmental parameters for the nine SOM nodes with the distribution for the full set of soundings (Figs. 5 and 13). Inevitably, the tails of the distribution are not captured. This is a limitation with any data reduction technique, although it can be worse for SOMs due to the smoothing between neighboring nodes that gives rise to their characteristic topological ordering. Future work could explore other clustering approaches, such as the two-stage method used by Obba et al. (2015).

We see two potential applications for our results: first, in severe weather forecasting, and second, in idealized supercell simulations. On the forecasting side, one could envisage a product, similar to the Sounding Analog Retrieval System (SARS; Jewell 2010), that takes gridpoint soundings from a model analysis or forecast and matches them to the nodes of our two SOMs. Based on the relative frequency of hazards associated with each node (Figs. 6 and 14) it would then provide a probabilistic forecast of tornadoes, hail, and wind, conditional on the occurrence of a supercell. This idea was previously proposed by NJ18, but with a specific focus on distinguishing nontornadic, tornadic, and significantly tornadic
also thank attendees at the AMS of SOM configurations as part of a CLEX summer project. We acknowledge Rounak Dalal, who performed preliminary testing for Climate Extremes (CLEX; CE170100023). The authors Australian Research Council through the Centre of Excellence for this paper.

We suggest to reduce the number of nodes from 12 to im-

As discussed in the introduction, the majority of previous idealized supercell simulations have used a small number of observed or analytical proximity soundings to represent the near-storm environment. We contend that the quasi-idealized SOM profiles presented in Figs. 4 and 12 could serve as a useful intermediate between real-world complexity and analytical simplicity in future idealized modeling studies. In particular, they permit an exploration of the multidimensional convective parameter space without the need to independently vary a large number of potentially correlated variables. Say, for example, a researcher wants to explore the sensitivity of low-level mesocyclone strength to variations in a particular aspect of the microphysics parameterization. They could run an initial set of simulations using a single combination of W-SOM and T-SOM nodes and then a series of sensitivity tests, using other node combinations, to explore the impact of realistic variations in the storm environment. Since it would generally be unfeasible to simulate all 81 possible node combinations, we recommend focusing on those that occur most frequently (Fig. 18) or combinations involving the corner nodes, since these represent the most distinct profiles. Used in this way, we believe our results can contribute to both more realistic and more robust modeling of severe convective storms in the future.

Acknowledgments. Author Warren was supported by the Australian Research Council through the Centre of Excellence for Climate Extremes (CLEX; CE170100023). The authors acknowledge Rounak Dalal, who performed preliminary testing of SOM configurations as part of a CLEX summer project. We also thank attendees at the AMS 29th Conference on Severe Local Storms for useful feedback on an earlier version of this work and, in particular, Yvette Richardson for her suggestion to reduce the number of nodes from 12 to improve the interpretability of the results. We also thank Chris Nowotarski, two anonymous reviewers, and Editor Christopher Weiss for their insightful comments during the preparation of this paper.

Data availability statement. The filtered storm mode and environment database (comprising the 11 483 severe right-moving supercell proximity soundings and associated convective parameters, severe weather reports, and storm classifications) together with the SOM outputs (feature vectors, BMU labels, and QE values) is publicly available via the institutional repository “Bridges” (https://figshare.com/s/14bc7786a821522e975d).

APPENDIX

SOM Training

The SOM algorithm uses an iterative learning procedure to create a topologically ordered set of nodes that collectively describe the multidimensional distribution function of the input data. At each step of the iteration, the feature vectors describing the input data are presented to the SOM and assigned to the “closest” node (BMU) based on some distance metric (usually Euclidean distance). The reference vectors describing the nodes are then adjusted toward the feature vectors assigned to them. Critically, this adjustment is applied to both the BMU and surrounding nodes within some neighborhood kernel. This is how the SOM achieves its characteristic topological ordering. In the original stepwise recursive version of the SOM algorithm, each input sample is presented sequentially, with the SOM nodes updated every time. However, in practice the input samples can be presented in batch, with the node reference vectors adjusted based on the kernel-weighted mean of all feature vectors within their neighborhood. Relative to the sequential approach, the batch SOM algorithm results in much faster convergence and is less sensitive to how the nodes are initialized. Training of the SOM proceeds in two stages. During the first stage, the width of the neighborhood kernel is gradually reduced from an initial value that is close to the smallest dimension of the SOM to a final value that is typically a fraction of the grid spacing. The resulting map is then fine-tuned during the second stage, in which the width of the neighborhood kernel is maintained at its final value from the first stage.

For the MATLAB implementation of the batch SOM algorithm used in this study (Vesanto et al. 2000), the following parameters must be specified: 1) map size (i.e., the SOM dimensions); 2) map shape (“sheet,” “cylinder,” or “toroid”); 3) lattice type (“rectangular” or “hexagonal”); 4) initialization method (“random” or “linear”); (5) neighborhood function (“Gaussian,” “bubble,” “cutgrass,” or “ep”); 6) initial and final neighborhood widths; and 7) number of iterations in each training phase. Consistent with the vast majority of previous applications of SOMs in atmospheric science, we use a rectangular sheet lattice and a Gaussian neighborhood function. While Liu et al. (2006) recommended the “ep” (Epanechnikov) neighborhood function for training relatively small SOMs, Jiang et al. (2015) found comparable results to the Gaussian function. As discussed in section 3, the size of the SOMs was set to 3 × 3 as a trade-off between SOM quality (measured by QE and TE) and interpretability. Linear initialization (where the nodes are initialized as a regular, two-dimensional sequence of vectors along a hyperplane spanned by the first two principal components of the input data) was used as this leads to much faster ordering and convergence relative to a random initialization (where the nodes are initialized as random vectors) (Kohonen 2001). The initial and final neighborhood widths were set as 2 and 0.5, respectively, with 40 iterations for both training stages. These choices were informed by the recommendations of Kohonen (2013) and refined through trial and error to minimize QE and TE.
REFERENCES


——, and M. D. Parker, 2020: Insights into supercells and their environments from three decades of targeted radiosonde


