ELSEVIER

Atmospheric Research 67–68 (2003) 273–299

ATMOSPHERIC RESEARCH

www.elsevier.com/locate/atmos

Thunderstorm predictors and their forecast skill for the Netherlands

Alwin J. Haklander, Aarnout Van Delden*

Institute for Marine and Atmospheric Sciences, Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

Accepted 28 March 2003

Abstract

Thirty-two different thunderstorm predictors, derived from rawinsonde observations, have been evaluated specifically for the Netherlands. For each of the 32 thunderstorm predictors, forecast skill as a function of the chosen threshold was determined, based on at least 10280 six-hourly rawinsonde observations at De Bilt. Thunderstorm activity was monitored by the Arrival Time Difference (ATD) lightning detection and location system from the UK Met Office. Confidence was gained in the ATD data by comparing them with hourly surface observations (thunder heard) for 4015 six-hour time intervals and six different detection radii around De Bilt. As an aside, we found that a detection radius of 20 km (the distance up to which thunder can usually be heard) yielded an optimum in the correlation between the observation and the detection of lightning activity.

The dichotomous predictand was chosen to be any detected lightning activity within 100 km from De Bilt during the 6 h following a rawinsonde observation. According to the comparison of ATD data with present weather data, 95.5% of the observed thunderstorms at De Bilt were also detected within 100 km.

By using verification parameters such as the True Skill Statistic (TSS) and the Heidke Skill Score (Heidke), optimal thresholds and relative forecast skill for all thunderstorm predictors have been evaluated. It was found that Heidke reaches a maximum for more thundery index values than the TSS. In order to arrive at a single optimal threshold value, the TSS and Heidke were combined to form the Normalized Skill Score (NSS). When comparing forecast skill in a dichotomous forecasting scheme, the Lowest 100 hPa Lifted Index scores best, although other versions of the Lifted Index have relatively good performance as well. Even though the Boyden Index does not account for any moisture, it serves surprisingly well as a dichotomous thunderstorm predictor.

The Rank Sum Score (RSS) has been used to assess relative forecast skill without the use of a dichotomous forecasting scheme. Again, the Lowest 100 hPa Lifted Index scored best.

* Corresponding author. Tel.: +31-30-2533168; fax: +31-30-2543163.
E-mail address: a.j.vandelden@phys.uu.nl (A. Van Delden).

doi:10.1016/S0169-8095(03)00056-5

Finally, we have estimated the 'thundery case' probability as a function of the various thunderstorm predictors. We found that thunderstorm probability depends most on latent instability (especially near the surface), next on potential instability, and least on conditional instability.

Keywords: Thunderstorm predictors; Forecast skill; The Netherlands

1. Introduction

Since the late 1940s meteorologists have tried to assess thunderstorm risk with the help of thunderstorm indices and parameters that are deduced from the vertical temperature, moisture and wind profiles. As a simple example, we mention the 'Vertical Totals Index', which is defined as the temperature difference between the air at 850 hPa and the air at 500 hPa. For an index to become a useful thunderstorm predictor at a certain location, its statistical properties and the relation of these properties to thunderstorm occurrence must be determined for that particular location. This has been done for northeast Colorado (Schultz, 1989), for Cyprus (Jacovides and Yonetani, 1990) and for Switzerland (Huntrieser et al., 1997).

The main objective of this study has been to gain statistical information on many of the available indices and parameters, specifically for the Netherlands. By considering 32 different indices and parameters and comparing their ability to forecast thunderstorms, more insight is gained in the vertical, thermodynamic structure of the preconvective troposphere over the Netherlands.

The Royal Dutch Meteorological Institute (KNMI) provided us with 6-hourly, high-resolution rawinsonde observations at De Bilt. These were used to calculate the various index and parameter values. Thunderstorm activity was monitored by the Arrival Time Difference (ATD) sferics lightning location system from the UK Meteorological Office.

The structure of this article is as follows. In Section 2, we describe the data sets. In Section 3, we evaluate the reliability of the data obtained from the lightning detection system. In Section 4, we present the predictors and define the predictand. Section 5 is devoted to presenting the results, while Section 6 discusses the results. Finally, in Section 7, we conclude with a summary of the results and with some recommendations for future research.

2. Data sets

2.1. High-resolution rawinsonde data for De Bilt

The high-resolution data from the Royal Dutch Meteorological Institute (KNMI) gave us the opportunity to interpolate and calculate integral parameters with high precision. During the ascent of the rawinsonde, pressure, temperature and relative humidity are measured every 10 s. These data are then used to calculate the dew-point temperature and the geopotential height according to hydrostatic equilibrium. Wind direction and speed

data are calculated by tracking the rawinsonde's movement in those 10 s with the so-called LORAN-C (Long Range Navigation) navigation system.

The accuracy for pressure measurements is ±0.5 hPa, temperature measurements ±0.2 °C, relative humidity measurements ±3%, and wind measurements ±0.2 m/s (Väisälä, 1999). The mean ascent velocity of the rawinsondes is 5 to 6 m/s, implying a mean vertical resolution of 50 to 60 m. The reliability of temperature and relative humidity data is often overestimated. Ice deposition on the sensors can be the cause of too high relative humidity values. However, too low relative humidity values are observed quite frequently as well, especially at greater heights. Radiation effects can cause temperature errors of several degrees.

Exactly 11 495 soundings (00, 06, 12 and 18 UTC) were available between 6 January 1993, 12 UTC and 31 December 2000, 18 UTC, i.e., more than 98.5% of all possible soundings in that period. However, for some soundings, the wind or humidity data were incomplete or absent. We have calculated the values of all 32 thunderstorm predictors for each sounding. When it was impossible to calculate the value of a particular index, the index value for that sounding was marked 'unknown' and was excluded from further calculations. Integral parameters were calculated with integration steps of 1 hPa.

2.2. The Arrival Time Difference (ATD) sferics lightning location system

The UK Meteorological Office provided us with sferics/ATD data to give geographic and temporal spread of lightning strikes for the area 40 °W to 40 °E, 30 °N to 70 °N. Holt et al. (2001) describe it as an archive of reliable and continuous lightning fix data. The flash location accuracy for the Netherlands is estimated at 5 to 7 km (Lee, 1986), which is satisfactory for our purposes. The system probably performs better, because measured r.m.s. ATD errors are smaller than is assumed in this calculation. Like the rawinsonde dataset, this dataset is very large, covering the period between January 1990 and June 2000 with time intervals of approximately 6 min. During this period of more than 10 years, system outages or other disruptions have occurred only occasionally. We have tried to detect these disruptions by assuming a system outage whenever no sferics were reported in the whole detection area for more than 6 h. The outage is then assumed to have started at the time of the last sferic before the time gap, and stopped at the time of the first sferic after the time gap. An important limitation imposed within the ATD system is that the location of only 400 lightning flashes can be fixed per hour (Holt et al., 2001). Because of this limitation, an estimated 10% to 25% of all lightning flashes occurring in the atmosphere have been registered. However, our project only covers the occurrence of thunderstorms, not their intensity. Since we are concerned with a relatively large area and thunderstorms usually generate several strikes, this implies that the ATD system provides a way to determine whether or not there was any thunderstorm activity over a certain area during a certain period of time. Presently, the fixing rate of the ATD system has been increased to 10 000 flashes per hour. Key variables that were present in the original dataset are: time of the sferic [day, month, year, hour, minute], latitude of the sferic [°N] and longitude of the sferic [°E]. For each sferic, the distance between the sferic location and De Bilt was calculated. We assume a perfectly spherical earth, with a radius a of 6371 km. One degree latitude therefore equals $(2\pi a)/(360°) = 111.2$ km. One degree longitude represents a

longitudinal distance of $(2\pi a \cos\phi)/(360^\circ) = 111.2 \cos\phi$ km, where ϕ denotes degrees latitude. To account for the curvature effect, the mean latitude of De Bilt and the sferic location is taken in the cosine argument. The spatial resolution of the fixes is 0.01° latitude and 0.01° longitude. For De Bilt (52°6′N 5°11′E), this implies a resolution of 1.11 km in the meridional direction and 0.68 km in the zonal direction.

2.3. Hourly "present weather" (ww) codes for De Bilt

Hourly observations were available from 31 May 1997 to 2 October 2000. With these data we were able to compare the ATD lightning detection data with human-made thunderstorm observations at De Bilt. In the code, which is approved by the World Meteorological Organization, 100 different present and past weather types are represented by two-digit numbers (00–99) (see Dai, 2001 for a complete list). For our purposes, we are only interested in the codes that have anything to do with thunderstorm activity, i.e., code ww = 13, 17, 29, 91–99. If one of the codes ww = 17, 29, 91–99 is given at a particular hour, we know that thunder is heard. This usually implies that a lightning discharge has occurred within a radius of about 20 km around the observer, since thunder is audible up to a distance of 8 to 20 km (Huffines and Orville, 1999). For ww = 13 (lightning visible, no thunder heard), the distance between the observer and the thunderstorm is much more uncertain. Furthermore, ww = 13 will occur much more frequently during nighttime, when lightning is visible at much greater distances. For the sake of consistency, we have chosen to exclude ww = 13 from the list of thundery present weather codes, defining a thunderstorm observation as a situation where thunder was heard.

3. Comparison between hourly observations and the ATD system

We have mentioned that the ATD system has registered an estimated 10% to 25% of all lightning flashes. Therefore, it is useful to compare the human-made synoptic observations with the simultaneous lightning detection data. To make this comparison, we have divided the period for which both datasets are available into 6-h intervals. The overlapping period for the ATD and synoptic datasets is 31 May 1997 to 30 June 2000. This is a time period of 1127 days, i.e., 4508 six-hour time intervals. Not all these time intervals are taken into account, for we only want to consider those intervals for which both datasets are complete. All 6-h intervals partly or entirely overlapping with an ATD system outage are omitted. Six-hour intervals, for which not all 6 hourly present weather observations are available, are omitted from this comparison as well. With these conditions set, we are left with 4015 (89%) 6-h time intervals, which is still a period of more than a thousand days.

Let us consider such a time interval of 6 h, e.g., from 0 to 6 UTC. If thunder was heard in this time interval, there will at least be a thundery synoptic code (ww = 17, 29, 91–99) at 1, 2, 3, 4, 5, or 6 UTC. We define the variable SYN to indicate whether or not thunder was heard: SYN = "Yes" if so, and SYN = "No" if not. The synoptic data are compared with lightning detection data for circular areas around De Bilt with different radii R. We have chosen to consider radii of $R = 10$, 20, 50, 100, 150 and 200 km around De Bilt, for which we define six variables: $LD_{10\ km}$, $LD_{20\ km}$, $LD_{50\ km}$, $LD_{100\ km}$, $LD_{150\ km}$, and $LD_{200\ km}$ (see

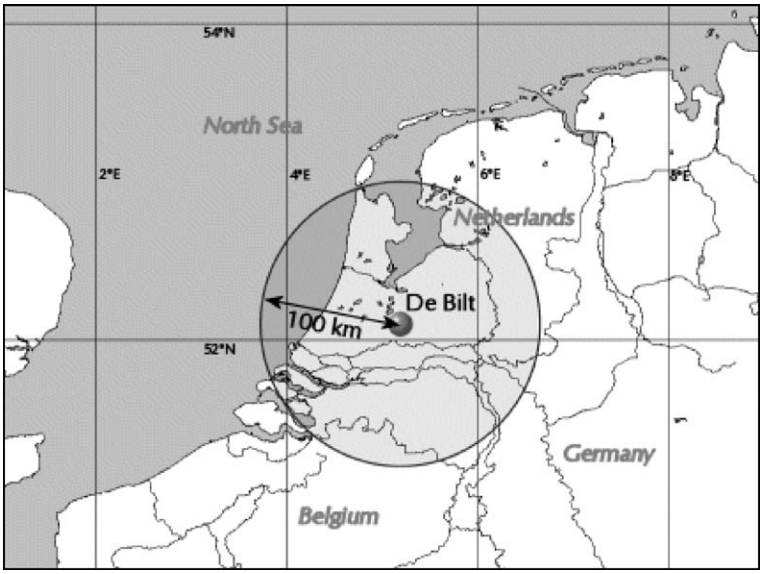

Fig. 1. Map of the detection area surrounding De Bilt, using a detection radius of 100 km.

Fig. 1 for a map of the area). LD_{xkm} = "Yes" if one or more sferics were fixed between 00:00 and 05:59 UTC within a distance of x km from the station; LD_{xkm} = "No" if this is not the case.

Table 1 shows the POD, FAR, CSI, TSS and Heidke values for all six detection radii. These verification parameters are defined in Appendix B. Obviously, the POD increases with increasing detection radii and already reaches 93.9% for $R = 50$ km. It is interesting to note the difference between the TSS, the CSI, and Heidke. The TSS reaches its highest value at $R = 50$ km, and a high FAR does not seem to affect it much. The CSI and Heidke reach their maximum values at $R = 20$ km, which confirms the distance at which thunder can still be heard. With an estimated ATD location error of 5–7 km, the FAR of 11.5% at $R = 10$ km seems acceptable. Furthermore, we should be aware of the fact that synoptic observations are imperfect. It is well possible that not all thunder that occurs within 10 km from De Bilt is heard. Van Delden (2001) discusses the many factors that may have a negative influence on the reliability of observations of thunder and lightning.

Table 1
Skill scores of the ATD system for various detection radii compared with thunderstorm observations.

Detection radius (km)	POD [%]	FAR [%]	CSI [%]	TSS	Heidke
10	34.8	11.5	33.3	0.35	0.49
20	69.7	18.6	60.1	0.69	0.74
50	93.9	57.1	41.8	0.90	0.57
100	95.5	74.4	25.3	0.86	0.37
150	97.0	81.6	18.3	0.82	0.27
200	97.7	85.1	14.8	0.79	0.21

4. Definition of predictors and predictand

4.1. Use of thunderstorm indices as dichotomous predictors

A forecast is always based on certain information, which is called the predictor. This predictor may be the output of an atmospheric model, but might as well be a coin tossing method. The verifying observations are called the predictand (Von Storch and Zwiers, 1999). Whereas our predictand—thunderstorm occurrence vs. no thunderstorm occurrence—is dichotomous, the various predictors can take on a very large (or even continuous) set of values. One way to assess the forecast skill of a thunderstorm predictor is to make it dichotomous as well, i.e., to divide its range of values into two parts in which a thunderstorm event is either forecast or not forecast. Thus, we assume that thunderstorm probability increases monotonically with either decreasing or increasing index values, where we choose some benchmark value to distinguish between 'thunderstorm' or 'no-thunderstorm' forecasts. This benchmark value is called an upper threshold value if we would only forecast a thundery case when the index value lies at or below that benchmark value, and a non-thundery case is forecast otherwise. If we would forecast a thundery case at or above this benchmark value and a non-thundery case otherwise, we call it a lower threshold value.

Suppose, for instance, that the Vertical Totals Index (VT), defined as the temperature at 850 hPa minus the temperature at 500 hPa, ranges between 10 and 35 °C at a certain location. Obviously, high VT values increase the chance for conditional instability, and therefore, increase thunderstorm probability. Now, using a lower threshold value of 25 °C, we could turn VT into a dichotomous thunderstorm predictor, defining a thunderstorm event forecast by $VT \geq 25\,^{\circ}C$ and a non-event forecast by $VT < 25\,^{\circ}C$.

When using the various thunderstorm indices and parameters in such a two-class categorical forecasting scheme, it makes sense to associate them with their optimal threshold values. For that purpose, the forecast skill of the predictor is evaluated for many different associated threshold values within the predictor's range. Eventually, the threshold value with the best performance is chosen to be associated with that particular predictor.

4.2. Choosing a predictand

Since all thunderstorm predictors were derived from the 6-hourly rawinsonde data for De Bilt, the thunderstorm index and parameter updates were available every 6 h. Hence, it seems natural to choose a predictand that describes the 6 h after a particular rawinsonde observation. It is assumed that the rawinsonde observations at De Bilt represent the preconvective conditions at any point within a 100-km distance from De Bilt. We have shown that, using a detection radius of 100 km, 95.5% of all synoptic thunderstorm observations at De Bilt are confirmed by the ATD system. Consequently, we have chosen a dichotomous predictand that answers the following question. Was at least one sferic fixed by the ATD system at less than 100 km from De Bilt (Fig. 1) during the 6 h after the rawinsonde observation? From now on,'thundery' cases will be defined as rawinsonde observations for which this question can be answered positively.

5. Results

In this section, the general results of our study are shown. We use the Surface-based Lifted Index as an example.

5.1. Threshold value test

Fig. 2 shows a scatter-plot for the Surface-based Lifted Index (LI_{sfc}). For 10433 cases, LI_{sfc} values could be calculated and compared with ATD lightning detection data. Low index values are obviously associated with higher thunderstorm probability and vice versa. In the summertime, values are generally lower, implying more latent instability at the surface than in winter.

If we would like to use LI_{sfc} as a dichotomous thunderstorm predictor, with a certain upper threshold value α for a 'thundery case' forecast, we could try to estimate the optimal value for α from Fig. 2. A more objective way to find α would be compute the skill score (see Appendix B) for the total range of index values whereupon the α with the best skill score is chosen. For the 10433 rawinsonde observations, LI_{sfc} values ranged between -8.9 and 32.8 °C, which implies a total range size of 41.7 °C. Starting with the minimum LI_{sfc} value of -8.9 °C, we test the total range stepwise with increments of 1‰ of the total range size, until we reach the maximum LI_{sfc} value of 32.8 °C. A contingency table can be constructed for each α, where only index values $\leq \alpha$ imply that a thundery case is forecast.

Fig. 3 shows the values of the four entries in our contingency table as a function of α. The four lines are easily interpreted in terms of dots in Fig. 2. The line labeled *h* denotes

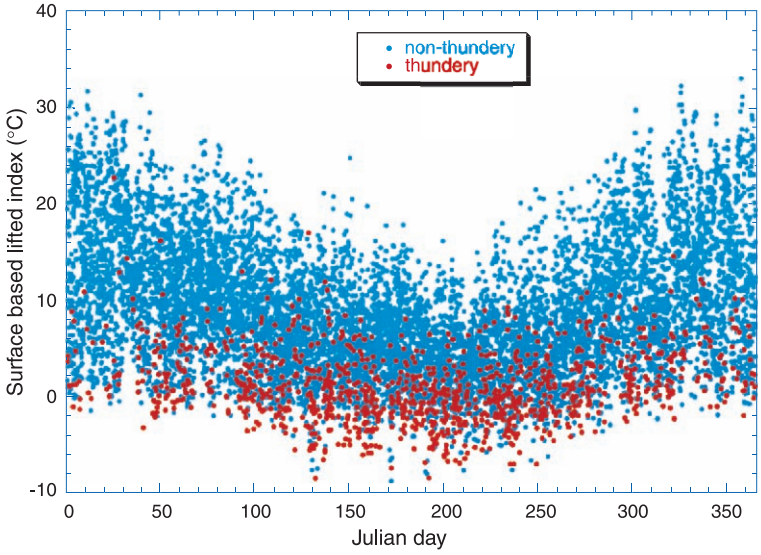

Fig. 2. The dependence of thunderstorm occurrence on the Surface-based Lifted Index and the Julian day. The 1131 thundery cases are indicated by red dots, the 9302 non-thundery cases by cyan dots.

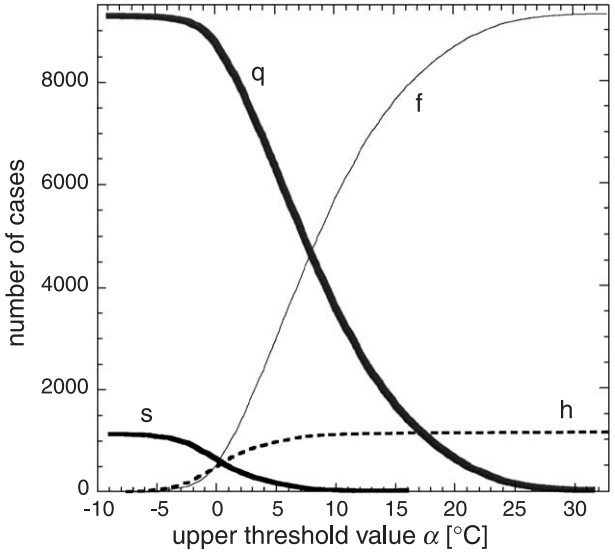

Fig. 3. Contingencies (*h*: number of hits; *s*: number of surprises; *f*: number of false alarms; *q*: number of quiescent cases) as a function of the upper threshold value (α) of the Surface-based Lifted Index.

the number of red dots (thundery cases) at LI_{sfc}-values $\leq \alpha$, the line labeled *s* the number of red dots at LI_{sfc}-values $>\alpha$, the line labeled *f* the number of cyan dots (non-thundery cases) at LI_{sfc}-values $\leq \alpha$, and the line labeled *q* the number of cyan dots at LI_{sfc}-values $>\alpha$. Because of the large datasets, very smooth curves are obtained. The thundery and non-thundery medians can also be read from Fig. 3. For the LI_{sfc}, they are 0.5 and 8.0 °C, respectively.

Now, suppose we would forecast an event—a thundery case—at any LI_{sfc} value, setting $\alpha = 32.8$ °C. Then, there would be no surprises, implying a POD of 100%. All non-thundery cases would be false alarms, and the FAR would equal the non-event frequency. The CSI would equal the event frequency, since all events would be hits and all non-events would be false alarms. Finally, since $s = 0$ and $q = 0$, both TSS and Heidke would be zero.

Since both the TSS and the Heidke score are frequently used in literature, we have decided to use the *best of both*. A new skill score, the Normalized Skill Score (NSS), is defined as follows:

$$\mathrm{NSS}(\alpha) = \frac{1}{2}\left(\frac{\mathrm{TSS}(\alpha)}{\mathrm{TSS}_{\max}} + \frac{\mathrm{Heidke}(\alpha)}{\mathrm{Heidke}_{\max}}\right),$$

where TSS_{max} is the maximum TSS value for any α and $Heidke_{max}$ is the maximum Heidke score for any α. Note that the maximum values for TSS and Heidke do not have to be reached at the same α.

Fig. 4 shows the result of an upper threshold value test for LI_{sfc}. The POD increases monotonically with increasing α, since for higher α, more and more events will have

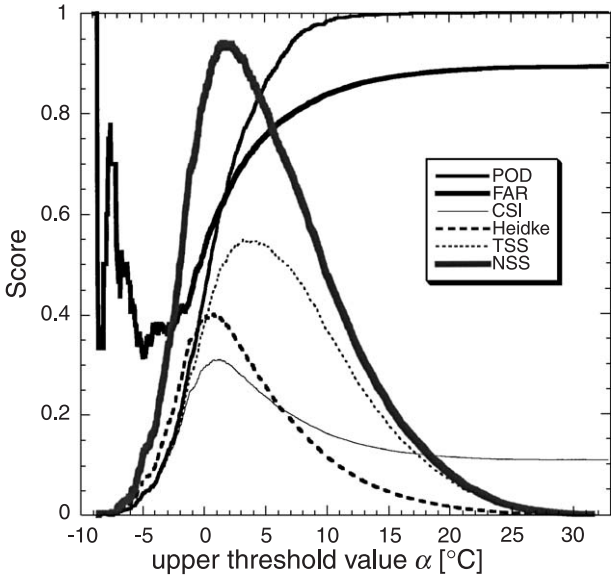

Fig. 4. Assessment of the quality of a two-class categorical forecasting scheme as a function of α, using several verification parameters for the Surface-based Lifted Index.

been forecasted. However, both the number of false alarms and the number of event forecasts increase with increasing α, and the FAR does not increase monotonically. In fact, for our dataset, the FAR decreases with increasing α for $\alpha \leq -5.0$ °C. This erratic behavior is due to the fact that the FAR only considers event forecasts ($h+f$), i.e., cases for which $LI_{sfc} \leq \alpha$. From Fig. 3 we can see that the FAR will be based on only a few event forecasts at low α, making this score very sensitive to individual cases.

The CSI, the ratio of hits to the total number of events and false alarms reaches its maximum of 31.1% at $\alpha = 1.1$ °C. However, the TSS suggests a much higher optimal threshold value, at $\alpha = 3.3$ °C where the TSS reaches 55.0%. Furthermore, according to the TSS, it does not matter if we would choose $\alpha = 1.1$ °C or $\alpha = 7.5$ °C; the TSS would be 47.6% for both α. Nonetheless, considering Fig. 2, $\alpha = 1.1$ °C does seem to be a better upper threshold value than $\alpha = 7.5$ °C, as was indicated by the CSI.

The maximum Heidke Skill Score of 40.0% at $\alpha = 0.8$ °C suggests an even lower optimal threshold than the CSI. At $\alpha = 0.8$ °C, the TSS is 45.1%, which was also reached at an α as high as 8.2 °C! Although the TSS and Heidke are both frequently used in literature as forecast skill parameters, there seems to be quite a difference between their characteristics. Apparently, the TSS pursues a relatively high POD, whereas the Heidke Skill Score attempts to reduce the FAR to reasonable values.

We therefore turn to the Normalized Skill Score, to get the best of both "worlds". The maximum NSS of 94.1% was reached at $\alpha = 1.6$ °C, where the TSS and Heidke were at 91.3% and 96.9% of their maximum scores, respectively. There were 711 thundery and 1177 non-thundery cases with $LI_{sfc} \leq 1.6$ °C. On the other hand, there were 420 thundery

and 8125 non-thundery cases with LI_{sfc}>1.6 °C. We would then have 711 hits, 1177 false alarms, 420 surprises and 8125 quiescent cases. This implies POD = 62.9%, FAR = 62.3%, CSI = 30.8%, TSS = 50.2%, and Heidke = 38.8%.

5.2. Optimal thresholds according to the Normalized Skill Score

The NSS is used to find optimal benchmark values for all parameters. Results are shown in Table 2. As mentioned above, both TSS and Heidke will be near their maximum values where the NSS reaches its maximum. Since the NSS is normalized, NSS values near 100% only imply that the TSS and Heidke agree on that particular threshold. We still have to turn to the TSS and Heidke scores to assess forecast skill, when that particular

Table 2
Optimal thresholds according to the maximum NSS, from high to low μ values. Thundery cases are always and only forecast when the value of a certain index satisfies the condition given in the second column. The four rightmost columns give the associated number of hits, surprises, false alarms and quiescent cases, respectively

Index	Threshold	μ	Heidke	TSS	POD	FAR	CSI	h	s	f	q
LI_{100}	≤ 3.0 °C	0.95	0.42	0.54	0.66	0.59	0.34	748	383	1099	8203
LI_{50}	≤ 3.1 °C	0.94	0.41	0.55	0.69	0.62	0.33	781	350	1261	8041
LI_{MU}	≤ 1.4 °C	0.86	0.38	0.55	0.71	0.65	0.31	800	330	1479	7820
$SWISS_{12}$	≤ 1.3	0.83	0.38	0.53	0.67	0.65	0.30	774	354	1420	7854
LI_{sfc}	≤ 1.6 °C	0.80	0.39	0.50	0.63	0.62	0.31	711	420	1177	8125
BOYD	≥ 94.6	0.80	0.33	0.57	0.80	0.70	0.28	911	222	2149	7155
ADED2	≥ − 0.9 °C	0.80	0.39	0.50	0.63	0.62	0.31	707	424	1164	8140
KO	≤ 1.9 °C	0.67	0.31	0.52	0.74	0.71	0.26	830	299	2001	7293
PII	≥ − 0.52 °C km^{-1}	0.65	0.34	0.47	0.63	0.67	0.28	712	417	1454	7839
$CAPE_{MU}$	≥ 168 J kg^{-1}	0.63	0.37	0.43	0.53	0.61	0.29	589	531	922	8250
SHOW	≤ 4.2 °C	0.61	0.31	0.50	0.70	0.71	0.26	787	343	1900	7401
TT	≥ 46.7 °C	0.60	0.29	0.51	0.76	0.73	0.25	860	270	2315	6986
ADED1	≥ − 2.2 °C	0.58	0.32	0.46	0.63	0.69	0.26	713	417	1577	7724
YON_{MOD}	≥ 1.0	0.54	0.28	0.49	0.73	0.73	0.24	828	302	2276	7025
RACK	≥ 30.8 °C	0.53	0.27	0.49	0.74	0.74	0.24	837	293	2347	6954
YON	≥ − 1.0	0.53	0.27	0.48	0.72	0.73	0.24	818	312	2251	7050
SI	≥ 39.2 °C	0.51	0.26	0.49	0.75	0.74	0.24	852	278	2471	6828
THOM	≥ 17.2 °C	0.51	0.29	0.45	0.63	0.71	0.25	715	415	1736	7563
TEI_{925}	≤ − 1400 J kg^{-1}	0.51	0.27	0.48	0.74	0.74	0.24	830	299	2377	6916
$CAPE_{sfc}$	≥ 173 J kg^{-1}	0.50	0.35	0.38	0.47	0.61	0.27	521	599	809	8363
BRAD	≤ 1.8 °C	0.47	0.26	0.46	0.70	0.74	0.23	787	342	2220	7074
JEFF	≥ 26.0 °C	0.43	0.25	0.45	0.70	0.75	0.23	791	339	2337	6962
TEI_{850}	≤ 200 J kg^{-1}	0.40	0.24	0.45	0.73	0.76	0.22	820	309	2565	6729
$SWISS_{00}$	≤ 7.6	0.40	0.27	0.41	0.61	0.72	0.23	683	445	1798	7472
$CAPE_{50}$	≥ 16 J kg^{-1}	0.39	0.29	0.38	0.52	0.69	0.24	580	541	1283	7890
VT	≥ 25.3 °C	0.37	0.23	0.45	0.74	0.76	0.22	834	297	2714	6590
KI_{MOD}	≥ 32.5 °C	0.36	0.29	0.36	0.48	0.68	0.24	543	587	1141	8158
CT	≥ 20.7 °C	0.33	0.21	0.44	0.78	0.78	0.21	877	253	3082	6219
$CAPE_{100}$	≥ 11 J kg^{-1}	0.32	0.28	0.35	0.48	0.69	0.23	533	588	1166	8007
KI	≥ 21.0 °C	0.18	0.21	0.35	0.59	0.77	0.20	667	463	2193	7106
SWEAT	≥ 134	0.10	0.18	0.34	0.62	0.79	0.18	664	412	2522	6682
DCI	≥ 6.8 °C	0.07	0.22	0.28	0.44	0.74	0.19	498	632	1452	7849

threshold value is used. For the purpose of determining the relative performance of the benchmark values according to the NSS, we again combine the TSS and the Heidke Skill Score. We rescale the TSS and Heidke values in Table 2, so that the maximum scores become 1 and the minimum scores become 0. We define a new parameter μ,

$$\mu \equiv \frac{1}{2}\left(\frac{\mathrm{TSS}-\mathrm{TSS}_{\mathrm{poorest}}}{\mathrm{TSS}_{\mathrm{best}}-\mathrm{TSS}_{\mathrm{poorest}}}\right)+\frac{1}{2}\left(\frac{\mathrm{Heidke}-\mathrm{Heidke}_{\mathrm{poorest}}}{\mathrm{Heidke}_{\mathrm{best}}-\mathrm{Heidke}_{\mathrm{poorest}}}\right)$$

where TSS is the TSS for the particular index in Table 2, $\mathrm{TSS}_{\mathrm{best}}$ is the highest TSS in Table 2 ($\mathrm{TSS}_{\mathrm{best}}=0.57$), $\mathrm{TSS}_{\mathrm{poorest}}$ is the lowest TSS in Table 2 ($\mathrm{TSS}_{\mathrm{poorest}}=0.28$), Heidke is the Heidke Skill Score for for the particular index in Table 2, $\mathrm{Heidke}_{\mathrm{best}}$ is the highest Heidke in Table 2 ($\mathrm{Heidke}_{\mathrm{best}}=0.42$), and $\mathrm{Heidke}_{\mathrm{poorest}}$ is the lowest Heidke in Table 2 ($\mathrm{Heidke}_{\mathrm{poorest}}=0.18$).

The four versions of the Lifted Index score best in a thundery/non-thundery forecasting scheme. SWISS_{12} is highly dependent on $\mathrm{LI}_{\mathrm{sfc}}$, and therefore should not receive too much credit for its relatively good performance. Right after $\mathrm{LI}_{\mathrm{sfc}}$ comes the Boyden Index. Note, however, that BOYD scores much better according to the TSS than according to Heidke. This is because TSS focusses more on a high POD than on a low FAR while Heidke emphasizes a low FAR. Still, even the Heidke Score is not too bad.

5.3. Rank Sum Scores (RSS) and climatology

The forecast skill of an index can also be assessed in another way, without dividing its range into 'thundery' and 'non-thundery' classes. If an ordered list is compiled with high to low values of a thunderstorm index, its Rank Sum Score can be defined as the absolute difference between the mean rank of all 'no thunderstorm' cases and the mean rank of all 'thunderstorm' cases, divided by the sample size (Schultz, 1989). The more a predictor is capable of separating the thundery from the non-thundery cases, the more forecast skill it will have and the higher its Rank Sum Score will be. Different Rank Sum Scores can only be compared if event frequencies are equal. Schultz (1989) uses the Rank Sum Score in order to compare the performances of eight forecasting schemes for severe and significant weather. He states that the Rank Sum Score "has the useful characteristic of being independent of the shapes of the frequency curves that characterize the populations from which the samples were taken". An advantage of using Rank Sum Scores is that the RSS enables us to assess relative performance of the various thunderstorm predictors without choosing any threshold values.

The rightmost column in Table 3 shows the RSS values for all 32 thunderstorm predictors. Although this comparison of forecast skill uses quite a different method, once again, LI_{100} and LI_{50} are found on top of the list. The different CAPE versions have relatively poor Rank Sum Scores, since $\mathrm{CAPE}_{\mathrm{i}}$ cannot take on negative values. This means that forecast skill is underestimated for $\mathrm{CAPE}_{\mathrm{i}}$. We will see in the next section that these parameters can still be very useful in evaluating thunderstorm probability. Many other useful statistics are shown in Table 3.

Table 3
Other statistics for all 32 thunderstorm indices and parameters. The sample size for each index can be derived from Table 2. Medians, means and standard deviations were evaluated for all cases, for all thundery cases, and for all non-thundery cases, separately (Mdn = Median, Stddv = Standard deviation, RSS = Rank Sum Score)

Index	Extreme values		All cases			Thundery cases			Non-thundery cases			RSS
	Max	Min	Mdn	Mean	Stddv	Mdn	Mean	Stddv	Mdn	Mean	Stddv	
LI_{100} (°C)	30.2	− 4.8	7.7	8.38	5.56	1.9	2.51	2.83	8.4	9.09	5.39	0.37
LI_{50} (°C)	31.3	− 5.7	7.8	8.54	5.92	1.8	2.37	3.00	8.6	9.29	5.74	0.37
$SWISS_{12}$	42.4	− 10.5	7.4	8.63	8.55	− 0.6	− 0.11	4.35	8.6	9.70	8.32	0.36
LI_{sfc} (°C)	32.8	− 8.9	7.1	8.01	6.82	0.5	0.97	3.74	8.0	8.87	6.61	0.36
ADED2 (°C)	5.3	− 24.0	− 4.3	− 5.02	4.49	− 0.3	− 0.61	2.30	− 4.8	− 5.55	4.39	0.36
LI_{MU} (°C)	20.7	− 8.9	4.5	4.93	4.40	0.0	0.19	3.09	5.1	5.50	4.18	0.35
BOYD	100.8	81.6	93.5	93.30	2.19	95.4	95.51	1.34	93.2	93.04	2.12	0.35
KO (°C)	29.2	− 18.1	5.7	6.16	6.28	0.0	− 0.16	4.05	6.6	6.93	6.06	0.34
PII (°C km^{-1})	3.26	− 7.43	− 1.78	− 1.91	1.58	− 0.28	− 0.32	1.01	− 2.01	− 2.10	1.53	0.34
SHOW (°C)	24.0	− 5.1	7.4	7.91	4.66	3.0	3.42	2.75	8.0	8.46	4.55	0.33
ADED1 (°C)	3.0	− 15.9	− 4.4	− 4.81	2.95	− 1.8	− 2.06	1.70	− 4.8	− 5.15	2.89	0.33
TT (°C)	64.3	− 11.1	42.4	40.7	10.5	49.6	49.6	5.4	41.4	39.6	10.4	0.32
THOM (°C)	41.7	− 100.7	3.6	− 0.8	22.2	20.9	18.5	11.2	0.7	− 3.2	22.1	0.32
YON_{MOD}	8.2	− 37.7	− 1.3	− 3.58	7.21	2.4	2.03	2.77	− 1.9	− 4.27	7.28	0.32
YON	6.1	− 39.6	− 3.3	− 5.60	7.19	0.4	− 0.02	2.80	− 3.9	− 6.28	7.27	0.32
RACK (°C)	41.2	15.4	28.9	28.87	3.92	32.3	32.66	3.03	28.4	28.41	3.76	0.32
TEI_{925} (10^2 J kg^{-1})	310	− 269	28	34	86	− 42	− 47	60	39	44	83	0.31
$CAPE_{MU}$ (J kg^{-1})	4427	0	5	104	282	188	397	542	1	68	204	0.31
SI (°C)	60.6	− 58.5	29.9	26.3	19.8	44.8	42.9	9.8	27.5	24.3	19.7	0.31
JEFF (°C)	35.7	− 12.1	21.3	19.78	8.79	28.3	27.12	4.63	20.1	18.89	8.76	0.30
BRAD (°C)	14.6	− 4.7	3.4	3.78	2.95	1.0	1.21	1.83	3.7	4.09	2.91	0.30
KI_{MOD} (°C)	46.8	− 51.5	20.4	17.0	16.2	32.0	30.0	8.6	18.5	15.4	16.2	0.30
TEI_{850} (10^2 J kg^{-1})	294	− 263	39	45	77	− 18	− 23	57	48	53	75	0.29
VT (°C)	34.4	11.1	24.0	23.79	3.47	26.7	26.76	2.48	23.6	23.43	3.39	0.29
$SWISS_{00}$	39.0	− 3.5	12.0	12.58	6.14	6.5	7.53	4.66	12.7	13.19	6.01	0.28
CT (°C)	32.1	− 30.8	19.2	16.88	8.83	23.1	22.87	3.83	18.6	16.16	8.99	0.28
KI (°C)	37.8	− 74.9	11.7	7.7	18.1	23.2	20.8	9.5	9.7	6.1	18.2	0.27
$CAPE_{sfc}$ (J kg^{-1})	4427	0	6	96	271	141	361	540	4	63	193	0.24
$CAPE_{50}$ (J kg^{-1})	2036	0	0	21	87	18	87	189	0	12	60	0.23
DCI (°C)	35.8	− 81.4	− 7.3	− 9.2	18.1	4.5	3.5	14.0	− 8.6	− 10.7	17.9	0.23
SWEAT	454	0	97	113	69	152	163	78	92	107	65	0.22
$CAPE_{100}$ (J kg^{-1})	1633	0	1	15	71	9	63	153	0	9	49	0.22

5.4. The use of indices to estimate the probability of a thundery case

We have discussed the use of indices in a dichotomous forecasting scheme. But why not estimate the 'thundery case probability' (TCP) as a function of each thunderstorm parameter? We again consider the Surface-based Lifted Index, LI_{sfc}. To describe the TCP as a function of LI_{sfc}, we calculate the percentage of thundery cases for a number of cases with approximately the same LI_{sfc}, for the total range of LI_{sfc} values. Obviously, the probability distribution will become more erratic if only a few cases around a certain LI_{sfc} are considered. To obtain a somewhat smooth distribution, we have chosen to base the estimate of TCP at a certain LI_{sfc} value on 200 cases around that value. First, an ordered list of high to low LI_{sfc} values is compiled, which are all marked 'thundery' or 'non-thundery'. We then take the 1st to the 200th cases in that ordered list and calculate the mean thundery event occurrence (TCP) as well as the mean index value and standard deviation for those 200 cases. We repeat this procedure for the 2nd to the 201st cases and so on. If we have a total of N index values in our dataset, we obtain $N-199$ mean values, standard deviations and corresponding thundery case probabilities. Since $N=10\,433$ for LI_{sfc}, this implies that the probability plot will contain 10 234 data points.

The plot is shown in Fig. 5. As expected, the TCP clearly increases with decreasing LI_{sfc}. Standard deviations increase also, as was already illustrated in Fig. 2, but they are still relatively small, owing to the enormous sample size. A maximum TCP of 65.5% was reached, since 131 out of 200 cases with a mean LI_{sfc} of -4.1 °C and a standard deviation of 0.9 °C were thundery. It should be noted that the total thundery case frequency was 1131 out of 10 433 cases (10.8%) (Table 2).

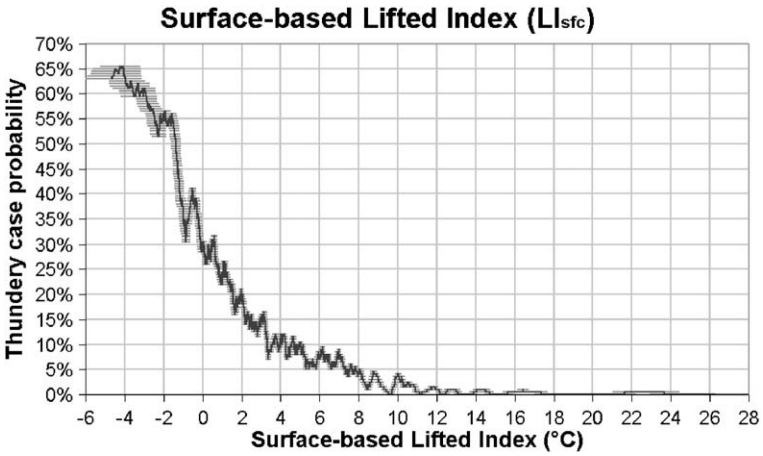

Fig. 5. Thundery case probability (TCP) as a function of the Surface-based Lifted Index. An ordered list of high to low LI_{sfc} values is made, and each case in that list is marked 'thundery' or 'non-thundery'. For the first 200 LI_{sfc}-values, the mean LI_{sfc}-value, the standard deviation and a mean thundery case probability were determined. Subsequently, this was done for the 2nd to 201st cases in the list, and so on. The resulting thundery case probabilities are plotted as a function of the mean LI_{sfc} values for each set of 200 cases, where the error bars denote the standard deviations.

5.5. An estimated probability table

Bringing some of the information in Fig. 5 into tabular form facilitates the comparison of thundery case probability as a function of the various indices. With some criteria, it is possible to estimate benchmark values of all thunderstorm predictors for which the TCP reaches a certain critical percentage Π, such as 50%, 55%, 60%, 65%. As an example, LI_{sfc} is again considered. For each critical TCP, Π, we sort out the least thundery index value for which the TCP reaches Π, and the least thundery index value beyond which the TCP stays $\geq \Pi$, according to the procedure in the previous section. So, for each Π, we have a low and a high index value, λ_{low} and λ_{high}, and their associated standard deviations, σ_{low} and σ_{high}. We then estimate the benchmark value λ for that Π, and its 'uncertainty' σ, by taking

$$\lambda(\Pi) \equiv \frac{1}{2}[(\lambda_{high} + \sigma_{high}) + (\lambda_{low} - \sigma_{low})],$$

$$\text{with } \sigma(\Pi) \equiv \frac{1}{2}[(\lambda_{high} + \sigma_{high}) - (\lambda_{low} - \sigma_{low})].$$

For example, considering $\Pi = 55\%$, the least thundery LI_{sfc} value for which the TCP reaches 55% is $\lambda_{high} = -1.54$ °C ($\sigma_{high} = 0.23$ °C), and the least thundery LI_{sfc} value for which the TCP stays $\geq 55\%$ is $\lambda_{low} = -2.46$ °C ($\sigma_{low} = 0.38$ °C). This can also be verified with Fig. 5. Therefore, the benchmark value for $\Pi = 55\%$ is estimated at $LI_{sfc} = -2.08 \pm 0.77$ °C.

Table 4
Estimated benchmark values and their 'uncertainties' for the "best" indices as a function of thundery case probability, shown in the upper row of the table. Index and parameter units are the same as in, e.g., Table 3

Index	50%	55%	60%
ADED1	0.04 ± 0.38		
ADED2	0.84 ± 0.12	1.22 ± 0.45	1.94 ± 0.49
$CAPE_{100}$	109 ± 26	174 ± 58	
$CAPE_{50}$	130 ± 21	209 ± 64	360 ± 170
$CAPE_{MU}$	670 ± 160	830 ± 150	
$CAPE_{sfc}$	670 ± 130	770 ± 120	980 ± 180
DCI	24.3 ± 2.3		
KI_{MOD}	40.2 ± 1.1		
LI_{100}	0.61 ± 0.42	0.09 ± 0.21	
LI_{50}	0.39 ± 0.37	0.02 ± 0.18	
LI_{MU}	− 1.30 ± 0.17	− 2.09 ± 0.84	
LI_{sfc}	− 1.36 ± 0.20	− 2.08 ± 0.77	− 3.30 ± 0.89
SHOW	− 0.06 ± 0.68		
$SWISS_{00}$	1.6 ± 1.1		
$SWISS_{12}$	− 3.89 ± 0.32	− 4.33 ± 0.54	− 4.73 ± 0.53
THOM	31.5 ± 1.3	32.1 ± 1.7	
YON	3.44 ± 0.48		
YONMOD	5.61 ± 0.61		

Table 5
The highest thundery case probability Π_{max} (in multiples of 5%) that was reached for all indices.

Π_{max} (%)	Index/parameter
60	ADED2, $CAPE_{50}$, $CAPE_{sfc}$, LI_{sfc}, $SWISS_{12}$
55	$CAPE_{100}$, $CAPE_{MU}$, LI_{100}, LI_{50}, LI_{MU}, THOM
50	ADED1, DCI, KI_{MOD}, SHOW, $SWISS_{00}$, YON, YONMOD
45	BRAD, JEFF, KO, PII, TEI_{925}, TEI_{850}
40	BOYD, KI, RACK, SWEAT
35	CT, SI, TT, VT

This procedure may seem cumbersome. Why not simply take the highest and lowest index values with TCP = Π instead? This method has been tried as well, but resulted in too thundery index values for the various Π 's. By using our more cumbersome criteria, this effect was reduced. For instance, the $\Pi = 5\%$ threshold would have been set at $LI_{sfc} = 6.70 \pm 1.49$ °C, instead of $LI_{sfc} = 7.71 \pm 0.48$ °C, while the latter seems to be better (Fig. 5).

We can construct a table, in which all index benchmark values are given as a function of Π. The result is shown in Table 4. Only the 18 indices for which the TCP reaches 50% are included in the table. Five of these indices (ADED2, $CAPE_{50}$, $CAPE_{sfc}$, LI_{sfc}, and $SWISS_{12}$) reach a thundery case probability of 60%. It is interesting to note that all these five indices represent latent instability at, or very near, the surface. Although $CAPE_i$ did not excel as a dichotomous predictor, it is capable of reaching high thundery case probabilities. The Boyden Index does not consider moisture at all, and this weakness is reflected in a low Π_{max} (only 40%) (see Table 5). It is clear that if we were to estimate thunderstorm probability from the three different instability types only (Appendix A), we should first consider low-level latent instability—especially near the surface—then potential instability and finally conditional instability. Nonetheless, we stress that these three instability types are certainly not independent.

6. Discussion

In this section, we discuss the disadvantages of the method used and the reasons why we should interpret the results carefully.

6.1. Diurnal variation

In our study, we have made quite a few generalizations, as all available rawinsonde observations were included and no distinction was made between different types of thunderstorms. An important generalization is that we have not distinguished between 00, 06, 12, or 18 UTC soundings. Solar heating will increase latent instability near the surface and in the atmospheric boundary layer. Since the predictand is based on the 6-h period after the rawinsonde observations, indices that include levels near the surface will reach higher thundery case probabilities than indices that use other levels. For example, LI_i (and $CAPE_i$) will frequently exhibit a relatively large diurnal variation, with more thundery

values in the 12 and 18 UTC soundings. Nighttime thunderstorms may be associated with other thermodynamic configurations than daytime thunderstorms. Huntrieser et al. (1997) already suggested different optimal thresholds for 00 and 12 UTC rawinsonde observations and concluded that indices have different relative forecast skill for 00 and 12 UTC soundings. Both SWISS indices have therefore been designed for 00 and 12 UTC soundings separately.

6.2. Seasonal variation

Seasonal variations were also not considered in the assessment of forecast skill. However, some of the indices were designed for a particular time of year, e.g., the S Index (SI) was meant for use from April to September. The KI scatterplot (not shown) seems to reveal that the optimal lower threshold for forecasting a thundery case with the K Index is lower in winter than in summer.

6.3. Different thunderstorm types

Furthermore, not each index and parameter was intended as an aid in the forecasting of any type of thunderstorm. Some were designed to forecast severe thunderstorms (e.g., DCI and SWEAT), while others were only meant for forecasting air-mass thunderstorms (e.g., KI and YON). The Jefferson and K index increase with increasing relative humidity at 700 hPa, which is thought to be favorable for the occurrence of air-mass thunderstorms in summer. However, relatively humid air at mid levels is unfavorable for thunderstorms that are associated with potential instability. The Boyden Index was initially designed to assess thunderstorm risk at frontal passages over the UK. In view of all this, some indices will have scored relatively poorly in our study, while they may have been very useful in forecasting the type of weather that they were designed to forecast.

6.4. Reliability of the data

Due to ice deposition and thermal radiation rawinsonde observations are not always reliable. However, owing to the huge dataset that was available to us, we believe that bad rawinsonde observations have been smoothed out more than sufficiently.

It is easier to determine whether or not the ATD system detected an observed thunderstorm than it is to determine for which cases the ATD system made a false detection of thunderstorm activity. For a few very unlikely thundery cases we have quickly verified the thundery case by viewing synoptic weather maps and satellite pictures in that 6-h interval. We had to conclude that the ATD system will occasionally have detected a sferic where thunderstorm activity was out of the question. For example, the highest LI_{sfc} that yielded a thundery case according to the ATD system was 22.6 °C at 00 UTC, January 26, 2000 (see Fig. 2). The ATD system fixed a sferic at about 38 km distance (51.77°N 5.30°E) from De Bilt at 0345 UTC that night. Another fix was made 11 min earlier, at 49.87°N 5.60°E, about 250 km from De Bilt, over the southeastern part of Belgium. However, this night was quiet and frosty over the Benelux region, with a high-pressure area of more than 1035 hPa situated close by and the presence of a distinct inversion at

about 2 km above ground level (AGL) at De Bilt ($\partial T/\partial z = +5.2$ °C km^{-1} between 1795 and 2687 m AGL). The sounding seems to be reliable, after having compared it with the rawinsonde observations 6 h before and after. Hence, these must have been false detections. It is impossible for us to determine exactly how many false detections we have mistaken for actual thundery cases. Considering Fig. 2, we see that there are not too many thundery cases that are extremely unlikely. Note that this apparent unlikelihood might as well be due to erroneous rawinsonde data.

6.5. Lifting mechanism

Severe thunderstorms are often associated with forced lifting along a line of moisture flux convergence (e.g., Van Delden, 1998). Unfortunately, it is very difficult to obtain information on possible sources of lift by one single rawinsonde observation. Some authors assume that warm air advection implies upward motion (see the definition of SWEAT in Appendix A). According to many other authors, canceling effects between warm air advection and differential vorticity advection are too large to make this imprudent assumption. Hence, lifting cannot be taken into account when estimating the thundery case probability based on a single sounding.

6.6. Representativeness of a sounding

Our predictand may seem quite strict, as it only covers the area within 100 km from De Bilt during the 6 h following a rawinsonde observation. However, a rawinsonde observation may still not be representative for this area and for this period in some situations. Especially in winter, when thunderstorms may develop along rapidly moving cold fronts, the synoptic situation may change a great deal in only 6 h time. Furthermore, since squall lines and weather fronts occur on a scale of 10 to 100 km (Holton, 1992), there may be another air mass present in the area that the sounding does not represent.

6.7. Advection of thunderstorms

Once a storm has formed, it can sometimes reach an area where conditions are less favorable for the formation of thundery showers. In other words, the sounding at De Bilt need not be representative for the pre-convective environment of a thunderstorm that is advected from a region far away. This will, of course, have a negative effect on the forecasting skill of an index.

7. Conclusions and recommendations

We have discussed how pressure, temperature, moisture and wind data from a single rawinsonde observation at De Bilt can be used as an aid in estimating the probability of thunderstorm occurrence within 100 km from De Bilt during the 6 h following the sounding. We have been able to create a comprehensive and reliable climatology for 32 different thermodynamic and kinematic parameters, covering a total period of more than

seven full years, or more than 10,000 six-hour time intervals. The correlations between these parameters and the occurrence of thundery cases have been evaluated.

7.1. Preliminary findings

First we examined the correlation between thunderstorm occurrence according to synoptic observations and that according to the ATD system. Of six different lightning detection radii (10, 20, 50, 100, 150 and 200 km), a radius of 20 km around De Bilt rendered the highest agreement with the synoptic observations, according to Heidke and the CSI, confirming the 8–20 km distance up to which thunder is normally heard. We found that in 95.5% of all 6-h intervals in which thunder was heard, at least one sferic was fixed at less than 100 km distance from De Bilt. We can therefore be quite certain that an observed thunderstorm at De Bilt will also have been registered as a thundery case. False detections by the ATD system were relatively rare and will not have affected the results noticeably.

7.2. Main results

For use of the indices in a dichotomous forecasting scheme, the optimal thresholds between thundery vs. non-thundery forecasts have been estimated. Two skill scores were considered that are commonly used in literature: the True Skill Statistic (TSS) and the Heidke Skill Score (Heidke).

Combining the properties of the TSS and Heidke to form the Normalized Skill Score, an optimal threshold has been derived for all indices. We conclude from the results in Table 2 that in a dichotomous thundery case forecasting scheme, $\mathrm{BOYD} \geq 94.6$ and $\mathrm{LI_{sfc}} \leq 1.6\ ^{\circ}\mathrm{C}$ can equally well be used to predict a thundery case ($\mu = 80\%$). However, $\mathrm{LI_{100}} \leq 3.0\ ^{\circ}\mathrm{C}$ performs best ($\mu = 95\%$)! Relatively poor dichotomous thunderstorm predictors are the DCI, SWEAT and KI. It should, however, be noted that these three indices were never meant as a predictor for all types of thunderstorms in the first place.

The rightmost column in Table 3 showed the Rank Sum Scores for all 32 thunderstorm predictors. Although this comparison of forecast skill used quite a different method, $\mathrm{LI_{100}}$ was again found with the best performance, while the other indices and parameters do not seem to be rated much differently from what we saw in Table 2.

In Sections 5.4 and 5.5, we estimated thundery case probability as a function of the various thunderstorm indices and parameters. Probability plots (not shown) exhibited a monotonic behavior for most indices. Latent instability at or very near the surface yields the highest thundery case probabilities, whereas a much less definite answer is obtained by considering conditional instability.

7.3. Recommendations

In this study, we have assessed thunderstorm possibility and probability in general, based on predictors that can be derived from a single rawinsonde observation. This very general approach of thunderstorm forecasting can be improved. For this purpose, a few recommendations are made here.

(1) Discriminate between summer and winter soundings. The thickness of an atmospheric layer increases with increasing temperature. Since many indices use different pressure levels, stability is not measured equally in all seasons. For instance, the difference between the ambient lapse rate and the dry-adiabatic lapse rate is a measure of conditional instability. If conditional instability is measured by evaluating the temperature difference between two pressure levels, it will be overestimated in winter and underestimated in summer. For example, VT attains more thundery values in winter than in summer.
(2) Evaluate forecast skill for different thunderstorm types, such as air mass and frontal thunderstorms. Different thunderstorm predictors will yield better results for different thunderstorm types. For example, indices that specifically describe potential instability, such as BRAD, KO, and PII, will probably perform better when a synoptic lifting mechanism through a deep layer is present.
(3) Use data on lightning-frequency to distinguish between 'ordinary' and severe thunderstorms. There is a huge difference between a single wintry shower that produces one or two lightning strikes and an outbreak of severe thunderstorms in summer. Some indices, such as the DCI, may be more able than others to distinguish between 'ordinary' and severe thunderstorms.
(4) Include model derived parameters that cannot be derived from a single rawinsonde observation, such as vertical velocities at different levels and (low-level) moisture flux convergence to account for continuous moisture supply and a source of lift. In our opinion, this will lead to a significant improvement of forecast skill.

Acknowledgements

We would like to thank the Royal Dutch Meteorological Institute (Sander Tijm) and the UK Meteorological Office (Paul Hardaker, Nicola Latham) for providing us with the huge amount of data on which this study was based.

Appendix A. Definitions of indices and parameters

The indices and parameters which are investigated in this paper represent conditional instability and/or latent instability and/or potential instability. Let us consider an 'in situ' vertical temperature profile as it was measured during a rawinsonde ascent. Suppose the measured lapse rate in a certain vertical atmospheric layer was $\gamma = -\partial T/\partial z$. If the dry adiabatic lapse rate exceeds γ? but the pseudo-adiabatic lapse rate does not exceed γ, the layer is said to be *conditionally unstable*, since the instability is conditional to saturation of the air. Clearly, high values of γ will favor the development of strong vertical (convective) motions. If the air inside a conditionally unstable atmospheric layer is far from saturated, which is usually the case, conditional instability can still be released by a rising parcel from below the layer. Since the environmental lapse rate γ inside the layer exceeds the lapse rate pseudo-adiabatic lapse rate of a rising, saturated parcel from below, the parcel may be become positively buoyant as it rises through the layer. The possibility of pseudo-

adiabatically lifted air parcels to become unstable is called *latent instability* (e.g., Galway, 1956). The level at which this happens is called the Level of Free Convection (LFC) of the lifted parcel. Once a parcel has reached its LFC inside a conditionally unstable layer, it will accelerate upward owing to a positive buoyancy force. Suppose that an entire atmospheric column is lifted pseudoadiabatically, until it is completely saturated. If the lower part of the column reaches its lifting condensation level first, subsequent lifting will cause a rapid destabilisation, since the column cools according to pseudo-adiabatic lapse rate in the saturated lower part and according to dry adiabatic lapse rate aloft. It can be shown (Emanuel, 1994) that lifting the entire column to saturation would only result in a lapse rate γ which is greater than the pseudoadiabatic lapse rate if θ_e decreases with height. Hence, lifting a column of air in which $\partial\theta_e/\partial z<0$ to saturation throughout, yields (at least) conditional instability. Remember that conditional instability is conditional to saturation of the air and since the entire column has been lifted to saturation, the condition is already satisfied. This is called *potential instability*. Some authors use the wet-bulb potential temperature θ_w to define potential instability. If an ascending air parcel can become warmer than its environment at its LFC, its θ_e will always exceed the ambient θ_e at that level, regardless of the ambient relative humidity there. Therefore, latent instability always implies potential instability. However, this statement cannot be reversed! In the absence of latent instability, an atmospheric layer can still be potentially unstable due to a rapid decrease in relative humidity with height.

In the following, a list of definitions is given of the 32 indices and their relation to the three types of instability is discussed.

Adedokun1 Index, $\mathrm{ADED1} \equiv \theta_{w-850} - \theta_{s-500}$, where θ_w is the wet-bulb potential temperature, θ_s is the saturated wet-bulb potential temperature and where the subscript refers to the (pressure) level. ADED1 (Adedokun, 1981, 1982) provides a measure for the buoyancy of an air sample that was lifted pseudoadiabatically from 850 to 500 hPa. Therefore, it is an indicator of latent instability.

Adedokun2 Index, $\mathrm{ADED2} \equiv \theta_{w-sfc} - \theta_{s-500}$. ADED2 (Adedokun, 1981, 1982) is a measure for the buoyancy of an air sample that was lifted pseudoadiabatically from the surface (2 m) to 500 hPa. It therefore considers the latent instability of an air parcel at the surface.

Boyden Index (BOYD), $\mathrm{BOYD} \equiv 0.1(Z_{700} - Z_{1000}) - T_{700} - 200$, where Z is the geopotential height of a particular (pressure) level. Unlike most instability indices, the Boyden Index (Boyden, 1963) does not take moisture into account. It merely describes the vertical temperature profile between 1000 and 700 hPa and was originally designed to assess thunderstorm risk at frontal passages over the UK.

Bradbury Index, $\mathrm{BRAD} \equiv \theta_{w-500} - \theta_{w-850}$. The Bradbury Index (Bradbury, 1977) is also referred to as the Potential Wet-Bulb Index (Peppler and Lamb, 1989). It assesses the potential instability between 850 and 500 hPa.

Convective Available Potential Energy, $\mathrm{CAPE_i} \equiv R_d \int (T'_v - T_v)\mathrm{d}(\ln\ p)$, where only increments with $T'_v > T_v$ are integrated. Here, p_i denotes the initial pressure level from which an air parcel is lifted pseudoadiabatically, p_{EL} is the pressure at the parcel's final equilibrium level, T'_v is the virtual temperature of the lifted parcel, and T_v is the virtual temperature of the environment. There exist many different definitions of CAPE in literature. We have defined it as the total amount of work done by the upward buoyancy

force that is exerted on an air parcel as it is lifted from its initial level to its final Equilibrium Level (EL). Negatively buoyant areas below the EL are not taken into account. We consider pseudo-adiabatic parcel ascent, in which all condensed water is instantly removed from the parcel. Furthermore, it is assumed that the environment does not contain any liquid water (no clouds). $CAPE_{sfc}$ is the CAPE of a surface parcel, or rather, a parcel with the temperature, dew-point temperature and pressure at 2 m height above the surface. $CAPE_{50}$ is the CAPE of a parcel with the ln p weighted average temperature, dew-point temperature and pressure of the layer 50 hPa above the surface. $CAPE_{100}$ is the CAPE of a parcel with the ln p weighted average temperature, dew-point temperature and pressure of the layer 100 hPa above the surface. $CAPE_{MU}$ is defined as the CAPE of a parcel with the temperature, dew-point temperature and pressure at the level where θ_e reaches its highest value in the layer 250 hPa above the surface ($CAPE_i$ was only calculated if dew-point temperature data were present up to 200 hPa and temperature data were available up to 170 hPa.).

Cross Totals Index, $CT \equiv T_{d-850} - T_{500}$, where T_d is the dewpoint temperature. Since the mixing ratio can be expressed in terms of dew-point temperature at a certain pressure level, the Cross Totals Index (Miller, 1967) increases with a combination of moisture at low levels (850 hPa) and relatively cold air at upper levels (500 hPa). It is usually combined with the Vertical Totals Index (VT) to yield the Total Totals Index (TT).

Deep Convective Index, $DCI \equiv T_{850} - T_{d-850} - LI_{sfc}$, where LI_{sfc} is defined later in this section. The DCI (Barlow, 1993) attempts to combine the properties of equivalent potential temperature (θ_e) at 850 hPa with latent instability at the surface (2 m). It should indicate the potential for strong thunderstorms.

Jefferson Index, $JEFF \equiv 1.6\theta_{w-850} - T_{500} - 0.5(T_{700} - T_{d-700}) - 8$. The Jefferson Index (Jefferson, 1963a,b, 1966) as it is used nowadays is a modified version of the original, which was developed by Jefferson in 1963 to make the Rackliff Index (RACK) less temperature-dependent. Jefferson used the temperature at 850 hPa instead of 900 hPa and included the 700 hPa dew-point depression, which increases with decreasing relative humidity, to account for the fact that rising air parcels might lose some of their buoyancy due to entrainment of drier air. On the other hand, the fact that dry air at 700 hPa may indicate decreasing θ_e with height, which implies potential instability, is neglected.

K Index, $KI \equiv (T_{850} - T_{500}) + T_{d-850} - (T_{700} - T_{d-700})$. George (1960) developed the K Index for forecasting air mass thunderstorms. This index increases with decreasing static stability between 850 and 500 hPa, increasing moisture at 850 hPa, and increasing relative humidity at 700 hPa.

Modified K Index, $KI_{MOD} \equiv (\bar{T} - T_{500}) + \bar{T}_d - (T_{700} - T_{d-700})$. The Modified K Index (Charba, 1977) should be an improvement on KI since now the mean temperature, $\bar{T}$, and mean dew-point temperature, $\bar{T}_d$, between the surface and 850 hPa are used ($\bar{T}$ and $\bar{T}_d$ are ln p averages).

KO Index, $KO \equiv 0.5(\theta_{e-500} + \theta_{e-700}) - 0.5(\theta_{e-850} + \theta_{e-1000})$. The KO Index (Andersson et al., 1989) was designed by the German Weather Service to account for low and mid-level potential instability by comparing θ_e values at low (1000 to 850 hPa) and mid (700 to 500 hPa) levels. If surface pressure is lower than 1000 hPa, θ_{e-1000} is replaced by θ_{e-sfc} in our calculations.

Lifted Index, $LI_i \equiv T_{500} - T'_{i \rightarrow 500\ hPa}$. The Lifted Index LI is defined as the difference between the observed temperature at 500 hPa and the temperature($T_i{}'_{\rightarrow 500\ hPa}$) of a parcel after it has been lifted pseudo-adiabatically to 500 hPa from its original level. Therefore, it focuses on the latent instability of an air sample. The Lifted Index can be calculated for any sample of air at pressure p_i>500 hPa if the ambient temperature at 500 hPa is known. It should be noted that the Lifted Index depends on the properties of the particular air parcel that was used. Originally, Galway (1956) developed the Lifted Index for the prediction of latent instability during afternoon hours by using the forecast maximum temperature. Since our aim has been to assess thunderstorm risk by means of actual soundings, we have used LI as an observed static index instead of a forecast index, following other studies on this subject (e.g., Peppler and Lamb, 1989; Huntrieser et al., 1997). LI_{sfc} is the LI for a surface parcel, or rather, a parcel with the temperature, dew-point temperature and pressure at 2 m height above the surface. LI_{50} is the LI for a parcel with the ln p weighted average temperature, dew-point temperature and pressure of the layer 50 hPa above the surface. LI_{100} is the LI for a parcel with the ln p weighted average temperature, dew-point temperature and pressure of the layer 100 hPa above the surface. LI_{MU} is defined as the LI for a parcel with the temperature, dew-point temperature and pressure at the level where θ_e reaches its highest value in the layer 250 hPa above the surface.

Potential Instability Index, $PII \equiv (\theta_{e-925} - \theta_{e-500})/(Z_{500} - Z_{925})$. The PII (Van Delden, 2001) is a measure the potential instability of the atmospheric layer between 925 and 500 hPa.

Rackliff Index, $RACK \equiv \theta_{w-900} - T_{500}$. By comparing the wet-bulb potential temperature at 900 hPa with the dry-bulb temperature at 500 hPa, RACK assesses the latent instability of an air parcel at 900 hPa. Note that 900 hPa is not a standard pressure level. Jefferson (1963a,b, 1966) later developed a less temperature-dependent version of this index (see the definition of JEFF).

Showalter Index, $SHOW \equiv T_{500} - T'_{850\ hPa \rightarrow 500\ hPa}$. The development of the Lifted Index was inspired by the Showalter Index (Showalter, 1953). This index is defined as the difference between the observed temperature at 500 hPa (T_{500}) and the temperature ($T_{850\ hPa \rightarrow 500\ hPa}$) of an air parcel after it has been lifted pseudoadiabatically to 500 hPa from 850 hPa. Showalter himself called it the Stability Index and according to his experience, values of +3 °C or less were "quite likely to produce thunderstorms". If low-level moisture does not reach as high as the 850 hPa level, SHOW will underestimate thunderstorm probability. In literature, the abbreviation 'SSI' is commonly used, but we have used 'SHOW' to avoid confusion with the S Index (SI).

S Index, $SI \equiv TT - (T_{700} - T_{d-700}) - A$, where A is defined as follows. If VT>25, then $A=0$; if $VT \geq 22$ and ≤ 25, then $A=2$; if $VT < 22$, then $A=6$. TT and VT are defined later in this section. The S Index was developed by the German Military Geophysical Office as an improvement on the Total Totals Index (TT), including the dew-point depression at 700 hPa and a variable parameter 'A' that is based on the Vertical Totals Index (VT). The S Index is considered useful from April to September (Reymann et al., 1998). It takes the same variables into account as the K Index, but in other proportions.

Severe Weather Threat Index, $\mathrm{SWEAT} \equiv 12T_{d-850} + 20(TT-49) + 2f_{850} + f_{500} + 125(\sin[d_{500}-d_{850}]) + 0.2$, where f_{850} is the wind speed at 850 hPa in knots, f_{500} is the wind speed at 500 hPa in knots, d_{850} is the wind direction at 850 hPa (0–360°), d_{500} is the wind direction at 500 hPa (0–360°). Only positive terms are added and the last term is set to zero, unless ALL of the following conditions are met: $210° \le d_{500} \le 310°$, $130° \le d_{850} \le 250°$, $d_{500} > d_{850}$ and $f_{500} \ge 15$ knots AND $f_{850} \ge 15$ knots. Some authors do not consider the above conditions, enabling SWEAT to take on negative values. The SWEAT index (Miller, 1972) was designed to assess severe weather potential, such as severe storms and tornadoes, rather than ordinary thunderstorms. Especially damaging convective weather seems to be targeted, since the Total Totals Index (TT) and the 850 hPa dew-point temperature are included. SWEAT is one of the few indices that attempt to include dynamical effects. Veering of the wind with height between 850 and 500 hPa suggests warm air advection in that layer, which could imply upward motion through the "omega equation" (Eq. (6.29) in Holton, 1992). In practice, however, we should be very careful in associating warm air advection with upward motion, since there is often a significant amount of cancellation between the terms in the omega equation. For instance, in a case study by Van Delden (1998), warm air advection is associated with downward motion.

$SWISS_{00}$ Index, $\mathrm{SWISS}_{00} \equiv \mathrm{SHOW} + 0.4\mathrm{WSh}_{3-6} + 0.1(T_{600} - T_{d-600})$. Huntrieser et al. (1997) introduced two non-dimensional thunderstorm indices as a guidance for forecasting thunderstorms in the northern part of Switzerland. They were called SWISS (stability and wind shear index for thunderstorms in Switzerland) indices and include stability, wind shear and relative humidity. SWISS_{00} was designed to be applied to 00 UTC rawinsonde observations, which is about 1 h after local midnight in Switzerland. The wind shear term (WSh_{3-6}) was calculated by interpolating the rawinsonde wind measurements at 6 km and 3 km AGL (above ground level). WSh_{3-6} is the length (or magnitude/norm) of the wind vector [m/s] that should be added to the wind vector at 3 km AGL to obtain the wind vector at 6 km AGL. According to SWISS_{00}, low SHOW values, low wind shear between 3 and 6 km AGL and relatively humid air at 600 hPa should favor the possibility of thunderstorm occurrence.

$SWISS_{12}$ Index, $\mathrm{SWISS}_{12} \equiv \mathrm{LI}_{sfc} - 0.1\mathrm{WSh}_{0-3} + 0.1(T_{650} - T_{d-650})$. The SWISS_{12} index (Huntrieser et al., 1997) was designed to be applied to 12 UTC rawinsonde observations in Switserland. The wind shear term (WSh_{0-3}) was calculated by interpolating the rawinsonde wind measurements at 3 km AGL. WSh_{0-3} is the length of the wind vector [m/s] that should be added to the wind vector at 10 m AGL to obtain the wind vector at 3 km AGL. According to SWISS_{12}, low LI_{sfc} values, increased wind shear between 10 m and 3 km AGL and relatively humid air at 650 hPa should favor the possibility of thunderstorm occurrence.

Thompson Index, $\mathrm{THOM} \equiv \mathrm{KI} - \mathrm{LI}_{50}$. The Thompson Index should be an improvement of KI, since KI neglects latent instability below 850 hPa.

Total Energy Index (925 hPa), $\mathrm{TEI}_{925} \equiv h_{500} - h_{925}$, where h is the static energy in J kg^{-1}, defined as $h \equiv (c_{pd} + rc_l)T + L_v r + (1+r)gZ$, where c_{pd} is the heat capacity of dry air at constant pressure, c_l is the heat capacity of liquid water, r is the mixing ratio of water vapor, L_v is the latent heat of vaporization and Z is the geopotential height. Since the static energy is conserved in adiabatic displacements, TEI_{925} gives a

measure of the potential instability of the atmospheric layer between 925 and 500 hPa (Darkow, 1968).

Total Energy Index, $\mathrm{TEI}_{850} \equiv h_{500} - h_{850}$. Properties are the same as for TEI_{925}, but now applied to the 850–500 hPa layer.

Total Totals Index, $\mathrm{TT} \equiv \mathrm{VT} + \mathrm{CT}$. TT (Miller, 1967) is a commonly used convective index in many parts of the world, but was originally designed for application in the U.S. (Peppler and Lamb, 1989). It fails to consider latent instability below 850 hPa.

Vertical Totals, $\mathrm{VT} \equiv T_{850} - T_{500}$. Like the Boyden Index, the Vertical Totals Index (Miller, 1967) does not consider moisture and only assesses conditional instability between 850 and 500 hPa. Since the 850–500 hPa layer thickness increases with increasing temperature, the actual lapse rate will be underestimated in summer and overestimated in winter.

Yonetani Index, $\mathrm{YON} \equiv 0.966\Gamma_{\mathrm{L}} + 2.41(\Gamma_{\mathrm{U}} - \Gamma_{\mathrm{W}}) + 9.66\mathrm{RH} - 15$ (if RH>0.57); $\mathrm{YON} \equiv 0.966\Gamma_{\mathrm{L}} + 2.41(\Gamma_{\mathrm{U}} - \Gamma_{\mathrm{W}}) + 9.66\mathrm{RH} - 16.5$ (if $\mathrm{RH} \leq 0.57$), where Γ_{L} is the lapse rate between 900 and 850 hPa, Γ_{U} is the lapse rate between 850 and 500 hPa, Γ_{W} is the pseudo-adiabatic lapse rate at 850 hPa, RH is the ln p-weighted average of the relative humidity between 900 and 850 hPa, where RH-values are between 0 and 1. With this index Yonetani (1979) attempted to combine conditional instability with low-level moisture as he designed the Yonetani Index for air mass thunderstorms in the northern Kanto Plain, Japan.

Modified Yonetani Index, $\mathrm{YON}_{\mathrm{MOD}} \equiv 0.964\Gamma_{\mathrm{L}} + 2.46(\Gamma_{\mathrm{U}} - \Gamma_{\mathrm{W}}) + 9.64\mathrm{RH} - 13$ (if RH>0.5), $\mathrm{YON}_{\mathrm{MOD}} \equiv 0.964\Gamma_{\mathrm{L}} + 2.46(\Gamma_{\mathrm{U}} - \Gamma_{\mathrm{W}}) + 9.64\mathrm{RH} - 14.5$ (if $\mathrm{RH} \leq 0.5$), where the notation is the same as for YON. This slightly modified version of the Yonetani Index was tailored specifically to the Greater Cyprus environment (Jacovides and Yonetani, 1990).

Summarizing, BOYD and VT account for pure conditional instability of a certain atmospheric layer. In contrast with the other predictors, these two indices neglect moisture. ADED1, ADED2, $\mathrm{CAPE}_{\mathrm{i}}$, LI_{i}, RACK, and SHOW account for pure latent instability at a certain level. BRAD, KO, and PII account for pure potential instability of a certain atmospheric layer. The other indices describe a combination of one or more of these three instability types with moisture. Only SWEAT, SWISS_{00}, and SWISS_{12} contain kinematic information.

Appendix B. The 2×2 contingency table and verification parameters

In literature, dichotomous (event/non-event) forecasts are often verified by using contingency tables. We follow the discussion of Doswell et al. (1990), in order to describe the use and usefulness of these tables. Consider an example of a 2×2 contingency table (Table 6). The letters h, s, f and q denote the number of correct event forecasts (hits), surprise events, false alarms and correct non-event forecasts ('quiescent' cases), respectively. In the rightmost column and lowest row of the table, the number of event forecasts (EF), non-event forecasts (NEF), events (E), and non-events (NE) are given, as well as the sample size (SS).

Table 6
Schematic 2 × 2 contingency table.

		Event observed?		
		Yes	No	
Event forecast?	Yes	h	f	$EF = h + f$
	No	s	q	$NEF = s + q$
		$E = h + s$	$NE = f + q$	$SS = h + s + f + q$

Different skill scores can be derived from the four independent entries in the table. Doswell et al. (1990) discuss eight ways in which ratios can be formed, involving each of these four entries with their associated marginal sums. These are the first eight ratios listed in Table 7. The Probability of Detection (POD) and the False Alarm Ratio (FAR) are the most commonly used ratios in literature. The POD gives the percentage of all events that are forecast. However, it is possible to create an artificially high POD by forecasting the event (too) often. For instance, if the meteorologist always forecasts the event ($E = h$), the POD will be 100%. To put things in the right perspective, the False Alarm Ratio (FAR) can then be used. The other six ratios are given here for completeness and reference. Somewhat more complicated skill scores are often more useful verification parameters.

The Critical Success Index (CSI) (Donaldson et al., 1975) is the ratio of hits to the total number of events and false alarms. It does not account for situations where a non-event was forecast correctly. Therefore, the CSI is often regarded as an index that only considers those situations where a forecasting problem existed. However, this is not entirely true, since in some cases, it can take much effort to finally conclude correctly that the event will not occur. The CSI is not increased due to these cases, which seems unfair. Furthermore, it is a biased score, because it inflates warning skill with increasing event frequency. Schaefer (1990) refers to the CSI as "a valid indicator of the relative worth of different forecast techniques", when applied to the same environment with the same event frequency.

Table 7
Skill scores and their interpretations.

Skill score		Percentage of	Definition
POD	Probability of detection	Expected events	h/E
FOM	Frequency of misses	Surprise events	s/E
FAR	False alarm ratio	False event forecasts	f/EF
FOH	Frequency of hits	Correct event forecasts	h/EF
PON	Probability of a null-event	Expected non-events	q/NE
POFD	Probability of false detection	Unexpected non-events	f/NE
DFR	Detection failure ratio	False non-event forecasts	s/NEF
FOCN	Frequency of correct null-forecasts	Correct non-event forecasts	q/NEF
CSI	Critical success index	Events and event forecasts that were hits	$h/(SS - q)$
TSS	True skill statistic	Expected events minus unexpected non-events	$h/E - f/NE$
Heidke	Heidke skill score	Correct forecasts not due to chance	See text for definition

The True Skill Statistic, TSS, does take the quiescent cases into account. Its definition is quite simple, namely the probability that an event is indeed forecast (POD) minus the probability that a non-event occurs unexpectedly (POFD). Murphy and Daan (1985) recommend it as a "proper formulation of a skill score", but this is contradicted by Doswell et al. (1990) where rare events (< 1%), such as severe local storms, are forecast. Doswell et al. (1990) show that the TSS approaches the POD if the number of quiescent cases dominates. Huntrieser et al. (1997) used the TSS to determine the optimal guidance values for their various thunderstorm predictors. These guidance values were taken at the point where the TSS reached a maximum. A perfect dichotomous forecast would have neither false alarms nor surprises, i.e., $f = 0$ and $s = 0$, implying TSS = 1. For totally random forecasts, TSS = 0, since there would be as many hits as surprises and as many false alarms as quiescent cases.

Finally, the Heidke Skill Score (Brier and Allen, 1951), in literature also referred to as S, gives credit for all correct forecasts that were not merely due to chance. This yields a computation method with Heidke=(CF − CFC)/(SS − CFC), where the number of correct forecasts is denoted by CF and the expected number of correct forecasts due to chance by CFC. Obviously, $CF = h + q$. The expected number of correct forecasts due to chance (CFC) is the number of event forecasts times the event frequency plus the number of non-event forecasts times the non-event frequency. In symbolic notation: CFC = EF(E/SS) + NEF(NE/SS). With these definitions made, writing out the results and rearranging terms, the Heidke Skill Score can be written as

$$\text{Heidke} = \frac{(h+s)(q-f)+(h-s)(q+f)}{(h+s)(s+q)+(h+f)(q+f)} = \frac{(\text{E})(q-f)+(\text{NE})(h-s)}{(\text{E})(\text{NEF})+(\text{NE})(\text{EF})}.$$

We see that all correct forecasts (h and q) are rewarded, whereas all incorrect forecasts (s and f) are penalized, but everything in a controlled way. A perfect forecast, $s = 0$ and $f = 0$, would imply Heidke = 1. For a totally random forecast, $q = f$ and $h = s$, and Heidke = 0. Von Storch and Zwiers (1999) describe the Heidke Skill Score as "a useful measure of the skill of a two-class categorical forecasting scheme".

References

Adedokun, J.A., 1981. Potential instability and precipitation occurrence within an inter-tropical discontinuity environment. Arch. Meteor. Geophys. Bioklimatol., A 30, 69–86.

Adedokun, J.A., 1982. On an instability index relevant to precipitation forecasting in West Africa. Arch. Meteor. Geophys. Bioklimatol., A 31, 221–230.

Andersson, T., Andersson, M., Jacobsson, C., Nilsson, S., 1989. Thermodynamic indices for forecasting thunderstorms in southern Sweden. Meteorol. Mag. 116, 141–146.

Barlow, W.R., 1993. A new index for the prediction of deep convection. Preprints, 17th Conf. on Severe Local Storms. Amer. Meteor., St. Louis, MO, pp. 129–132.

Boyden, C.J., 1963. A simple instability index for use as a synoptic parameter. Meteorol. Mag. 92, 198–210.

Bradbury, T.A.M., 1977. The use of wet-bulb potential temperature charts. Meteorol. Mag. 106, 233–251.

Brier, G.W., Allen, R.A., 1951. Verification of weather forecasts. Compendium of Meteorology. Amer. Meteor. Soc., Boston, pp. 841–848.

Charba, J.P., 1977. Operational system for predicting thunderstorms two to six hours in advance. NOAA

Technical Memo. NWS TDL-64. 24 pp. [Techniques Development Laboratory, National Weather Service, Silver Spring, MD.].
Dai, A., 2001. Global precipitation and thunderstorm frequencies: Part I. Seasonal and interannual variations. J. Climate 14, 1092–1111.
Darkow, G.L., 1968. The total energy environment of severe storms. J. Appl. Meteor. 7, 199–205.
Donaldson, R.J., Dyer, R.M., Krauss, M.J., 1975. An objective evaluator of techniques for predicting severe weather events. Preprints, 9th Conf. on Severe Local Storms. Amer. Meteor., Norman, OK, pp. 321–326.
Doswell III, C.A., Davies-Jones, R., Keller, D.L., 1990. On summary measures of skill in rare event forecasting based on contingency tables. Weather Forecast. 5, 576–585.
Emanuel, K.A., 1994. Atmospheric Convection. Oxford Univ. Press, New York. 580 pp.
Galway, J.G., 1956. The lifted index as a predictor of latent instability. Bull. Am. Meteorol. Soc. 37, 528–529.
George, J.J., 1960. Weather Forecasting for Aeronautics. Academic Press, New York. 673 pp.
Holt, M.A., Hardaker, P.J., McLelland, G.P., 2001. A lightning climatology for Europe and the UK, 1990–99. Weather 56, 290–296.
Holton, J.R., 1992. An Introduction to Dynamic Meteorology, 3rd ed. Academic Press, San Diego. 507 pp.
Huffines, G.R., Orville, R.E., 1999. Lightning ground flash density and thunderstorm duration in the continental United States: 1989–96. J. Appl. Meteor. 38, 1013–1019.
Huntrieser, H., Schiesser, H.H., Schmid, W., Waldvogel, A., 1997. Comparison of traditional and newly developed thunderstorm indices for Switzerland. Weather Forecast. 12, 108–125.
Jacovides, C.P., Yonetani, T., 1990. An evaluation of stability indices for thunderstorm prediction in Greater Cyprus. Weather Forecast. 5, 559–569.
Jefferson, G.J., 1963a. A modified instability index. Meteorol. Mag. 92, 92–96.
Jefferson, G.J., 1963b. A further development of the instability index. Meteorol. Mag. 92, 313–316.
Jefferson, G.J., 1966. Letter to the editor. Meteorol. Mag. 95, 381–382.
Lee, A.C.L., 1986. An experimental study of the remote location of lightning flashes using a VLF arrival time difference technique. Q. J. R. Meteorol. Soc. 112, 203–229.
Miller, R.C., 1967. Notes on analysis and severe storm forecasting procedures of the Military Weather Warning Center. Tech. Rep. 200, AWS, U.S. Air Force. 94 pp. [Headquarters, AWS, Scott AFB, IL 62225].
Miller, R.C., 1972. Notes on analysis and severe storm forecasting procedures of the Air Force Global Weather Center. Tech. Rep. 200 (Rev.), AWS, U.S. Air Force. 102 pp. [Headquarters, AWS, Scott AFB, IL 62225].
Murphy, A.H., Daan, H., 1985. In: Murphy, A.H., Katz, R.W. (Eds.), Forecast Evaluation. Probability, Statistics, And Decision Making in the Atmospheric Sciences. Westview Press, pp. 379–437.
Peppler, R.A., Lamb, P.J., 1989. Tropospheric static stability and central North American growing season rainfall. Mon. Weather Rev. 117, 1156–1180.
Reymann, M., Piasecki, J., Hosein, F., Larabee, S., Williams, G., Jimenez, M., Chapdelaine, D., 1998. Meteorological techniques. Tech. Note, AFWA, U.S. Air Force. 242 pp. [Air Force Weather Agency, AFWA, Offutt AFB, NE 68113].
Schaefer, J.T., 1990. The critical success index as an indicator of warning skill. Weather Forecast. 5, 570–575.
Schultz, P., 1989. Relationships of several stability indices to convective weather events in northeast Colorado. Weather Forecast. 4, 73–80.
Showalter, A.K., 1953. A stability index for thunderstorm forecasting. Bull. Am. Meteorol. Soc. 34, 250–252.
Väisälä, 1999. RS80 Series GPS Radiosondes. Upper-air systems product information, available at http://www.vaisala.com, Väisälä, 100 Commerce Way, Woburn, MA 01801.
Van Delden, A., 1998. The synoptic setting of a thundery low and associated prefrontal squall line in Western Europe. Meteorol. Atmos. Phys. 65, 113–131.
Van Delden, A., 2001. The synoptic setting of thunderstorms in Western Europe. Atmos. Res. 56, 89–110.
Von Storch, H., Zwiers, F.W., 1999. Statistical Analysis in Climate Research. Cambridge Univ. Press, Cambridge. 484 pp.
Yonetani, T., 1979. Instability index which takes account of the stabilities of a lower atmospheric layer for the purpose of thunderstorm prediction in the northern Kanto Plain. Rep. Natl. Res. Center Disaster Prev., 35–44 (in Japanese).