Friulian Journal of Science **2**. 2002, 59-76

The Analysis of the Thunderstorm Forecasts on the Friuli-Venezia Giulia Plain

DARIO GIAIOTTI[*] & FULVIO STEL[*]

Abstract. This work looks at four years of objective and subjective thunderstorm forecasts on the Friuli Venezia Giulia plain. This analysis is based on the usual attributes derived by the contingency table, that is the Probability Of Detection (POD), the False Alarm Rate (FAR) and the Hit Rate (HR), as well as more sophisticated skill indexes such as the Heidke skill Score (HSS). The categorical approach with the Brier Skill Score (BS) is adopted only for subjective forecasts. Results show that objective thunderstorm forecasts have an higher POD than subjective forecasts, but unfortunately even an higher FAR. This produces comparable HRs for both forecasts, although subjective forecasts are globally better than objective ones. Based on their behavior, to obtain a better product it is worth using a combination of both of them in the operative weather forecasts routinely issued by the Regional Meteorological Observatory of ARPA-FVG. The BS computed for the subjective forecasts highlights the positive feedback obtained from this analysis, carried out yearly by the authors. In particular, the subjective forecasts over the last year are much more calibrated.

Keywords. Thunderstorm forecasts, Forecasts verification, Weather forecasting.

1. Introduction. Friuli Venezia Giulia (FVG), is a small region of nearly 7800 km^2 in the north-east of Italy, bordering with the Carnic and Julian Alps on the north and with the Adriatic Sea – the lagoons of Grado and Marano – on the south. Even if the highest peaks in the Carnic and Julian Alps are not of particularly impressive heights (Mt. Coglians is less than 3000 m), these two chains are a natural and efficient barrier for the moist winds coming from south-west (Libeccio) and south-east (Scirocco)

[*] Regional Meteorological Observatory of ARPA-FVG, Cervignano del Friuli, Udine, Italy.

and making this region one of the rainiest of the whole Europe. The FVG plain, intensively under crop, gently decreases from the foothills (nearly 150 m above m.s.l.) toward south up to the lagoons in less than 80 km.

Due to its peculiar orography and geography, in spite of its smallness, this region hosts several kinds of meteorological phenomena, some of which could even be dangerous to people and property (Bechini et. al., 2001). Among these phenomena, the most important in the warm season are thunderstorms (Morgan 1989, 1990, 1991, 1997). In fact, from April to September, in the Friuli Venezia Giulia plain thunderstoms show an average frequency of 0.6 days (Giaiotti & Stel 2001). Because of their high frequency and their deep impact on human activities, since 1998 the Meteorological Observatory of the Regional Agency for the Environmental Protection of FVG (ARPA-OSMER) has been actively involved in a campaign whose aim is to improve operative thunderstorm forecasts in the short term (range of hours) as well as in the medium term (range of days). The latter will be looked at in the present work.

2. Materials and Methods

2.1 Definition of Stormy Day. The first step of every forecasting procedure is to clearly define the object of the forecasts (Murphy 1993). For this reason, it is important to provide a univocal definition of the observing variable linked to the object of the forecasts, that is the *observable,* and of the variable used for the forecasts, that is the *predictand.*

2.1.1 The observable: "daily storminess". The observable used in this study is "daily storminess", a dichotomous variable that can only assume two values: "yes" (the day is stormy)

Table 1. The Joint probability function for Subjective forecasts in the period 1998-2001.

Summer Season 1998		
	Observation	
Forecast	Yes	Not
Yes	0.25	0.03
Not	0.17	0.55

Summer Season 1999		
	Observation	
Forecast	Yes	Not
Yes	0.29	0.07
Not	0.21	0.43

Summer Season 2000		
	Observation	
Forecast	Yes	Not
Yes	0.30	0.06
Not	0.19	0.45

Summer Season 2001		
	Observation	
Forecast	Yes	Not
Yes	0.38	0.08
Not	0.11	0.43

and "no" (the day is not stormy). Because the final users of the forecasts are people, it is worthwhile to define the observable in a fashion which is close to common people's sensitivity. In fact, thunderstorms are sometimes very local phenomena and they could just be a small perturbation over a region without any significant precipitation. For this reason, in this work, daily storminess is defined considering thunderstorms only when they are a feature of the day on a sufficiently wide area over the FVG plain. To define daily storminess we used the observations of cloud to ground lightning strikes from the CESI/SIRF detection system (Iorio & Ferrari, 1995). The FVG plain is virtually covered by a regular network of elementary boxes whose mesh size is 4 km x 4 km (see Figure 1). When during a day, that is between 00:00 UTC and 23.59 UTC, at least one lightning strike is recorded in a box, that box is considered to be activated on that day. A threshold on the number of activated boxes is set, that is: a day is considered as stormy if more than 4 boxes have been activated, otherwise it is considered not stormy. On the basis of this, it has been shown (Giaiotti & Stel 2001) that stormy days have a high probability of being characterized by lightning and rainfall over a non-negligible area of the FVG plain.

2.1.2 The predictand in subjective and objective thunderstorm forecasts. The predicatands used at ARPA-OSMER for the medium-term forecasts are of two different kinds, according to the two different types of forecasts, that is *subjective forecasts* and *objective forecasts*.

Subjective forecasts have as predictand the probability of thunderstorm occurrence in the FVG plain, which is issued by a person, the forecaster, comparing the present with the past

Table 2. The Joint probability function for Objective forecasts in the period 1998-2001.

Summer Season 1998

	Observation	
Forecast	Yes	Not
Yes	0.33	0.12
Not	0.12	0.43

Summer Season 1999

	Observation	
Forecast	Yes	Not
Yes	0.25	0.11
Not	0.23	0.41

Summer Season 2000

	Observation	
Forecast	Yes	Not
Yes	0.45	0.20
Not	0.06	0.29

Summer Season 2001

	Observation	
Forecast	Yes	Not
Yes	0.42	0.21
Not	0.03	0.34

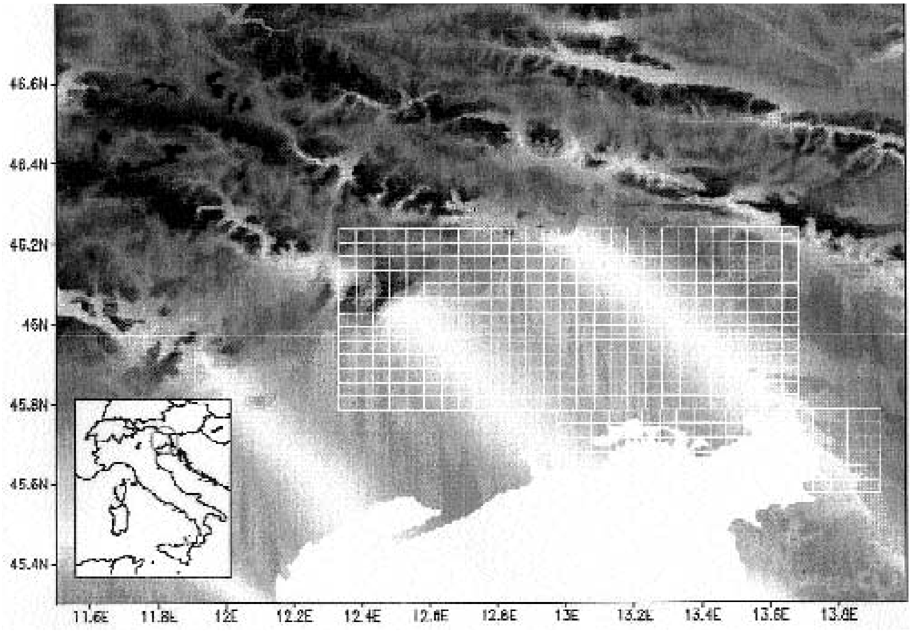

Figure 1. The network of boxes covering the plain of Friuli Venezia Giulia. The sizes of each elementary box is 4 km x 4 km.

situations on the basis of his own experience and his knowledge of the physical laws driving the atmosphere. Currently, at ARPA-OSMER, the information available to forecasters is obtained by five numerical models (see Appendix), two global and three local, from the network of synoptic stations managed by the Observatory and the radiosounding carried out by Italian Air Force at Campoformido (Udine). The two global models belong to the European Center for Medium Range Weather Forecasts (ECMWF) and the Deutscher Wetterdienst (DWD)[1]. The three local models are the ALADIN model (from Hidrometeoroloski Zavod – Slovenia), the Lokal modell of DWD and the HRM model (Italian Air Force). There are 30 synoptic ground stations managed by ARPA-OSMER which sample the atmosphere every hour, providing information about pressure, temperature, moisture, wind speed and direction. The radiosounding is launched every 6 hours almost in the middle of the FVG plain and gives information about the stability of the atmosphere. These local data sources provide fundamental information on the state of the atmosphere over the FVG plain, that is of the environment where thunderstorms could develop, and they are pivotal for the forecaster's decision. This determines the probability of thunderstorms' occurrence for the following day on FVG.

Objective forecasts have an index called coin index as predictand. It was developed on the basis of a statistical analysis made on historical cases. The values of the index are linked to the probability of thunderstorms' development on the FVG plain one day after the index has been issued. The index was defined using the upper-level data obtained by the radiosounding carried out by the Italian Airforce at Campoformido, while today the index is computed extracting the needed data from the ECMWF numerical model. The meteorological variables involved in this analysis are the geopotential height, the equivalent and saturated potential temperature (see Appendix) and the south-north and west-east wind components, all taken at the four mandatory levels of 1000 hPa, 850 hPa, 700 hPa and 500 hPa at fixed hours: 00:00 UTC, 12:00 UTC and the following day 00:00 UTC. To take into account a possible dependence of thunderstorms's occurrence on the year's period, even the sine and cosine of the Julian day are inserted in the data set, which is composed of 62 variables. To obtain the index, all the 62 variables are combined in a multilinear regression.

2.2 Forecast verification. The first analysis performed on forecasts is the comparison between the two conditional empirical distributions achieved by fixing observations connected to the forecast. In particular, for each predictand, forecasts were grouped into two classes: one including all forecasts referring to the observation of "not stormy day" and the other gathering all forecasts referring to "stormy day". Comparison between distributions was based on nonparametric Kolmogorov-Smirnov test (see Ledermann 1984). The second step in evaluating the quality of the thunderstorm forecast is the categorical approach. It is assumed that there is a discriminant value in the domain of forecasts predictand and that each forecast can be grouped into one of the two complementary classes: forecasts of "stormy day" and forecasts of "not stormy day". The discriminant value splits the whole domain of the predictands into two subsamples; in case of one dimensional monotone predictand a simple criterion to separate forecasts is the following:

$$\begin{cases} x_D \leq x \Rightarrow & \text{stormy day} \\ x < x_D \Rightarrow & \text{NOT stormy day} \end{cases} \qquad (1)$$

where x is the predictand variable, i.e. the index value or the probability issued, and x_D is the discriminant value. The categorical approach compresses the whole predictand information in a more compact two-state variable. In this way there is loss of information but there is the advantage of a very simple empirical joint probability density function of forecasts and observations can be made. Below is the general form of the joint probability density function:

	Observation	
Forecast	Yes	Not
Yes	**a**	**b**
Not	**c**	**d**

a is the fraction of forecasts of “stormy day” that have been successful since a stormy day was observed, **b** is the fraction of “stormy day” forecasts for which “a not stormy day” was observed, **c** is the fraction of forecasts of “not stormy day” for which a “stormy da” was observed and **d** is the fraction of successful forecasts of “not stormy day”. Of course, **b** and **c** are failed forecasts because observations have not matched the forecasts. From the above empirical probability function some important attributes of forecasts can be derived such as the probability of detection (POD), the false alarm rate (FAR), the bias (BIAS), hit rate (HR) and the skill scores. For further details see the Appendix. Subjective forecasts indicate the probability of stormy day occurrence. Therefore, it is important to estimate also the quality of the issued value of that predictand and this is not possible by a simple categorical approach. To access the whole subjective forecast information the attribute diagram and the Brier skill score were investigated; for further details see the Appendix.

2.3 Calibration of predictands. In the categorical approach, the discriminant value x_D plays an important role, since all the attributes, and the skill of the forecasts too, are a function of the selected discriminant value. x_D should be selected based on unambiguous criteria. For subjective forecasts the discriminant value arises naturally from the definition of the predictand, i.e. the probability of stormy day. That variable is defined in the domain [0%,100%] and the 50% value is a common used discriminant value to split it into two complementary subsamples. More complex is the identification of the discriminant value for objective forecasts. In that case, coin index has not a natural milestone in its domain,. It is thus necessary to calibrate the index. The calibration of coin index relied on the 1998 summer season observations. At the end of the season the empirical joint probability density function was computed, as shown in (2), as a function of the discriminant value x_D. Letting x_D varying throughout its domain, the best empirical probability function was found according to the maximization of the hit rate (HR) that in turn implies maximizing the Heidke skill score (HSS), see Appendix. The maximum skill in 1998 was reached for $x_D = -0.1$ and that was the discriminant value applied since then.

3. Results

3.1 Comparison between conditional distributions. Objective and subjective conditional empirical distributions do exist from the 1998 summer throughout 2001, as described in section 2.2. These were compared in pair according to the predictand selected. In all cases and for all the years the Kolmogorov-Smirnov test allows to reject the null hypothesis that the two distributions are the same, with confidence levels higher than 99%. These results indicate that for both types of forecasts the occurrence of stormy and not stormy days can be sufficiently correctly forecast. The evolution of the empirical distribution over the years needs some consideration. With regard to *objective forecasts* there are some differences in the median values from year to year and only slight differences in the shape of the distributions, that is they are close to be the same but with median values shifted only. With regard to *subjective forecasts* the distributions referring to "not stormy day" are more similar to bell-shaped ones but those of "stormy day" are almost flat, especially in 1998. In the following years the stormy days distributions of the predictand became slightly bell-shaped and this can be interpreted as a feedback action of forecasts verification as pointed out in the Discussion section.

3.2 Attributes of the forecasts. It is interesting to look at the evolution of forecast quality over the years for both *subjective* and *objective* thunderstorm forecasts. First of all, it is worth noting that *subjective forecasts* usually underestimate the number of stormy days, i.e. BIAS < 1. On the other hand, *objective forecasts* mainly overestimate them even if in 1999 they behaved like *subjective forecasts*, see Figure 2a. Concerning the probability of detection, *objective forecasts* performed better than *subjective* ones except in 1999. Figure 2b shows the high values of POD (~ 90%) for coin index in the past years, but *subjective forecasts* show around 60% of stormy days. During the last summer season, *subjective forecasts* enhanced their POD (78%) coming closer to coin index performance. High values of POD are a necessary, yet not sufficient, condition for good forecasts. One can reach the maximum POD issuing forecasts of stormy day always, but this implies a lot of false alarms. Forecast become unreliable. The FAR of coin index is higher than that of *subjective forecasts*, see Figure 2c. So far*, subjective forecasts* have maintained a FAR between 0.1 and 0.2, while *objective forecasts* show a 0.3 FAR. By combining the probability of detection and the false alarm rate it is possible to investigate forecasts performance in a deeper way and this can be done by the Hit Rate. Coin index and *subjective forecasts* have comparable HR as the negative and positive effects that each of them have act as a compensatory mechanism. S*ubjective forecasts* have always slightly higher values than HR, see Figure 2d. The 1999 season was the worst of all four: *objective forecasts* had a very low POD and *subjective* ones a significantly high FAR, which resulted in low hit rates. The skill of the forecasts we are analyzing was obtained comparing them with reference random forecasts, that is forecasts generated randomly from the marginal distribution of the observations and the forecasts, as as follows:

$$HSS = \frac{HR - HR_r}{1 - HR_r} \qquad (3)$$

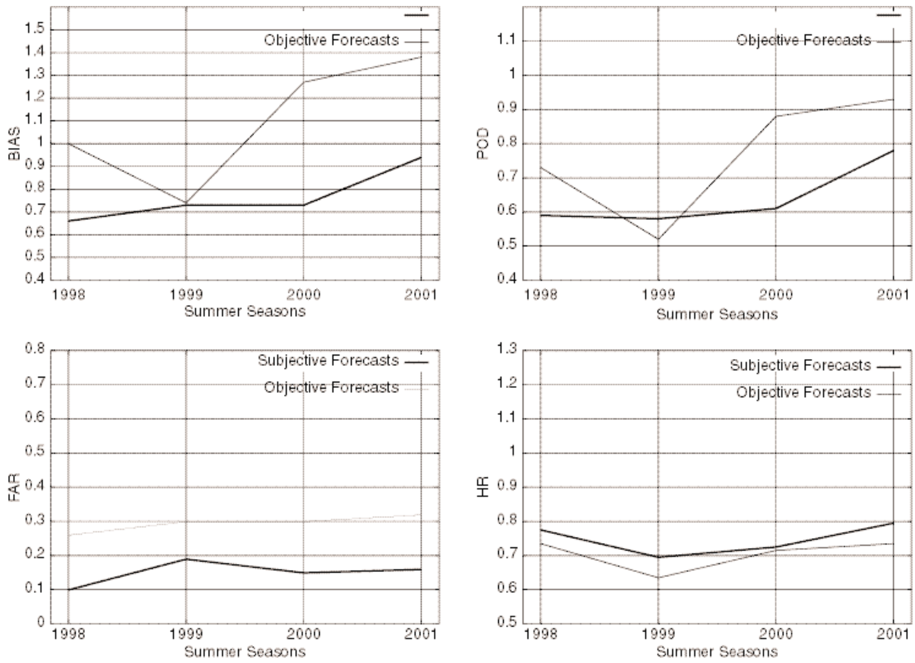

Figure 2. The evolution of some features of subjective (bold line) and objective forecasts (light line) throughout the period 1998-2001: a) The BIAS, b) POD, Probability of detection, c) FAR, False alarm rate, d) HR, Hit rate.

where HR is the hit rate of the forecasts for which we are evaluating the skill, that is coin index or *subjective forecasts*, HR_r is the hit rate of the random forecasts and 1 is the hit rate of perfect forecasts, see also Appendix. HSS is called the Heidke skill score. In case of perfect forecasts HSS = 1, otherwise HSS < 1. If HSS < 0 forecasts have less skill than random ones. In Figure 3a the Heidke skill scores for coin index and *subjective forecasts* are reported; their skill is comparable throughout the period of study, but *subjective forecasts* are a bit better than *objective* ones. Both have significantly more skill than random forecasts (HSS > 0).

3.3 Analysis of issued probabilities. *Subjective forecasts* provide more information than coin index since the issued probability of stormy day gives information on the forecaster's confidence of the occurrence of the event. In this case, verification of forecasts has been carried out by means of Brier score.

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2 \qquad (4)$$

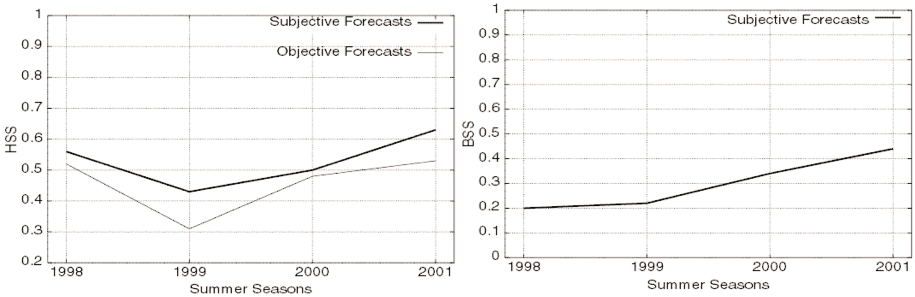

Figure 3. The skill of the forecasts in the period 1998-2001. Light line refers to Objective Forecasts while Subjective Forecasts are in bold line: a) Heidke skill score. b) Brier skill score. In this case subjective forecasts are present only since Brier skill score is not applicable to objective forecasts.

where: BS is Brier score, n is the number of forecast-observation pairs, p_i is the issued probability of stormy day, $p_i \in [0,1]$, and o_i is the observed weather, that is $o_i = 1$ in case of stormy day, otherwise $o_i = 0$. BS is always a number ranging between 0, in case of perfect forecasts, and 1, in case of completely wrong forecasts. According to the algebraic decomposition of Brier score derived by Murphy (1973), it is possible to split BS into three components that are: *reliability, resolution* and *uncertainty.* All these components are described in detail in the Appendix. In Figure 4, the Brier score, its components and the climatological probability of stormy day are reported for each of the considered years. It is worth noting that BS decreased from 1998 to 2001. This points to an improvement in subjective forecasts. This improvement is independent of the climatological features of stormy days since the climatological probability does not change too much from year to year. Rather, it limits the improvement of the skill to a great extent. Throughout the period of study, it was very close to 0.5, which is the value that maximizes the *uncertainty* component of Brier score. *Reliability* shows a decreasing trend from 1998 and it reaches values close to zero in 2001. This *means subjective forecasts* improved their calibration, that is issued probabilities are very close to the observed frequencies of stormy days. *Resolution* has increased, especially over the last two years. The combination of reduced *reliability* and increased *resolution* has produced lower values of BS in the last two years than in 1998 and 1999. The skill of the issued probabilities was computed by comparison of their BS and those expected from climatological forecasts, that is forecasts made every day issuing the climatological frequency of stormy day, as follows:

$$BSS = \frac{BS - BS_c}{0 - BS_c} \qquad (5)$$

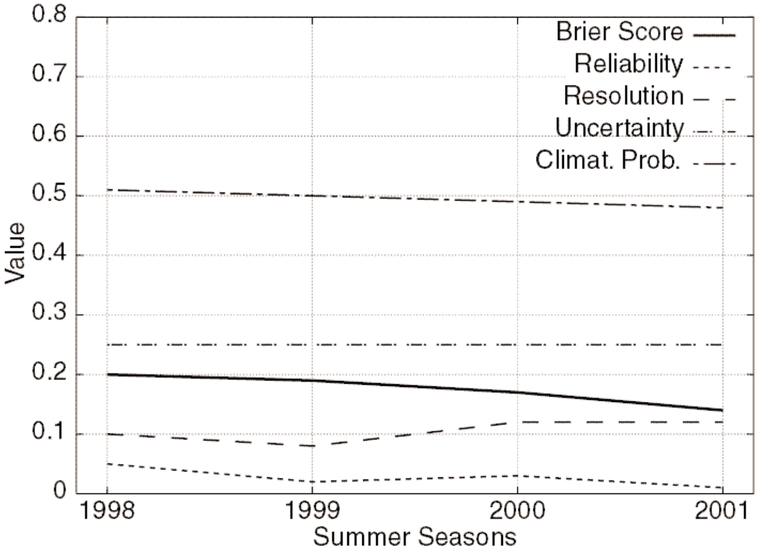

Figure 4. The Brier score and its components throughout the period 1998-2001. Solid bold line reports the brier score. Short dashed line is the reliability term, long dashed line is the resolution term and double points segment line is the uncertainty. It is also reported the climatological probability of stormy day over Friuli Venezia Giulia plain, long-short segments line. In the last two years, the increase of resolution and the decrease of reliability terms have produce the observed reduction of the Brier score, that in turn means an improvement of the subjective forecasts quality. Interesting is also the high an almost constant value of uncertainty term.

BSS is the Brier Skill Score, BS is the Brier score of the forecasts, BS_c is the Brier score produced by climatological forecasts and 0 (equation 5) is the BS for perfect forecasts. Figure 3b shows the big improvement in the skill of *subjective forecasts* by means of the significant increase in BSS values from 1999 onwards. It has doubled in the last two years. Figure 5 shows the improvement in *subjective forecasts* and reports the attributes diagrams (Hsu & Murphy 1986) for each summer season. Attributes diagrams report the observed frequency of stormy days as a function of issued probability; observations, o_i, are clustered according to the same forecast probability, then the average of the observations, $\bar{o}_j$, (see the Brier score decomposition in the Appendix) gives the observed frequency. Forecasts, for which issued probability match exactly the observed occurrence of stormy day, lay on the diagonal line crossing the diagram from the bottom left corner to the upper right one. *Reliability* is zero and they are perfectly calibrated forecasts. The contribution to the *resolution* of each class of issued probability is a function of the distance of the reported point from the no-resolution line. The constant value of that line is the overall climatological probability of stormy day. The calibration of *subjective forecasts* has increased over the years (Figure 5). In 1999 and 2000 forecasts generally underestimated the observed frequencies, points were almost above the diagonal, but in 2001 almost all were close to it. Spikes reported in the plots refer to classes of forecast probability including one or two cases only.

4. Discussion. This 4-year analysis clearly shows that *objective forecasts* have higher values of probability of detection (POD) than *subjective* ones, but the reverse is true for false alarms (FAR). The BIAS of forecasts, namely the overestimate of stormy days by coin index and the underestimate of *subjective forecasts*, plays a role in this too. In spite of those differences, *subjective* and *objective forecasts* have comparable hit rates (HR), those of *subjective forecasts* being a bit larger than those of coin index. The overall skill of both kinds of forecasts can be compared by the Heidke skill score (HSS). According to the definition of HSS, see equation (3), both have significantly more skill than random forecasts. Fluctuation of forecasts quality was observed, in particular the probability of detection is the most variable attribute; its variability is strictly connected to that of BIAS. Much more robust is the FAR. Both those features result in a rather limited variability of hit rate and Heidke skill score. The variability is greater for *objective forecasts* than for *subjective* ones. In 1999 there was a drop-off skill for both kinds of forecasts since coin index had very low POD values and *subjective forecast* showed a large number of false alarms. It is not clear why *objective forecasts* underestimated stormy days (BIAS < 1) in 1999; the climatological frequency of that year is slightly greater than 1998, the year of calibration (see the level of no resolution line in Figures 5). However, it was almost the same for the subsequent years when the BIAS is higher than one and the skill is higher. Regarding subjective forecasts, it is very likely that during 1999 forecasters had been influenced by the analysis of the forecasts issued in the previous year. At the end of 1998 almost all the forecasts issued during that summer season with a probability greater or equal 50% had an observed frequency equal to 1. Especially for the 50% class (see Figure 5), 1998 showed a very low calibration. Forecasters reported that most of the 50% issued values were related to situations of high difficulty in the evaluation of weather evolution. All recognized that better skill would have been reached avoiding the 40%-50% classes going towards the more extreme ones. In 1999 forecasters forced their judgment (shown in Figure 6) but too many difficult situations had issued percentages less than 40% and that resulted in a sharp underestimate of stormy days. In the later analyses this problem was pointed out, but forecasters deemed as advantageous the idea to force their forecasts and this is shown in the his-

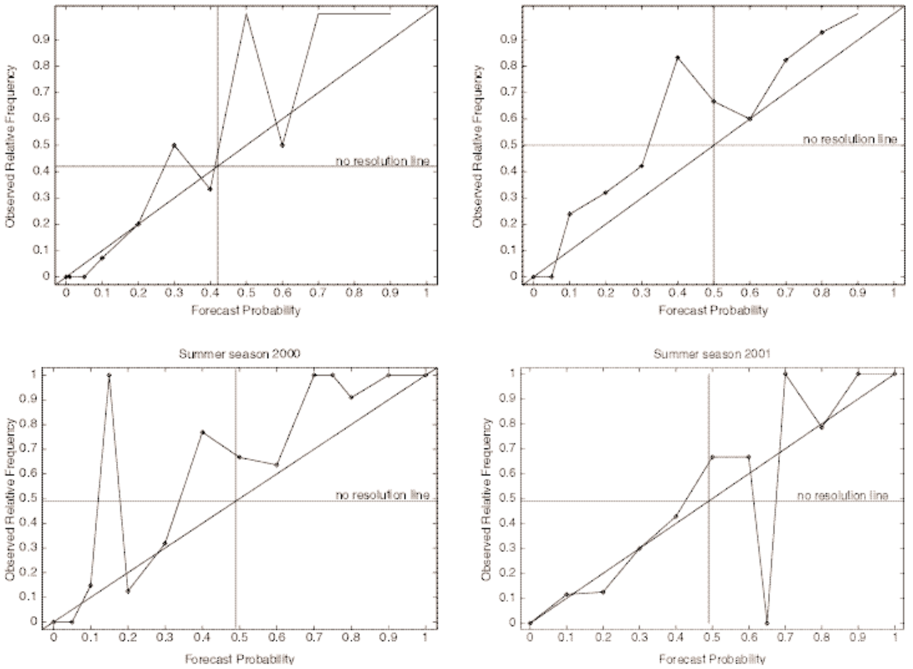

Figure 5. The attributes diagrams of subjective forecasts for each summer season in the period 1998-2001.

tograms for the 2000-2001 years (Figure 6). The above behavior is an example of the feedback effect on forecasts produced by their verification. Considering the low FAR values of *subjective forecasts* (Figure 2c), and the large relative frequencies in the high-percentage classes, namely 70%, 80% and 90%, it is clear that when forecasters strongly believe in the occurrence of a stormy day, then it will occur. Similarly, they are successful in issuing low probabilities when days will be not stormy. Difficulties arise in unclear weather situations for which stormy days are not recognized. Several of them are missing and they reduce the POD. To improve forecasts the authors suggest a combination between *subjective* and *objective forecasts*, that is the forecaster issues his probability of stormy day for tomorrow, not looking at coin index, then if his probability is outside the range [40%, 60%], the probability is retained. Otherwise, the coin index is checked and the forecasts become of stormy day (≥ 50%) or not stormy day (< 50%) according to *objective forecasts*. The issued probability is modified as a consequence of the limit imposed by coin, but the new value is upon the judgment of the forecaster. Virtual forecasts have been pro-

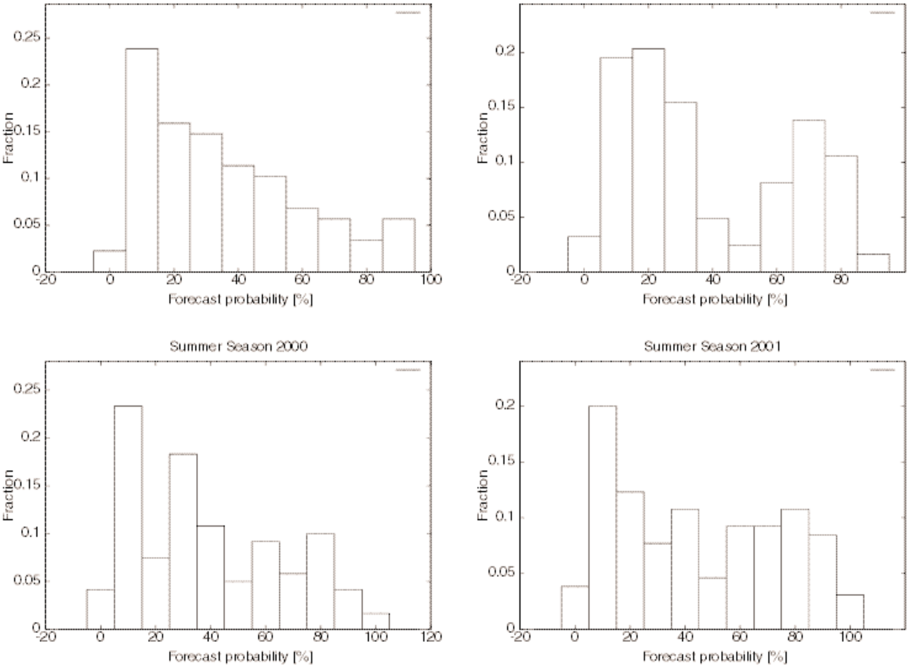

Figure 6. Histograms of issued probability for the subjective forecasts in the period 1998-2001.

duced for subsamples of the past summer seasons, with the application of the above- mentioned rule. In particular, results of *objective forecasts* were transformed in 60% issued probability in case of "stormy day" and 40% in case of "not stormy day". The evaluation of this virtual forecasts shows a significant improvement in the POD and an acceptable increase in FAR, with respect to pure *subjective forecasts* together with slight greater HR. This leads to small but higher values of Heidke skill score. Of course, the analysis of issued probabilities makes no sense since the algorithm used to modify the issued probabilities does not reproduce the human choice for the probability issued.

5. Appendix

5.1 Numerical models and meteorological variables. Numerical models are the most powerful tools for the modern weather forecasts. Essentially, they are huge and complex programs used to simulate the evolution of the atmosphere, knowing its present status and the laws that drive this system. Because it is not possible to reproduce the atmosphere in all its complexity and continuity, this system is discretely represented on a finite three-dimensional grid. At each point of the grid a set of variables (e.g. temperature, humidity, wind speed and direction, etc.) is defined with their initial conditions, then these variables evolve according to the laws of physics and taking into account the global forcings represented by orography, geography and solar radiation. The output of the numerical models is represented by the fields of the fundamental variables shown at standard (e.g. mandatory) levels and at fixed future times. Knowing this values, the forecaster is able to issue his weather forecasts.

Sometimes in meteorology it is useful to define new variables that, even if they are not close to the common sensitivity, they can reduce the computational time in the numerical models or are much strongly linked to the meteorological phenomena to which they are referring. One of these variables is geopotential height H, defined at the height h from the Earth's center as:

$$H(h_1) = \frac{1}{9.80665}\int_0^{h_1} \Phi(z)dz \qquad (m)$$

where z is the distance from the Earth's center, Φ(*z*) is the gravitational potential of Earth and 9.80665 is the standard acceleration of gravity at the mean sea level. This new variable is useful because a unit mass parcel of air moving along lines of equal geopotential does not change its mechanical potential energy. Another variable often used in meteorology and dealing with convective phenomena is the equivalent potential temperature. In fact, considering a unit mass parcel of air moving vertically in the atmosphere, it is necessarily subject to an expansion (upward movement) or compression (downward movement), which changes its temperature. Moreover, if the air mass is not dry, it changes its temperature in the vertical movement when condensation/evaporation occurs due to the release of latent heat. For this reason, meteorologists use a variable called equivalent potential temperature, which is the temperature that a mass of air will have if brought down to the sea level and if subject to a complete condensation of the water vapor. The formula used to retrieve the equivalent potential temperature θ_e is the following:

$$\theta_e = T\left(\frac{P_0}{P}\right)^{\frac{R}{C_p}} e^{\left(\frac{Lw}{C_p T}\right)}$$

where *T* is temperature in Kelvin degrees, *L* is the latent heat of condensation, *w* is the mixing ratio (mass of water vapor per unit of mass of air), P_0 is the pressure at sea level, *P* the pressure of the unit mass parcel of air, C_p is the specific heat of air at constant pressure and *R* is the gas constant. If temperature *Ts* at which a given mass of air is saturated by water vapor is used instead of its temperature *T,* then the equivalent saturated potential temperature θ_{es} is obtained. These temperatures are useful in weather analysis and forecasting because they facilitate comparison between masses of air at different heights and then calculation of instability.

5.2 The Joint Probability Density Function. The particular structure of a 2x2 joint probability density function for forecasts and observations is displayed in (6). It is the typical function for categorical forecasts by which the probability of occurrence or not occurrence of an event is reported.

Forecast	*Observation* Yes	Not
Yes	**a**	**b**
Not	**c**	**d**

a is the fraction of occurred events that have been forecast successfully, **b** is the fraction of events not occurred but forecast, **c** is the fraction of events occurred but not forecast and **d** is the fraction of events not occurred and not forecast. **b** and **c** are failed forecasts because observations have not matched the forecasts while **a** and **d** are successful forecasts. Several scalar parameters can be calculated from a 2x2 joint probability density function. Below are the parameters used in this work:

$$BIAS = \frac{a+b}{a+c}$$

BIAS is the ratio between the number of forecast events and the number of observed events. Forecasts with BIAS = 1 are unbiased ones, if 0 ≤ BIAS < 1 forecasts underestimate the number of occurred events: if BIAS > 1, forecasts overestimate it.

$$POD = \frac{a}{a+c}$$

POD is the probability of detection, that is the number of successful forecasts of the event over the whole amount of occurred events. If POD = 1 all occurred events were forecast successfully.

$$FAR = \frac{b}{a+b}$$

FAR, false alarm rate, is the number of forecast events that were not successful since the event did not occur. FAR values close to zero are typical of good forecasts, while larger FARs usually characterize forecasts that overestimate the number of occurred events.

$$HR = \frac{a+d}{a+b+c+d}$$

The hit rate, HR, evaluates the portion of correct forecasts over the whole amount of issued forecasts. This parameter takes into account forecasts of the event that were successful and also forecasts of the absence of the event that were successful. The closer the HR is to 1, the higher the quality of the forecasts. Of course, HR = 0 indicates completely wrong forecasts.

The skill of the forecasts is evaluated based on the comparison between the performance of the forecasts under evaluation and a reference forecasts Common reference forecasts are those obtained using climatological probability for the occurrence of the event or forecasts produced by random procedures, that is forecasts that do not benefit of any critical judgment of future weather. The comparison is made as follow:

$$SS = \frac{P - P_{ref}}{P_{perf} - P_{ref}}$$

and the skill score, SS, is a measure of the difference between the selected parameter of the forecast under evaluation, P, and the same parameter computed for the reference forecasts P_{ref}, relative to the difference between the perfect forecasts P_{perf} and the same reference. The Heidke skill score is defined considering the hit rate as parameter. See equation (3) in the article.

5.3 Decomposition of the Brier Score. According to the definition of Brier score, see equation (4) in the article:

$$BS = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$$

It can be split it into several components, Murphy (1973), and each of those components depends on some particular feature of the probabilities issued. This has the advantage that more information can be extracted on the performance of the probability forecasts. The splitting is based on the clustering of all n pairs (y_i,o_i) of forecasts and observations in N classes belonging to the same issued value of probability y_j $j \in \{1,2,3, ..., N\}$. For each class $\bar{o}_j$, the observed relative frequency, is computed as follows:

$$\bar{o}_j = \frac{1}{m_j}\sum_{k=1}^{m_j} o_k(j)$$

$o_k(j)$ are the observations in the class belonging to the issued probability y_j and m_j is the number of the observations in that class. Of course, $n = \sum_{j=1}^{N} m_j$ and the overall relative frequency of the event, that is the climatological occurrence of the event, is:

$$\bar{o} = \frac{1}{n}\sum_{i=1}^{n} o_i = \frac{1}{n}\sum_{j=1}^{N} m_j \bar{o}_j$$

According to the above definitions, the decomposition of Bier score is the following:

$$BS = \frac{1}{n}\sum_{j=1}^{N} m_j (y_j - \bar{o}_j)^2 - \\ - \frac{1}{n}\sum_{j=1}^{N} m_j (\bar{o}_j - \bar{o})^2 + \bar{o}(1 - \bar{o}) \qquad (7)$$

Brier Score can be split intohe sum of three terms. The first term onthe right of equation (7) is the reliability component. Reliability summarizes the calibration of the forecasts, that is how much each issued forecasts is closer to the observed relative frequency $\bar{o}_j$. Reliability should be close to zero for good forecasts;in particular, for perfect forecasts reliability equals zero. The second term is resolution. This component explains how much forecasters are able to discriminate periods in which the probability of the event is different from the overall climatological probability o. In this term issued probability does not appear but is a function of issued forecasts since $\bar{o}_j$ depends upon them.

The last term refers to uncertainty and is a component of Brier score not depending on the ability of the forecasters, since it is function of the overall climatological probability only. This term is not negative and reaches its maximum at $\bar{o} = 0.5$. In case of rare events $\bar{o} \cong 0$ or very frequent events $\bar{o} \cong 1$, the contribution of uncertainty to the Brier score is small, but for events with climatological probability close to 50% it is relevant. This means that there is large uncertainty in forecasting the event's occurrence.

Notes

[1] From year 2002 onwards, forecasters at ARPA-OSMER could use a new product from ECMWF, called "ensemble forecasts", which consists in 50 runs of the same numerical model obtained starting from 50 slightly different initial conditions. In this way, the forecasters could even obtain an estimate of the robustness of the numerical forecasts due to the uncertainties in the initial conditions.

References/ Bibliografie

Giaiotti B.D. & Stel F. (2001). A comparison between subjective and objective thunderstorm forecasts. *Atmos. Res.*, 6: 111-126.

Hsu W.R. & Murphy A.H. (1986). The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, 2: 285-293.

Iorio R. & Ferrari D. (1996). *Proceedings of the 23rd International Conference on Lightning Protection (ICLP)*. Florence, Italy.

Ledermann W. (1984). Distribution-Free methods. In Ledermann W. (Ed), *Handbok of Applicable Mathematics*, vol. 6. New York: John Wiley and s., pp. 603-630.

Murphy A.H. (1973). A new vector partition of the probability score. *J. Appl. Meteorol.*, 12: 595-600.

Murphy A.H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast*, 8: 281-293.

Wilks D.S. (1995). *Statistical Methods in the Atmospheric Sciences*. Academic Press.