# CORRESPONDENCE

## Comments on "Automated 12–36 Hour Probability Forecasts of Thunderstorms and Severe Local Storms"

STEVEN J. WEISS, CHARLES A. DOSWELL III AND FREDERICK P. OSTBY

*National Severe Storms Forecast Center, Kansas City, MO 64106*

20 March 1980 and 4 August 1980

The recent paper by Reap and Foster (1979) describes an automated system to forecast thunderstorms and severe local storms. Since their aim is to provide guidance that can be used in issuing convective outlooks by the National Severe Storms Forecast Center (NSSFC), it is appropriate to comment on the paper from the viewpoint of the operational forecasters at NSSFC. We shall focus our discussion on three major aspects which concern us: 1) the predictand samples, 2) the probability equations and 3) the verification results.

### 1. Predictand samples

The predictand data for localized severe storms consist of reports of tornadoes, large hail and convective gusts $\geq$ 93 km $h^{-1}$ or wind damage. While Reap and Foster state that the "identifiable sources of error" in the sample have been removed at NSSFC, this should not be taken to imply a predictand sample that is as accurate and unambiguous as, for example, temperature and precipitation records. Indeed, as pointed out in many papers (e.g., McNulty *et al.*, 1979; Galway, 1977; Asp, 1963; Court, 1970; Abbey, 1976), a myriad of problems associated with tornado climatological data exists. Foremost among these is the influence of population density. Since tornado verification requires either direct observation or evidence of tornadic damage, reported tornadoes are often clustered near urban areas. Further problems include the difficulty in identifying nocturnal tornadoes, and nonuniform reporting procedures.

Similar problems are associated with verifying large hail and wind damage/gust reports. Owing to the less spectacular nature of these phenomena when compared to the tornado, these events often go unreported. Also, non-tornadic severe thunderstorm reports can be "filtered out" during the process of transmitting severe storm reports to NSSFC for archival. Thus, it must be recognized that the severe local storm report statistics contain many biases and limitations and likely do not accurately reflect the true distribution of these events. Any weaknesses inherent in these data are bound to be carried over into any predictive technique which relies on the storm reports as input. Naturally, it is recognized that this data problem is shared by both subjective and statistical forecasts. However, it should be noted that the statistical methods used by Reap and Foster choose conditions relative to a "typical" severe weather event. A potential problem exists in doing so, since a large proportion of reported severe thunderstorms occur in "non-typical" conditions.

Turning to thunderstorms, the predictand data consist of MDR data; VIP values of 3 or greater are used by Reap and Foster to indicate the occurrence of a thunderstorm. Any statistical relationship between an actual thunderstorm and a particular VIP value remains questionable, especially when the radar echo height is not accounted for. Also, the thunderstorm predictand data are seriously affected by many factors, including the lack of VIP equipment on a number of radars during the sample period, a nonuniform state of radar calibration, variations between radar fields of view, and the interpolation of frequency values in data-void areas.

As of September, 1975, nearly half (23 out of 48) of the NWS network radars were *not* equipped with VIP. Lacking VIP equipment, the radar operator must estimate MDR values by manually attenuating the echo return. Reap and Foster acknowledge that this method can lead to MDR errors of one or two levels. However, they state "the errors are essentially random and should not bias the MDR sample."

Our operational experience at NSSFC in the Convective SIGMET Aviation program, and its predecessor, the Radar Analysis and Development Unit, is not consistent with this assumption. It

0021-8952/80/111328-06$05.50

is quite common for adjacent radars, one VIP-equipped and the other non-VIP, to disagree by 2 *or more* VIP levels when monitoring the same thunderstorm cell. Further, a nonuniform calibration state between radars, an operational fact of life, contributes to additional discrepancies. Even adjacent radars, identically equipped with VIP apparatus, may differ dramatically solely as a result of this problem. Of course, differences in radar operator skill also affect the storm detection reports.

When the radar reporting system is examined from an overall operational viewpoint, it becomes apparent that different radars see the same convective phenomena differently, over extended periods of time. Experience indicates that some radar stations consistently report VIP levels too high, while others have a low bias. Reap and Foster have not provided any evidence that those differences are, indeed, random or that the spatial biases can be dismissed.

This problem is reflected in the MDR-derived climatology for April shown by Reap and Foster (their Fig. 3), which depicts the maximum thunderstorm frequency in southeast Nebraska. This apparent maximum could be the result of a positive MDR bias at the Grand Island radar site (a non-VIP station during 1974–76). The problems with the MDR thunderstorm climatology are apparent when their Fig. 3 is compared to standard thunderstorm climatological statistics; e.g., NWS Local Climatological Data (LCD) indicates the maximum frequency of *reported* thunderstorms in April to encompass eastern Oklahoma, Arkansas, northeast Texas and south-central Missouri (Fig. 1). If, perhaps, the period 1974–76 is in some way unrepresentative of the period shown in Fig. 1, then the representativeness of the MDR climatology is also questionable.

Further, the maximum April frequency of reported thunderstorms coincides with a significant minimum as determined by MDR values (Reap and Foster's Fig. 3) over northeast Texas and adjacent areas. This MDR minimum could be the result of interpolation in a data-void area, since it corresponds almost exactly to the shaded blocks in the northeast Texas region of Reap and Foster (their Fig. 1).

It appears there are irreconcilable inconsistencies between MDR-indicated and actual thunderstorm frequencies. This suggests that the MDR data contain significant *nonrandom* errors and do not accurately portray the frequency of thunderstorm occurrence.

## 2. The probability equations

Weaknesses in the predictand data become quite significant when one examines the probability equations generated by the MOS technique. As indicated in Reap and Foster (1979, Tables 4 and 5), interactive predictors incorporating thunderstorm and severe local storm climatological frequencies are the leading predictors in the forecast equations. This may well be the result of the "non-typical" nature of severe thunderstorm reports, reflecting the influence of mesoscale perturbations unresolved by the models, and the wide variety of large-scale conditions under which severe thunderstorms can occur. Unlike the vast majority of MOS forecasts, the automated thunderstorm and severe local storm forecasts are highly dependent on modulating the climatology of these events.

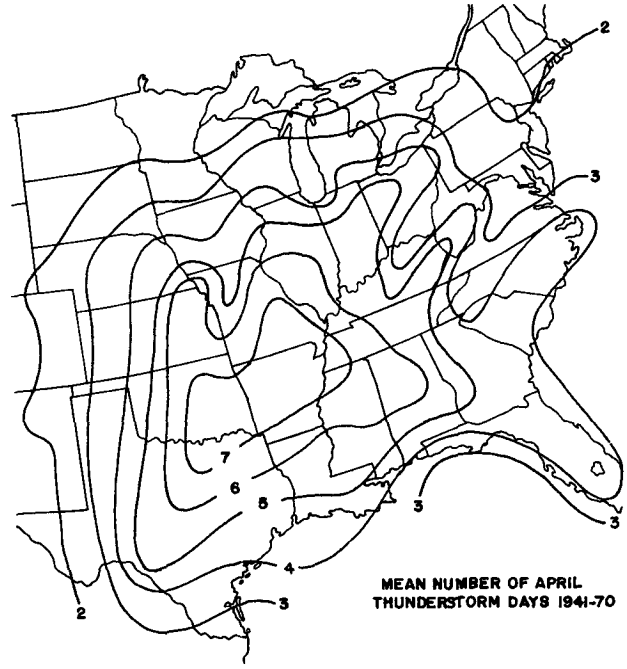


FIG. 1. Mean number of reported thunderstorm days for April. Data extracted from NWS Local Climatological Data (LCD) based on 1941–70 period.

This heavy reliance on climatology becomes even more evident when, despite switching to LFM model input in 1978 (the equations were developed from the 6-layer PE model input), "no noticeable differences were observed in the verification statistics for 1977 versus 1978." This is quite surprising, since the truncation error (and associated speed bias) of the LFM is significantly less than that of the 6-layer PE. For instance, this change to the LFM model altered the verification scores for the MOS-derived probabilty of precipitation amount (PoPA) forecasts (Zurndorfer, 1980).

Owing to the limited ability of the model output to provide information pertaining directly to the severe storms forecast problem (as revealed by Reap and Foster's low cumulative reductions of variance), it becomes increasingly important to develop

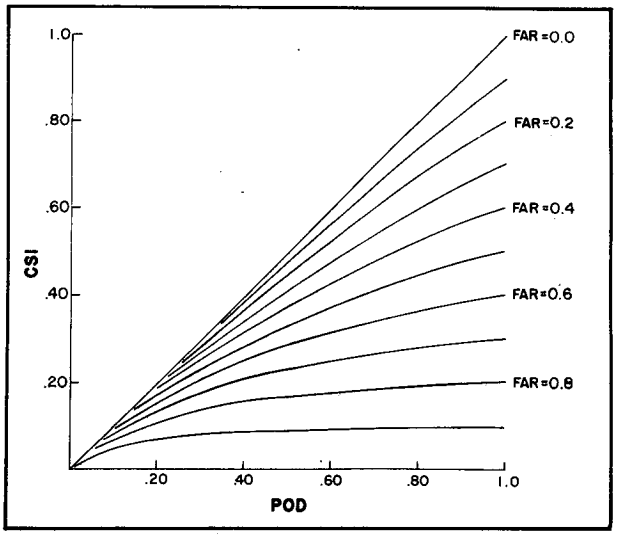


FIG. 2. Critical Success Index (CSI) as a function of the Probability of Detection (POD) of various False Alarm Ratios (FAR).

climatological frequencies which accurately reflect the occurrence of the predictands. Unfortunately, as indicated previously, we feel that both their severe local storm and thunderstorm frequencies are flawed in ways which limit the value of the forecast product.

### 3. Verification results

Finally, consider the verification results. Since SELS produces categorical forecasts of thunderstorms and severe local storms, the automated probability forecasts need to be transformed into similar categorical forecasts to provide meaningful guidance and comparison.

To accomplish the transformation from probabilistic to categorical thunderstorm forecasts, Reap and Foster develop statistics based on the Critical Success Index (CSI) technique of Donaldson *et al.*, 1975. Note however, that the threshold value that maximizes the CSI (35% probability) is only established *after the fact*, i.e., after analyzing the 1977–78 data, which in effect renders the sample no longer "independent", but is now really a "dependent" sample to be applied to some other body of data. This problem again arises when the severe local storm unconditional probability threshold is chosen for Fig. 8 of Reap and Foster. The 1977–78 data are used to establish the threshold which is then applied to those same data. This would seem to give an unfair advantage when computing comparative verification scores against some other technique that is not influenced by the data it is attempting to evaluate.

While the bias or reliability of the conditional severe local storm forecasts (Reap and Foster, 1979, Table 8) appears impressive, the use of standard statistical techniques for rare event phenomena must be questioned. As indicated by Foster and Reap (1978), "the higher the probability for a given area, the greater the expected coverage." This essentially agrees with Murphy's (1978) contention that an average point probability forecast is equivalent to an expected areal coverage forecast. Thus, for example, a 20% probability of precipitation (PoP) corresponds to 20% areal coverage, and is generally viewed as a relatively low probability event. Conversely, for severe weather this is a dense coverage, qualifying as an outbreak day (Weiss, 1977). It is not clear that statistical methods used successfully for common weather events (i.e., reliability tables for PoP forecasts) are applicable to forecasts of rare events like severe local storms. Operational experience suggests that, when interpreted in terms of coverage, the performance of the severe local storm conditional probabilities is not properly illustrated in Table 8.

The verification scores presented in Table 9 of Reap and Foster must also be evaluated in light of the verification technique used. As developed by Donaldson, the CSI is intended to verify point forecasts, not large area forecasts such as severe weather outlooks. When the CSI in its "pure" form is applied to the verification of outlooks, two major problems arise. First, the False Alarm Ratio (FAR) becomes excessively high (generally $>0.90$). Each of the MDR blocks within the forecast area, if observed nonsevere, contributes to category Z in their Table 7. However, this application assumes that every MDR block within the forecast area is expected to include at least one severe local storm. This is not a valid assumption. As defined in the NWS Operations Manual (1979), only a small fraction of the MDR blocks within the forecast area is anticipated to be affected by severe storms (generally $<10\%$).

Second, unless the FAR is redefined to reflect more accurately the true purpose of the forecast, the resulting CSI becomes quite meaningless. Note that if the FAR remains very high, the CSI is insensitive to variations in the Probability of Detection (POD). For example, when the FAR = 0.90, the CSI is essentially constant within a POD range from 0.3 to 1.0 (Fig. 2)! Therefore, an inappropriately high FAR dominates the CSI, limiting it to relatively low values. While the CSI can range from 0 to 1, with the higher numbers indicating a better forecast, the CSI scores in Reap and Foster's Table 9 reflect very little skill for either product, despite favorable POD values. An example of the problems with the authors' verification technique can be found by applying it to a hypothetical perfect NSSFC outlook for 3 April 1974 (Fig. 3). This perfect forecast has 31% density coverage (31% of the MDR blocks af-

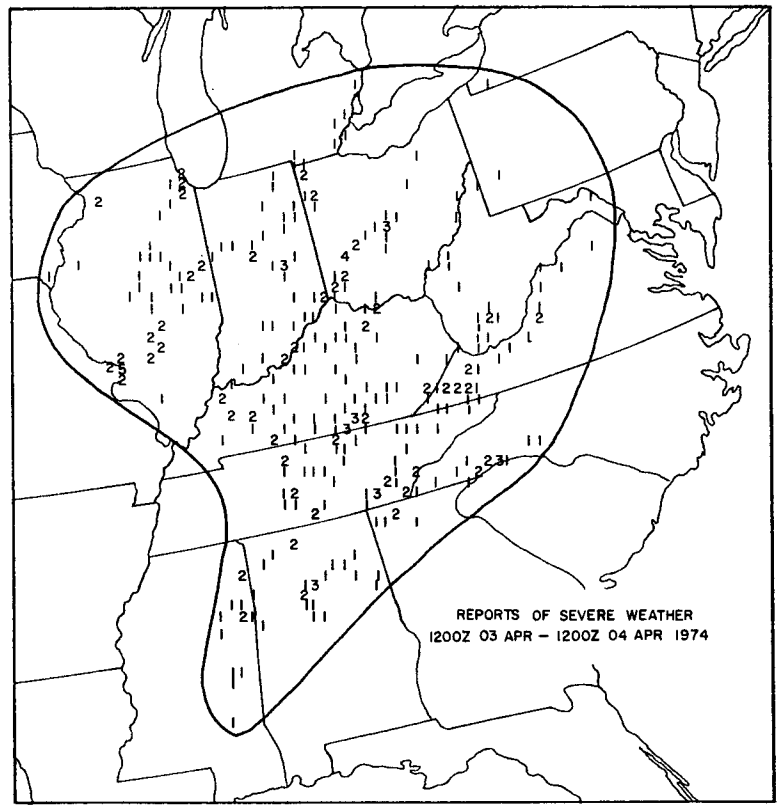


FIG. 3. Hypothetical perfect SELS-type severe weather outlook area (within the closed line) for the period 1200 GMT 3 April–1200 GMT 4 April 1974 and reported severe weather. Numbers indicate the total number of severe weather reports (including tornadoes, hail and severe windstorms) within an area of ~490 km$^2$. Roughly four such blocks are equivalent to an MDR block.

fected by severe events). This means that 69% of the blocks are devoid of severe storms, corresponding to an unmodified FAR of 0.69. Since all reports fall within the outlook area (POD is 1.0), the resulting CSI is only 0.31 on the biggest tornado day in recorded history! Clearly, it is difficult to attain a high CSI using this method even if an accurate forecast is issued. Therefore, this application of the CSI cannot reliably measure the skill involved in the forecast products, rendering the CSI and FAR statistics in Table 9 largely meaningless.

An improved CSI method of outlook verification that incorporates both the coverage (actual and forecast) and the areal distribution of reports within the outlook area into the FAR calculation is being developed at NSSFC by Weiss *et al.* (1980). Comparative verification of 1978 and 1979 data utilizing this new technique is currently being performed at NSSFC and the results will be reported in the future.

In an operational environment, the transformation of probabilistic forecasts of thunderstorms and severe local storms into their categorical counterparts is not an easy matter. Since June of 1978, NSSFC has outlined the TDL severe local storm forecast area by the 6% severe local storm conditional probability contour along with the 35% thunderstorm probability contour as suggested by Foster and Reap (1978). To define the density categories for the TDL forecasts, severe storm conditional probability thresholds provided by Reap (personal communication, 1978) have been applied (i.e., isolated $\geq$6%; few $\geq$15%; scattered $\geq$30%).

As indicated by Reap and Foster (1979), however, TDL has introduced a new method for determining a categorical severe local storm forecast. The authors derive "unconditional severe local storm probability forecasts by computing the product of the conditional severe local storm probabilities with the unconditional thunderstorm probabilities." Unfortunately, NSSFC was not aware of this newer definition and therefore did not apply it when attempting to use the automated product as guidance in real time.

Further, the authors "selected the unconditional severe local storm probability threshold that maximized one or more of verification scores. The threshold value selected was 4.0%." Examination of Reap and Foster's Fig. 8 reveals that *none* of the

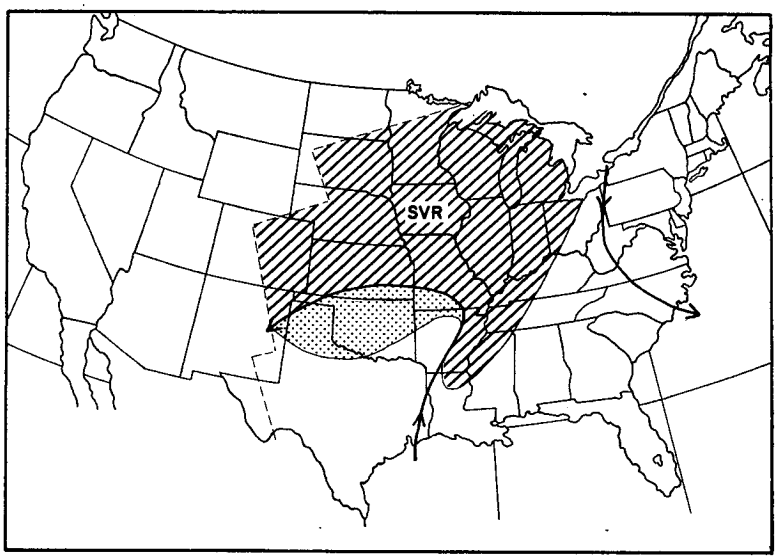


FIG. 4. Sample TDL thunderstorm and severe weather outlook on 25 June 1978. General thunderstorm forecast area is to the right of solid directed lines; hatched and stippled region denote forecast severe thunderstorm area. Stippled area denotes severe thunderstorm area outside of general thunderstorm area.

verification scores attain a maximum at 4%. In fact, it appears that the CSI maximizes at a threshold of around 6%. Additional explanation seems to be required to justify the selection of an unconditional severe local storm probability threshold value.

It is interesting to note that, when a 4% unconditional severe thunderstorm probability is used to delineate a severe thunderstorm forecast area and a 35% thunderstorm probability is used to outline a general thunderstorm area, one can conceivably forecast severe thunderstorms outside of the general thunderstorm area! Fig. 4 illustrates the TDL outlook for 25 June 1978 using the above criteria.

For operational use, the method for transforming the probability forecasts into categorical forecasts is apparently undergoing further development. According to TDL (1979), "tests indicate the threshold value should be raised (lowered) for those cases with high (low) probability values". Such a situation emphasizes the difficulty in applying large-scale model output and inadequate climatology to the prediction of mesoscale weather events. If variable thresholds are required, the probability numbers themselves must be subject to question. Thus, it appears difficult to justify the present statistical technique when the results cannot be interpreted in a consistent and meaningful fashion.

Since there are a number of ways to generate categorical forecasts of severe local storms from probabilistic forecasts, problems arise in interpreting the *significance* of various combinations of probability values. For example, should a thunderstorm probability of 80% coupled with a severe storm conditional probability of 5% be viewed any differently than a thunderstorm probability of 10% and a conditional severe probability of 40%, since they both yield unconditional severe storms probabilities of 4%? Or consider a 50% thunderstorm probability and an 8% conditional probability of severe. Is this latter combination more "significant" since both values exceed the previously defined lower limits of 35% thunderstorm probability and 6% conditional severe storm probability? These are questions that need to be addressed in order to make best use of the automated product.

## 4. Conclusions

In summary, the probability equations, in fact, do serve the operational forecaster by providing an additional data source. Our comments should not be construed as an effort to devalue the approach taken by Reap and Foster but, rather, to put it into the proper perspective. In doing so, we have raised a variety of questions which we feel need to be answered. These begin with several aspects of the predictand data. Problems with the predictand sample seriously affect the resulting probability equations, owing to the heavy reliance on modulated climatological frequencies in both sets of equations. The wide variety of large-scale environments capable of supporting severe thunderstorms, as well as unresolved mesoscale perturbations, seem directly responsible for the selection of climatology as the leading predictor.

Additional difficulties arise when statistical methods developed for point forecasts of routinely measured quantities are used without modification

for area forecasts of relatively rare events. Finally, the method used for converting the probabilistic product into categorical form has not been clarified, remaining the subject of continuing experimentation. Resolution of these crucial problems would result in an improved guidance product that will benefit operational meteorologists.

*Acknowledgments.* We wish to express our appreciation to Dr. Joseph Schaefer and Mr. Leslie Lemon of the Techniques Development Unit, NSSFC, for their helpful suggestions. Many valuable discussions with forecasters in the Severe Local Storms Unit, NSSFC, are also acknowledged. Finally, Mrs. Beverly Lambert's patience and typing skills have been invaluable.

## REFERENCES

Abbey, R. F., 1976: Risk probabilities associated with tornado wind speeds. *Proc. Symp. Tornadoes: Assessment of Knowledge and Implications for Man*, Lubbock, Texas Tech University, 177–236.

Asp, M. O., 1963: History of tornado observations and data sources. Key to Meteor. Records, Doc. No. 3.131, U.S. Weather Bureau, Washington, DC.

Court, A., 1970: Tornado incidence maps. ESSA Tech. Memo. ERLTM NSSL-47, National Severe Storms Laboratory, Norman, OK.

Donaldson, R. J., R. M. Dyer and M. J. Kraus, 1975: An objective evaluator of techniques for predicting severe weather events. *Preprints 9th Conf. Severe Local Storms*, Norman, Amer. Meteor. Soc., 321–326.

Foster, D. S., and R. M. Reap, 1978: Comparative verification of the operational 24-h convective outlooks with the objective severe local storm guidance based on model output statistics. TDL Office Note 78-7.

Galway, J. G., 1977: Some climatological aspects of tornado outbreaks. *Mon. Wea. Rev.*, **105**, 477–484.

Murphy, A. H., 1978: On the evaluation of point precipitation probability forecasts in terms of areal coverage. *Mon. Wea. Rev.*, **106**, 1680–1686.

McNulty, R. P., D. L. Kelly, and J. T. Schaefer, 1979: Frequency of tornado occurrence. *Preprints 11th Conf. Severe Local Storms*, Kansas City, Amer. Meteor. Soc., 222–226.

National Weather Service, 1979: *Operations Manual.* Chapter C-40, Severe Local Storm Warnings, 33 pp.

Reap, R. M., and D. S. Foster, 1979: Automated 12–36 hour probability forecasts of thunderstorms and severe local storms. *J. Appl. Meteor.*, **18**, 1304–1315.

Techniques Development Laboratory, 1979: Severe local storms prediction—medium range forecasting. *Mon. Prog. Rep.* (November), Silver Spring, p. 4.

Weiss, S. J., 1977: Objective verification of the severe weather outlook at the National Severe Storms Forecast Center. *Preprints 10th Conf. Severe Local Storms*, Omaha, Amer. Meteor. Soc., 395–402.

——, D. L. Kelly and J. T. Schaefer, 1980: New objective verification techniques at the National Severe Storms Forecast Center. *Preprints 8th Conf. Weather Forecasting and Analysis*, Denver, Amer. Meteor. Soc., 412–419.

Zurndorfer, E. A., 1980: A comparative evaluation of PE, LFM and probability of precipitation amount quantitive precipitation forecasts for the period 1975–1979. *Preprints 8th Conf. Weather Forecasting and Analysis*, Denver, Amer. Meteor. Soc., 19–22.