Preprints, 5th Australian Severe Thunderstorm Conference
Bureau of Meteorology, 29 July - 2 August, 1996
Avoca Beach, New South Wales, Australia

Verification of Forecasts of Convection: Uses, Abuses, and Requirements

Charles A. Doswell III
NOAA/ERL National Severe Storms Laboratory
Norman, Oklahoma (USA)

1. INTRODUCTION

This paper is focuses on verification of forecasts of convective phenomena, but many of the notions are tied to the more general question of forecast verification in general, irrespective of the forecast elements. Therefore, it will be necessary to set some context before dealing with the specific issues of convection and its associated weather events.

It is probably a safe bet that most forecasters are somewhat uncomfortable with forecast verification. Looking at how well the forecasts worked out can be an unsettling experience. We all feel frustrated at times with our inability to outthink the atmosphere, and it is not always a happy time to be confronted with the quantitative results of our efforts. Nevertheless, it is almost a platitude to say that a forecast not verified is a forecast not worth much. For instance, unverified forecasts might contain biases that would be easy to eliminate if they had been identified by verification. In these days when "quality assurance" and "total quality management" have become sacred phrases in the business world, can we escape the need to assess how well we are doing? There can be little doubt that this is a necessity, if we *care* about the quality of our output. If no one cares how good it is, then why bother issuing the product in the first place?

All right then, given that everyone agrees on the inevitability of verification, another issue becomes apparent fairly quickly. The notion of comparison rears its ugly head: one forecaster to another, one office to another, one nation to another, etc. Suddenly, verification has become politicized; an item of concern not for what it offers to us as professionals but because of what it implies about the "pecking order" within an office, or within a region, or within a nation. This means that the bureaucrats and stuffed shirts begin to take an interest in verification. These folks have their own agendas, mainly related to (a) not looking bad so that the next promotion is not threatened, (b) making everyone else look bad so that *their* next promotion is threatened. Moreover, within an office, verification can become a weapon wielded by a poor local manager, intoxicated with the power of position, to bash favorite targets and to bless the favorites. Forecasters may engage in pecking-order battles, using the numbers to humiliate their foes. There can be little doubt that all of these are unarguably negative by-products of the necessary task of verification, but they are real enough and can have the effect of making many forecasters dread the whole concept of verification.

Finally, consider the question of the numbers themselves. It has been said in many different ways that statistics lie. As pointed out in his delightful book, Hooke (1962, Preface) has clarified this in a compelling way:

> ...a person drawing inferences from data cannot choose between using statistics and not, as some seem to think. Such a person is engaged in statistics whether he likes it or not, and his only choice is between using good statistical procedures and using poor ones.

The analysis of data via statistical methods can be accomplished by a variety of tools, and a key point is to decide on what tools to use. Murphy (1993) has pointed out that what really matters in verification is *the joint distribution of the forecasts and the observations*. For a set of forecasts, this contains all the non-time dependent information (i.e., it does not consider how the forecasts varied from day to day, but rather considers them in the aggregate). This

usually takes the form of an m x n "contingency table" where m is the number of forecast categories and n is the number of observational categories. This is illustrated in Table 1, where the table elements form a matrix $c_{ij} = C$, with sums along the margins forming vectors $c_{\bullet i}$, and $c_{j\bullet}$. The total number of forecasts or observations is $c_{\bullet\bullet}$. If we are dealing with categorical (dichotomous) forecasts and events, the table is the familiar 2 x 2 version so commonly used in forecasting convection. If the forecasts are polychotomous (e.g., more than two probability categories), the events can still be dichotomous (as with PoP forecasts) or they also can be polychotomous [thus permitting a proper definition for the "fuzzy" severe event categories proposed by Alford et al. (1995)]. The development of this table is where verification issues truly begin.

Table 1. The joint distribution table of forecasts (f_i), i=1,2,...,m, and observations (x_j), j=1,2,...,n, also known as the forecast contingency table.

Observed	x_1	x_2	...	x_n	
Forecast					sum
f_1	c_{11}	c_{12}	...	c_{1n}	$c_{1\bullet}$
f_2	c_{21}	c_{22}	...	c_{2n}	$c_{2\bullet}$
.	.	.	.	.	.
.	.	.	.	.	.
f_m	c_{m1}	c_{m2}	...	c_{mn}	$c_{m\bullet}$
sum	$c_{\bullet 1}$	$c_{\bullet 2}$	...	$c_{\bullet n}$	$c_{\bullet\bullet}$

2. STATISTICS

So far, so good. It is from this simple table that controversy flows, however; the controversy usually swirls about what *measures* to develop from the table. Murphy and Winkler (1987) have argued persuasively that a *measures*-oriented verification is not as useful as a *distributions*-oriented verification. Brooks and Doswell (1996) have provided an example of how much more insightful it is to focus on the distributions rather than the measures. I am going to take that tack, no doubt to no one's surprise! Let me try to provide an analogy that is somewhat artificial but which illustrates the problem with a measures-oriented approach.

Imagine a professional baseball team, the Oklahoma Twisters. Their management has concluded they must improve their performance for the following year and have mandated they must trade one of two players, because of their high contract salaries. "Big Stick" Ozzie is a power hitter and bats 4th in the lineup with typical annual performance numbers of: 100 runs batted in, 40 homeruns, scores 60 runs, and has a 0.235 lifetime batting average. "Disco" Ian is a contact hitter who bats 1st in the lineup with typical annual numbers of: 60 runs batted in, 10 homeruns, scores 100 runs, and has a 0.335 lifetime batting average.

This is a classic example in sports of how statistical measures of the "value" of a team member can create problems. If the measure chosen weights power hitting over contact hitting, then "Disco" Ian gets the axe, and is traded to Far Rockaway, New Jersey. If the measure chosen weights batting average over homeruns, then "Big Stick" Ozzie is packing his bags for Minot, North Dakota. How easy is it to measure the "value" of these two players to the team? Each contributes in his own way (I am ignoring their defensive performances) and the team as a whole needs *both* of the abilities these players have.

Ideally, on a team with both these players, you'd like to have "Big Stick" improve his batting average a bit without sacrificing his power. At the same time, "Disco" could help the team by hitting another few home runs without sacrificing his batting average. It would be unrealistic to expect *both* of them to be hitting 40 home runs and hitting 0.335, each with 100+ runs batted in and

100+ runs scored. A realistic goal is for them to make incremental improvements in their weak points and to maintain their strengths at the same time. Each team member is valuable to the team in different ways, but that doesn't mean they can't become even more valuable.

This somewhat contrived and perhaps silly example does contain most of the essence of the issues surrounding verification. Any performance measure that is biased toward a particular aspect of performance within the verification matrix is necessarily a simplification of the reality, perhaps even an oversimplification. There can be no single number that contains all the information about performance that might be necessary to make improvements. Moreover, each forecaster on the staff contributes to the total forecast office performance in different ways. If a particular forecaster's unique contribution is not properly assessed by the chosen measure, are we to conclude that this forecaster offers nothing of value to the team?

My contacts with forecasters suggest that one of the reasons for hating verification is that they are afraid of how their value is going to be measured. Choosing the wrong measure means that individuals might end up low in the "pecking order." If the purpose of the verification is *not* the ranking of individuals, but to seek methods for improving the forecasts, then single measures clearly are insufficient. Although single measures seem to make comparative verification possible, it is also worth noting that comparative verification *increases the dimensionality* of the problem (see, e.g., Murphy 1991). The increase in dimensionality is associated with the fact that, in general, the forecasts are made under different circumstances (except perhaps in special experiments where controls are imposed to make the situations comparable). If it is going to be both useful and meaningful to compare forecast performance, all the extraneous factors need to be deconvolved from the performance matrices. It is this deconvolution of effects that makes comparative verification more difficult. This is not easily done and often is not even attempted, either out of ignorance or for reasons of economy, leading (quite properly) to a perception that the comparison is unfair.

3. NEEDS

If we adopt the viewpoint that verification ought to be designed to assist the forecasters in improving the quality of what they produce, then what do we really need? I believe there are several constraints that must be rigidly enforced if a verification effort is to have a "payoff" in terms of measurably improving forecasts.

1. Accurate information about the predictand must be available. This seems so obvious it is almost em-barrassing to have to say it. Nevertheless, it appears that this *must* be said, because it is clear that many of today's forecasts involve events for which we have no consistently reliable validating observa-tions. This is particularly true for severe convection, even in the relatively data-rich United States The severe convective storm report-sparse Australian continent makes verification a dubious business right at the outset. *Without reasonably accurate obser-vational information, there can be* ***no hope*** *of forecast improvement through verifica-tion.* Period. If we insist on forecasting for data-void areas, then we are simply casting our seed to the wind and there can be little or no point to doing serious, rigorous verification in such areas. It is pointless, in my opinion, to *forecast* anything that we do not observe, owing to the impossibility of *verifying* such forecasts.

Consider the following from my experience in verifying severe local storm watches. As noted in Doswell et al. (1990), there is a high linear correlation between watches and severe weather (Fig. 1).

At face value, this seems to say something quite favorable: verification is good where severe weather occurs. However, it also suggests that the verification is dominated by the data set being used to validate the forecasts.

In considering the evolution of the skill (as measured by the Heidke score), in another paper (Doswell et al. 1993), we observed that the skill level as a function of time was dominated by the "inflation" of

severe weather reporting in the U.S. (see Fig. 2).

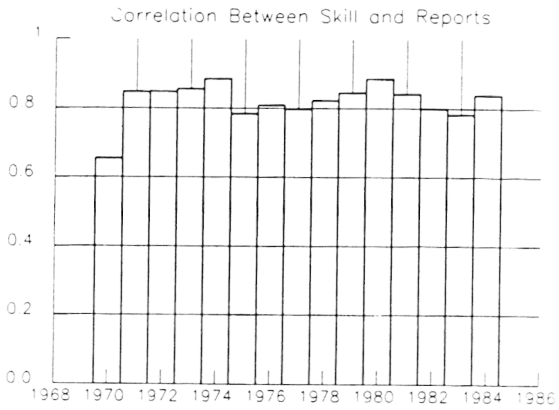

Fig 1. Linear correlation between observed severe weather reports and Heidke skill score for tornado or severe thunderstorm watch (from Doswell et al. 1990).

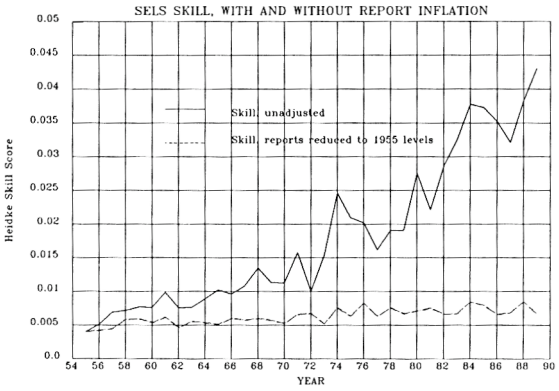

Fig. 2. Effect of compensation for "report inflation" on the Skill Score, where the verifying reports are adjusted to a common level (from Doswell et al. 1993)

The conclusion I have drawn from my study of long-term verification of convective weather events is that the verification exercise is dominated by the quirks in the verifying data set.

Thus, an apparently insurmountable problem has arisen. Without reasonable data, a meaningful verification is not pos-sible. But without verification, the whole point of forecasting in the first place is questionable. Perhaps the only solution to this dilemma is to make convective forecasts that include remote areas with little or no hope of reliable reports, but then one should only do serious, quantitative verification in those areas where one has confidence that if a severe event happens (or doesn't happen), it will be reported properly. That is, the quality of forecasting is only measured where it is plausible to do so, and the selected "index areas" become the basis for assessing the overall success of the forecasting. Doing a verification in areas of sparse and/or unreliable data can be done, but the numbers should not be used to guide a program of systematic forecast improvement.

2. The forecasts and the observations need to be matched care-fully.. Presumably, no one would attempt to use rainfall measurements to verify tem-perature forecasts. But if there are no veri-fication data for severe convective events (hail, convective wind gusts, tornadoes, heavy precipitation), it is tempting to use *proxies* for those events. For example, consider radar data, satellite data, lightning data, CG lightning flash data, etc.; all represent remotely sensed information that is tempting to use in lieu of the actual observed presence of some convective event. This is a practice to be avoided, in my opinion *unless* it is also decided a priori that what one is forecasting is *not* a convective weather event but some proxy such as reflectivity morphology, or Doppler radar signatures, or CG lightning wave forms, or whatever. There are simply too many dubious assumptions when there is a mismatch between the forecast and the verifying event.

3. The verification process should be multidimensional, rather than based on a single measure. I already have made this argument above. For an m x n contingency table, the dimensionality is $mn-1$; hence, for the 2x2 table, there are three independent numbers needed to describe the information con-tained in the table. Presumably, the task of putting together the forecasts and the observed events into a database can be done locally or at a central site. It is worth noting that a multifaceted forecast verification system is much harder for forecasters to

"play" in order to maximize their scores. Note that strictly proper scoring rules are only possible for probabilistic forecasts (A. Murphy, personal communication) In general, a forecaster's best strategy in a multidimensional verification is to put out the best forecasts possible, and then look at the verification as a process to improve on those forecasts. This means that:

4. The results of the verification need to be made available to forecasters, rapidly and regularly. If the purpose of this process is truly to aid forecasters, then they should be the primary recipients of the information. It definitely should not sit on some bureaucrat's computer until performance rating time! The forecasters need *feedback*, they need it *quickly*, and in a form that makes clear what are their weak and strong points within the panoply of aspects contained in the joint distribution matrix of forecasts and observations. Any other dissemination of the verification results is at best questionable and is certainly full of potential for abuse by bureaucrats and others. Any other use of the results is a luxury and may be a potential abuse of the data. In this day of powerful workstations, there is no excuse for not providing diagnostic verification results to forecasters. It simply requires a commitment to improving forecast quality through verification. Incidentally, this means that forecasters must be educated and trained how to interpret properly the output of the verification program.

5. Based on the results of the verification, there must be follow-up research. I call this follow-up research "closing the loop" and it is appalling to me how little of this is ever done. To me this indicates the uselessness of most of the verification programs in the terms I have suggested are the *only* reasons for doing verification: improving the forecasts! Suppose we have a forecaster, Paul Q. Vorticity, who has shown that he is quite adept at synoptic-scale forecasting and yet is doing quite badly at mesoscale problems. Forecaster Mary N. Brunt-Väisälä, on the other hand, does exceptionally well at the task of mesoscale forecasting. Is it a radical idea to suggest that Paul needs to find out what Mary is doing so that he can get his performance improved? Shouldn't Mary try to enhance her synoptic performance by learning from Paul? If the office as a whole has done badly in weakly baroclinic situations, don't we need to look at those situations and try to find out if there is something we can do collectively to improve our performance? Suppose at our forward-thinking office, we have looked at those specific situations and decided we don't know how we might have done better than we did. Then it might be time to call Dr. Helmholtz Theorem at nearby Goshwhatta University, since these weakly baroclinic events are his research stock-in-trade.

My point is that verification can do much more than offer insight into forecasting performance. If we choose to make the effort at "loop-closing," it can point out in *which* situations we have problems, and provide dates and times when we did poorly and even when we did well in those situations. and so can help to direct a *systematic* process aimed at forecast improvement. Most forecast post-mortems are chosen in an ad hoc fashion, rather than on the basis of what is most important for that forecaster's performance, or what the office as a whole needs to be studying. Closing the loop is the most neglected aspect of forecast verification, in my opinion.

4. DISCUSSION

It is probably quite foolish to expect that forecasting ever can be separated from politics. I believe, however, that forecasters need not be content with what "The System" provides for them vis-á-vis forecast verification. If the coneheads in administration insist on using some silly verification scheme that suits their agenda, fine. Let it go, but develop your own for your own use and information. If you can admit that a careful look at your performance is, indeed, worthwhile because you care about that performance, then there is no reason not to develop a system that serves *your* needs and let the bureaucrats do whatever they want. The access to data in a modern forecasting office, as well as the proliferation of personal computers, means you can do something useful with the available in-

formation no matter what "The System" chooses to do.

What I am trying to suggest is that forecasters should not leave this very important aspect of their job at the whims of others who do not share *their* agenda: improvement of the forecasts. Of course, for those whose view of the forecasting job is that of a "9 to 5" clock-punching exercise, I'm probably not going to engender a great deal of enthusiasm for extra work. But assuming that your forecast performance matters to you, I am trying to make it clear how to make verification work *for* you, not against you. In the future, I'd like to see more forecasters interested in verification. Yes, I know it's tedious and hard work, but the payoffs are there, if you're willing to invest the effort. I hope it goes without saying (but I am saying it anyway!) that verification should be designed so that it can't be "played" - i.e., forecasting to get the best scores rather than to put out the best forecast. Am I being naive to hope that even if your verification system is not "strictly proper" and can be "played," you will resist that temptation and continue to do the best job you can?

If you're an administrator and you've been squirming and angry with what I have said because it doesn't fit you, calm down! If the shoe doesn't fit, you don't *have* to wear it. A tip of the cowboy hat to you if you have continued to pursue the interests of bench forecasters in your administrative role! You are in a position, then, to make verification something other than a hated, neglected concept. If you still care about the quality of the product that goes out under your management, it behooves you to get on the ball and support a *meaningful* verification program, not a hollow exercise, in your office, region, or country.

Acknowledgments I appreciate the helpful comments offered by Dr. Allan Murphy on an earlier version of the manuscript. I also have benefited enormously from many stimulating discussions with my colleague, Dr. Harold Brooks.

REFERENCES

Alford, P., C. Ryan and J. Gill, 1995: Thunderstorms and severe thunderstorms: A forecasting perspective (3rd Ed.). Bureau of Meteorology, Melbourne, p. 14.3.

Brooks, H.E., and C.A. Doswell III (1996): A comparison between measures-oriented and distributions-oriented verification methods in forecast verification. *Wea. Forecasting*, **11**, (in press).

Doswell, C.A. III, D.L. Keller and S.J. Weiss, 1990: An analysis of the temporal and spatial variation of tornado and severe thunderstorm watch verification. Preprints, *16th Conf. Severe Local Storms* (Kananaskis Park, Alberta, Canada), Amer. Meteor. Soc., 294-299.

_____, S.J. Weiss, and R.H. Johns, 1993: Tornado forecasting: A review. *The Tornado, Its Structure, Dynamics, Prediction, and Hazards* (Church et al., Eds.), Geophys. Mongor. 79, Amer. Geophys. Union, 557-571.

Hooke, R., 1962: *Introduction to Scientific Inference*. Holden-Day, San Francisco, 101 pp.

Murphy, A. H., 1991: Forecast verification: Its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590-1601.

_____, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281-293.

_____, and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338