NOAA Technical Memorandum
NWS HYDRO 43

A CATEGORICAL, EVENT ORIENTED, FLOOD FORECAST VERIFICATION SYSTEM FOR NATIONAL WEATHER SERVICE HYDROLOGY

Office of Hydrology
Silver Spring, Md.
June 1988

U.S. DEPARTMENT OF COMMERCE / National Oceanic and Atmospheric Administration / National Weather Service

NOAA Technical Memorandum
NWS HYDRO 43

A CATEGORICAL, EVENT ORIENTED, FLOOD FORECAST VERIFICATION SYSTEM FOR NATIONAL WEATHER SERVICE HYDROLOGY

David G. Morris

Office of Hydrology
Silver Spring, Md.
June 1988

UNITED STATES
DEPARTMENT OF COMMERCE
C. William Verity

National Oceanic and
Atmospheric Administration
William E. Evans

National Weather Service
Elbert W. Friday, Jr.
Assistant Administrator

A CATEGORICAL, EVENT ORIENTED, FLOOD FORECAST VERIFICATION SYSTEM FOR NATIONAL WEATHER SERVICE HYDROLOGY

DAVID G. MORRIS
Southern Region
National Weather Service
West Gulf River Forecast Center
Fort Worth, Texas

June 1988

TABLE OF CONTENTS

A CATEGORICAL, EVENT ORIENTED, FLOOD FORECAST VERIFICATION SYSTEM FOR NATIONAL WEATHER SERVICE HYDROLOGY

David G. Morris

ABSTRACT. The National Weather Service River Forecast Centers deliver most of the flood prediction service in the United States, chiefly for the larger streams that are sufficiently networked with river and rain gages to permit the forecast of river rise well in advance of crest. However, a system of compiling information over time that permits the Agency to judge and track its flood prediction capability has never been established due to complexity of the verification problem. In particular, the historic reliance on observed versus forecast crest difference as the primary means of determining flood forecast accuracy has prevented agreement within the profession as to the best method of compiling verification data.

As a possible solution, a different approach is suggested whereby a river rise is viewed as an event, or series of events, each of which is categorized according to magnitude. The event is classified by stage, on a scale of one to six, ranging from no flood to record major flood. An event may pass through more than one classification before recession takes place. If the event is not correctly predicted, an error, measured in feet, is computed. It is an event error, not a stage or crest error. The site-specific flash flood event is similarly analyzed, but on a scale of one to four. The resulting data obtained over time permit a variety of simple statistics to be computed that should prove highly useful in evaluating flood prediction capability. Two examples of such statistics would be BIAS by Category (under or overforecasting an event), and False Alarm Ratio (the fraction of forecast threat events that did not verify). Also available would be a comprehensive tabulation of forecast and observed lead times for each event, number of rises nationwide in each category, average prediction error by category, and a fair comparison of demonstrated forecast skill.

I. INTRODUCTION

For more than a quarter of a century the National Weather Service (NWS) has provided the public with a flood forecast service, chiefly out of the Nation's River Forecast Centers, but has done so with limited means of compiling information over time that judges in some manner the Agency's flood prediction capability. Hydrologists of every stripe, and some of the best minds in the business, have debated verification from every angle, so it seems, and have not settled on a method that is acceptable to the profession. It is my contention that this failure is not due simply to the complexity of the verification problem, but rather is also a result of our collective failure to view flood phenomena in a perspective that lends itself to solution to the verification problem. In short, the historic insistence on the part of the river forecaster to include observed crest vs forecast crest, or observed stage vs forecast stage, as the sole basis for prediction accuracy, plus a warning lead time measure, creates more of a problem than it solves, thus resulting in our current state of no national flood prediction verification program. It also rivets our attention to one, and only one, aspect of a river rise, and prevents our "seeing" the flood in proper public perspective, that being an event of some consequence.

The verification method presented herein is simple and straightforward, is patterned after meteorological reasoning in verification, and, I believe, meets the Agency's need to compile data that answers the question "how did we do" in flood forecasting, whether that be a question addressed to a single site along a river, or for all streams in the United States for which verification is practical. It is a method that also tallies the number of floods and the magnitude of each flood, the latter being terribly significant to the public, but never explicitly addressed by other verification means, at least those brought to my attention. In this report the term "verification" is restricted to mean simply the comparison of the category or magnitude of flood predicted versus that category or magnitude of flood observed. By "evaluating the service or capability", in reference to categorical verification, we are only referring to resulting summary information on verificaton (data, tabulation, statistics), not value to the user of the flood forecasts. However, before detailing this different approach to flood verification, I believe it necessary to discuss other river forecast verification efforts, current and past, in order to demonstrate the need to adopt something new for assessing the

strengths and weaknesses in our flood prediction program.

2. VERIFICATION METHODS

2.1 TWENTY-FOUR (24) HOUR STAGE FORECASTS

Verifying twenty-four (24) hour stage forecasts presents little information to assess the strengths and weaknesses in flood prediction. Over a period of time, 24-hour data reflects mostly river recession, punctuated by rises (often unexpected), and, of course, the error typically increases whenever runoff occurs, as forecasting becomes more uncertain whenever it rains, with the error generally increasing with the frequency and magnitude of the rises. If a river forecast center (RFC) has a very low error for a year or so, it was likely because nothing hydrological was happening. Some offices may find a selected 24-hour stage verification useful for assessing how well we do in stage prediction (say, for navigation interests). If so, the office is free to collect and conclude whatever it wishes. However, to simply collect 24-hour stage prediction data blindly for 365 days each year is to assemble, I think, a data set that cannot address flood forecasting skill.

2.2 FORECAST CREST VS OBSERVED CREST

Hydrologists invariably strongly disagree when faced with the prospect of such information being assembled and used to assess forecast performance. And, worse yet, used to compare RFCs, which would be done by someone, no matter what anyone says to the contrary, just because the information was there. Hydrologists know that some rivers are "well behaved" (easy to forecast), and other rivers are not. Hydrologists know that some rises are complex, requiring substantial skill or subsantial luck, or a combination of both, while other rises are comparatively simple to deal with. Hydrologists know that, from either a technical achievement viewpoint or historical viewpoint, some floods are far more significant than others. Hydrologists know that the best effort in warning, coupled with the best public response, may not be reflected fairly by the final difference between observed and forecast crest values. In short, this difference, this number, cannot be used to judge prediction error with consistently interpretable meaning that is satisfactory to the RFC, the Agency, and the public. So no one is really comfortable with it. If we remove said emotions from our argument, the following logic still leads one away from serious consideration of crest numbers as a primary basis for verification.

1. A perfect forecast - the observed crest occurs close in time to forecast crest, and the crest values are the same, or within a few inches. Now all we have to do is always issue a perfect forecast, and there is no verification problem. But we can't do this, so the problem remains.

2. A less than perfect forecast. Even without worrying about flood warning lead time (subject addressed later on in this report), we find ourselves in an increasingly difficult "scoring" predicament as the observed moves away from forecast. Observed vs forecast crest or stage numbers, by themselves, cannot answer these kinds of questions:

a. Is it a flood or not?

b. If a flood, is it a serious flood?

c. How about catastrophic or record flood?

d. What is the significance of a one foot miss? How about five feet?

e. If a five foot error is "good forecasting" well upstream on the, say, Mississippi, is it also a "hit" of sorts at New Orleans? Why not? Who said a five foot miss was a good job anywhere under any circumstances?

So then, why even consider the numbers? Answer: Because hydrologists judge the accuracy of <u>their</u> prediction by that number. It may technically point to certain problems in the river forecast procedure that are correctable; he knows that a one foot miss is better than two, and what to reasonably expect out of the office development work that supports the forecast effort; his experience dictates it was excellent (maybe) forecasting under the circumstances, but the circumstances generally change, and all rises are different; hard data varies from storm to storm, etc., etc., etc. That forecast vs observed value makes him feel good or not so good, and it may subjectively, if not objectively, point the way to a better job on the next rise. But it is not likely meaningful information to the outsider.

Now, we could at this point discuss how it is that stage/flow and forecast error at one point along a river is not translatable to another site or another river for comparative purposes in skill scoring. But it would be lengthy and not add to this report. Suffice it to say, observed crest vs forecast crest are data that have very limited value, I believe, in

statistical analysis designed to assess how well we predict flood.

2.3 MEAN FORECAST LEAD TIME (MFLT)

Developed by Sittner (1977) while employed as a research hydrologist in the Office of Hydrology (O/H), MFLT was designed to "measure the overall effectiveness of the flood forecasting program" delivered by NWS. MFLT is a well reasoned, logical approach to evaluating the service, and is a credit to the creative thinking of its author. However, for a variety of reasons, MFLT died a slow death. MFLT deserved better. It is not the intent here to resurrect MFLT, as I am going to suggest a different approach to the problem of "monitoring the forecast service as a whole," but we need to understand what it is, as it is part of our verification history, and appreciate its merit or demerit, depending upon how one wishes to view the idea.

To quote Walt Sittner, "MFLT is, in essence, the average warning time provided by a group of forecasts, suitably adjusted for forecast inaccuracy. The basis for the adjustment is the effect of the inaccuracy on the user." MFLT is defined as "the average warning time that would be provided by a group of error-free forecasts that would have affected the users in the same manner as did the group of forecasts actually issued." If, for example, three stage forecasts are issued on a flood rise, and the time from issue to the time of forecast stage occurrence is 17, 20, and 23 hours respectively, MFLT computes (17+20+23)/3=20, so 20 hours is the verification score for this event. If, however, there are timing errors in the forecasts - the stage occurs earlier or later than predicted - Walt designed a simple computational adjustment to lower the MFLT score in some proportion to the error. When an event involves the issuance of a number of forecasts, one of them, likely the last, should correctly predict the observed crest. If not, the verification formula imposes a "penalty" that adjusts the score. This penalty objectively distinguishes a "hit" or "miss" forecast based upon a bracket; the bracket being an allowable range of error that is reasonable for the forecast point, if a bracket was not explicitly stated in the prediction itself. Walt's verification further consists of rules in MFLT computation that cover all possible combinations of rises and types of forecasts. Mr. Sittner's documentation should be consulted for detailed explanation of MFLT.

Why not MFLT in 1988? Well, I have a problem with MFLT as the sole basis for a national flood verification system. First

of all, and by far the biggest drawback to MFLT, is that it combines forecast lead time with a crest hit or miss into one number, a value in hours . It does not have a communicable meaning to anyone not intimately familiar with MFLT. Situation: We have a flood, and are asked "How did you do?". Answer: "Well, real good, our MFLT was 18.6 hours", or, ".....not so good, our MFLT was almost zero." Reply, "What?" Verification is always "how did we do," and in flood forecasting, how we did has three fundamental aspects: (1) Magnitude of the flood, (2) warning or lead time, at least roughly, and (3) how close, stage-wise, the flood was forecast. Number one is necessary to put the event in proper perspective. All three require separate answers, and there is no way to combine those answers into one and clearly communicate. While it is true that compiling MFLT numbers over years will indicate a trend in the service at the forecast site in question, and do so nicely, items one, two, and three referenced above remain largely masked. Ask the meteorologist how he did in forecasting a heavy snow event, and you will not get an MFLT type answer.

A second facet of MFLT that bothers me is the fact that a rationale is used based on users response which, for the forecast service monitoring (verification) goals set forth by me later in his report, is not always appropriate. It is assumed in MFLT that precautionary measures are taken by individuals in some stepped fashion up the rising limb of the forecast hydrograph, based on whether or not one is "above or below" the stage prediction, with precautions varying according to stage above some threshold level and the amount of time available. More simply stated, having learned one is going to get wet, what one decides to do, or can do, is based largely on the forecast lead time. The implication is, of course, that a longer lead time equates to a better (more effective) service. Well, all this may be largely true, and ideally is true, but I suspect is often not true. Also, it is possible, under some circumstances, that there be considerable (maybe equal) benefit to a flood warning with zero, or near zero, lead time. This is recognized by the meteorolgist in the flash flood forecast/warning service provided by the Agency: ".....warnings with zero lead time are legitimate and may in fact provide a nearby community with enough time to take effective action" (Campbell, 1985). For our purposes, explained shortly, user value, user response, is not germane to verification, if, by verification, we mean "how did we do", and we do mean that. It is simply a question of what was forecast vs what was observed. Customer satisfaction, social benefit, resulting from a flood warning is a facet of the service that should be evaluated separately. We will just assume our service has value, and get on with designing a

verification procedure.

None of the arguments presented are intended as criticism of MFLT. MFLT accomplishes what the author intended. It is an "engineering solution," and a very good one at that. If the purpose is to evaluate, in terms of value to the user, a series of forecasts for a single flood event, then MFLT is the way to go. It does measure forecast effectiveness in a manner that is physically meaningful and that is closely related to economic benefit derived from a flood prediction service. MFLT, in my opinion, would be an excellent "companion" to the categorical verification system presented herein. Each approach answers different, but equally important (so I think), questions regarding our ability to predict flood. My differences are solely in the area of verification philosophy and what I perceive to be verification goals of the National Weather Service. So where do we go now?

3. CATEGORICAL, EVENT ORIENTED, FLOOD FORECAST VERIFICATION

It has been said that "when all was said and done, more was said than done." This section, hopefully, does something. What follows is simple reasoning that I hope is acceptable to my colleagues, and that I hope leads to a verification program acceptable to the Agency. My primary interest is not in numbers that lead to conclusions that lead to better forecast procedures - local study can accomplish that, although, hopefully, a verification system is also an aid in that regard. My primary interest is not in how well the public likes the product, although this is terribly important, as I have seen poor forecasts well received and good forecasts poorly received. I have no interest in verification that attempts to take into account the fact some rivers require more skill or luck than others to forecast, just because every hydrologist, like every meteorologist, thinks his turf is toughest. Besides, I don't know how to do it. And, finally, I have no interest in attempting to create a verification system that answers in one big swoop every question we would like to have answered via verification, because (a) I do not know how to do that either, and (b) that is not a worthy goal, in my opinion, as it is surely true such an effort would lead to a verification algorithm terribly complex and data intensive, and operational hydrologists are not inclined to accept such a mix. Also, it is possible that many in the Agency would not be sure what all the resulting statistics meant, anyways. I've noticed over the

years that the creators of complex statistical formulations are often the only ones positive about what the formulation does. Everyone else gets to "interpret" the results, hoping the interpretation is correct. While this is not necessarily bad, I believe it can be avoided in flood verification.

Here is the interest, and the questions the Agency, first and foremost, wants answered:

1. Did you have a flood? - Yes, no.

2. If yes, how "big" a flood?

3. Did you forecast the flood? - Yes, no.

4. If no, how much did you miss it?

5. If yes, how far in advance was the public warned of the event?

Now, once those questions are answered, a variety of statistics may be employed to summarize and draw conclusions, that we all understand. How well it was possible to do, what the public did, what is reasonable error, or anything else that pops into mind is not relevant to items 1 through 5. Administrative investigation or other information may be employed to answer whatever, on a case by case basis, but keep it out of flood verification. The objective (goal) of this report is to answer the above five questions, and do so in a manner that is communicatable to anyone with some familiarity with flood, and a knowledge of the attempt by government and non-government services to predict the degree of flood. By addressing these questions categorically one does, in fact, establish a framework for evaluating the flood prediction capability rendered by anyone.

3.1 THE FLOOD EVENT

The flood is an event; it is a weather event, or, if you prefer, the result of weather events. The flash flood is a different event. The flash flood will be treated separately in this report, but prior discussion of flood does largely apply. Snowfall is an event; rain is an event. The flood and the weather all have one thing in common: They are events with duration, and they occur in degrees. Stated differently, meteorological and hydrological events have time and magnitude. Both may be classified or categorized, and are. The

meterologist finds it necessary and convenient to verify weather events according to a classification (snow vs heavy snow, wind vs high wind, etc.). The meteorologist forecasts a blizzard, and he has a definition for it. The meteorologist forecasts a dust storm, and he has a definition for it. And he verifies by the definition. The meteorologist, for years, has had a classification for rain events (light, moderate, etc.) that would serve nicely should verification be desired. Tornadoes are classified, hurricanes are classified, etc., etc. A flood is not something that needs to be treated differently. It is just another event. We have rain and no rain; we have flood and no flood. Events and non events, both. Like weather, flood can change classification rapidly or gradually. Flood gives way to major flood just as snow gives way to blizzard. When a hydrologist issues a crest forecast, and in some cases just a stage forecast during river rise, he is forecasting an event, plain and simple. And it has significance as an event. It has public significance, historical significance, Agency significance, and, also, verification significance.

In this report we later use the term verification event, meaning a flood of specified degree. A river rise, an event in itself, may very well pass through several levels of flooding, and may also display more than one crest before a recession takes place to normal levels. For verification purposes, such a storm may be viewed as more than one event.

3.2 THE FLOOD CLASSIFICATION

"Significant weather elements are most naturally verified as categorical events; contingency tables and the various scores derivable from them are the natural means of accumulating the verification information" (NWS, 1982). So says the Agency. No reason to exclude floods, that I can see. In 1983 I drafted some preliminary thoughts on categorical flood verification, and it is my intent to now expand on those thoughts (Morris, 1983). But first, I think it advisable to address certain opinions expressed on flood classification, either to me, or in my presence, although not necessarily in the context of verification. "One man's minor flood is another man's major flood." Or, how about, "there's no such thing as a minor flood when you're knee deep in water." Conclusion: There appears to be an aversion in some quarters of the river forecasting community to flood classificaton. It's the same as if the meteorologist refused, and he doesn't, to say "partly cloudy" because what is "partly" to one person is not to another; it's the same as refusing to use the term "fair weather" because the perception of fair varies from person to person; it's the same

as refusing to state temperatures as low because a segment of the public wears shortsleeve shirts in 40 degree, 20 mph wind; it's the same as refusing to call for light rain because one man's "light" is another man's "moderate"; and, after all, we know that a brief heavy shower somewhere, soaking some poor devil, sans umbrella, would make us look like fools. So hydrologists have been evasive when it comes to "putting a stamp" on their product. It is both the privilege and the obligation of the National Weather Service to define weather events, including hydrologic events, and it is long overdue, I think, for this Agency to quantify and standardize a flood prediction terminology. It may not be necessary to do so in communicating with the public, but in many (most?) cases it is likely advisable, and for the purposes of Agency internal communication, including verification, it is mandatory. Here's another sentiment: "I once stated in a public forecast that the river was going into major flood (and it did), and then got burned by the press for a bad crest forecast." Conclusion: There "ain't" no substitute for a real good crest forecast; a categorical statement as to the expected magnitude of flooding is not sufficient in itself. The concluson is off-target. As alluded to earlier in this report, what the public thinks of service rendered, while most important, is not what we are trying to accomplish with the verification goals set forth in this report. Also, nothing has been said, nor will be said, to imply that a definition or classification of anything is a substitute for good crest forecasts, temperature forecasts, wind forecasts, or any other kind of forecast. No matter what form the forecast service takes, no matter how good or bad the product may prove to be, the Agency is subject to boos and cheers. In fact, the lack of national definition in its flood prediction service leaves the Agency all the more vulnerable to criticism.

Table 1 is a listing of suggested flood categories, on a scale of 1 to 6, ranging from no flood to record flood. The flood definitions should be largely acceptable to the profession. Source document is an Eastern Region ROL, dated June 25, 1976, with slight modification to Category 6 by me. It is my intent to build a simple verification plan around these categories. Any point along a river, for which NWS issues a forecast, may be broken down into these degrees of flooding, if one chooses. We choose to. This is not to say that every category must be used - for some locations one may decide that the term "minor flood" or "moderate flood" has no basis or acceptance, but common sense dictates that everyone everywhere recognizes no flood, a major flood, and a record flood, and surely most everywhere there is such a thing as minor to

moderate flooding. And if the site has no established record flood, then obviously there may be a major flood that someday establishes one. Until then, a major flood is "as high as you can get," unless one wishes to consider the probable maximum flood (PMF) as an upper limit (stage) to some additonal category of flood whereby we might obtain a "measure" of our ability to predict very rare events.The PMF, if used to establish an extreme flood category, must be estimated from probable maximum precipitation (PMP) by hydrologic techniques (Linsley, 1975). If someone somewhere argues that a particuar river point defies any classfication, then I would say, fine, drop it from verification consideration, and then I would ask why the forecast is being issued. For the purposes of verifying public flood forecasts, and that is what we are trying to do, I have a difficult time imagining a forecast with any real meaning that does not fit a said category. Suffice it then to say, that, with relatively few exceptions, all six categories apply to a forecast point, and may be determined for that point. The more categories used, the better the definition of the event. Weather Service Form E-19, a multi-page document that describes a forecast site in just about every way possible, and that supposedly exists for all our "significant" forecast points, should or could contain much of the information necessary to establish flood categories at the gage in question. If not, a Service Hydrologist or RFC hydrologist must gather the information, based on a site visit/survey, or the flood levels may be arbitrarily established based on topographic charts and some common sense. One could argue that all this should be done even if verification was not the goal. Should the E-19 not exist, the information is likely available in some form elsewhere. If the forecast point is new, then flood categories should be established as preliminary or tentative, and then changed as necessary in future years. The proposed verification plan allows one to change stage assignments for flood categories at any time with no loss in verification history or degradation of the verification data base. We don't care, for example, that a major flood at Doonesville, USA, is realized at a lower stage today than yesterday because of flood plain encroachment. What we care about is that, for example, a major flood, in terms of general public impact, occurred, and what the agency forecast was, in those same terms, for the purpose of verifying the stage prediction. Categories 5 and 6, of course, could also change as years go by, and, again, this would be of no consequence to the verification system. It should be emphasized at this point that our concern for "public impact", in terms of verification, is so far confined to the defined levels/degree of flood so stated in Table 1. This is understandable and reasonable considering what we are trying to verify, which is our ability to predict certain

TABLE 1. CATEGORICAL FLOOD EVENTS

CAT	DEGREE	GENERAL DESCRIPTION
1	NO FLOODING	NO FLOODING EXPECTED
2	MINOR FLOODING	NEAR FLOOD STAGE - ONLY MINIMAL DAMAGE EXPECTED
3	MODERATE FLOODING	SECONDARY ROADS BLOCKED-TRANSFER TO HIGHER ELEVATIONS NECESSARY TO SAVE PROPERTY-SOME EVACUATIONS MAY BE REQUIRED
4	MAJOR FLOODING	EXTENSIVE INUNDATION AND DAMAGE-MANY PRIMARY ROADS AND BRIDGES CLOSED-MANY PEOPLE MAY BE EVACUATED
5	NEAR RECORD FLOODING	MAJOR FLOODING WHICH IS EXPECTED TO APPROACH THE RECORD FLOOD
6	RECORD FLOODING	MAJOR FLOODING WHICH IS EXPECTED TO EQUAL OR EXCEED FLOOD OF RECORD

TABLE 2. EXAMPLE OF STAGE ASSIGNMENTS FOR FLOOD CATEGORIES, THE TRINITY RIVER AT DALLAS

STAGE (FT)		FLOOD EVENT	FLOOD CATEGORY	STAGE RANGE
		NO FLOOD	1	00.0-29.9
30	FS	MINOR FLOOD	2	30.0-31.9
32		MODERATE FLOOD	3	32.0-39.9
40		MAJOR FLOOD	4	40.0 +
50		NR RECORD MAJOR FLOOD	4,5	50.0-52.5
52.6	FOR	RECORD MAJOR FLOOD	4,6	52.6 +

FS - FLOOD STAGE
FOR - FLOOD OF RECORD

degrees of flood. Public impact and associated river stage, within flood categories, is simply not germane to the earlier stated verification goals. The highest degrees of flood, "near record" and "record flood", are flood levels not defined in terms of public impact because by their very nature ("standard industry practice") they have other definitions. And it works out nicely that these kinds of flood events can be used to "tie down the upper end" of a flood classification system like suggested, whereby in all cases we refer to a flood event that has clear meaning. If the forecast site has no river gage, (we have a few of these) flood categories could still be established, but it would be best to simply not use the site as a verification point due to the difficulty in obtaining flood level information during a rise. If the river gage malfunctions with flood in progress at a verification site, then after-the-fact crest information must be utilized (high water marks, eye-witness testimony, or a big, flat reasonable guess) to obtain verification data. The meteorologist often has great difficulty obtaining observed information for verification of a warning. The hydrologist has a comparatively simple task - his warnings are always site specfic. In summary, all or nearly all river forecast points should have established flood categories, even if only tentative, if the forecast service has substantial public impact. This is, my guess, likely at least 75 percent of the forecast points serviced by an office. This requirement places no great burden on field personnel, so I think. A collaborative effort between an RFC hydrologist and Service Hydrologist should result in task completion in a matter of weeks, with only few exceptions. It would be, in my opinion, one of the most useful things we could do, and would also be something of lasting value. The verification plan requires it, and the Agency should require it. Flood classificaton at a river site is a privilege and obligation of the National Weather Service. The meteorologist did not seek public approval, by way of example, in classifying hurricanes on a scale of 1 to 5, and the hydrologist does not need to seek public consensus on flood categories in a flood plain. What is required is that we decide to do it and in some reasonably scientific manner (i.e., give it some thought and investigation). Then a number of good things can happen, like verification and the ability to communicate throughout the Agency with reference to matters of the flood forecast service, all based on a common classification of events.

3.3 EXAMPLE OF FLOOD CATEGORIES

Figure 1 illustrates flood category assignment for the Trinity River at Dallas, a staff-graph presentation out of document E-19. On the left-hand margin, the flood category is noted, and to the right the threshold stage and some pertinent information as to why. Near the right-hand margin, important historical floods are specified. More detailed information may be contained elsewhere in the E-19, but the staff-graph page is a handy birdseye view of all that we need to know to establish verification for Dallas. Table 2 is a summary stage "banding" for each flood event. Note that as we move into major flood, the crest may fall into two categories. The rationale is this: A major flood (Cat 4) is a major flood; a near record flood (Cat 5) is a major flood (Cat 4); and a record flood (Cat 6) is also a major flood (Cat 4). This is important for scoring purposes, and we will illustrate the value of such shortly.

Surely by now someone has asked the question "what constitutes a near record flood?" Answer, "What do you want?" I see three options here: (1) An arbitrary value of, say, one foot below the flood of record (FOR) for all forecast points. But there are problems with this idea, known only to the hydrologist, unless someone else compares rating curves. (2) An arbitrary stage value unique to each site, assigned by the hydrologist according to what he thinks is reasonable. Once established, the number does not change. (3) An assignment of stage equal to, say, 10 percent of the FOR discharge, not to exceed one (?) foot. Probably the best idea of all. The choice is a matter of Agency (Hydrology) decision. Makes no difference to the verification plan; it is a choice to be made prior to plan implementation (if that ever happens), but once made, we should stick to it nationwide.

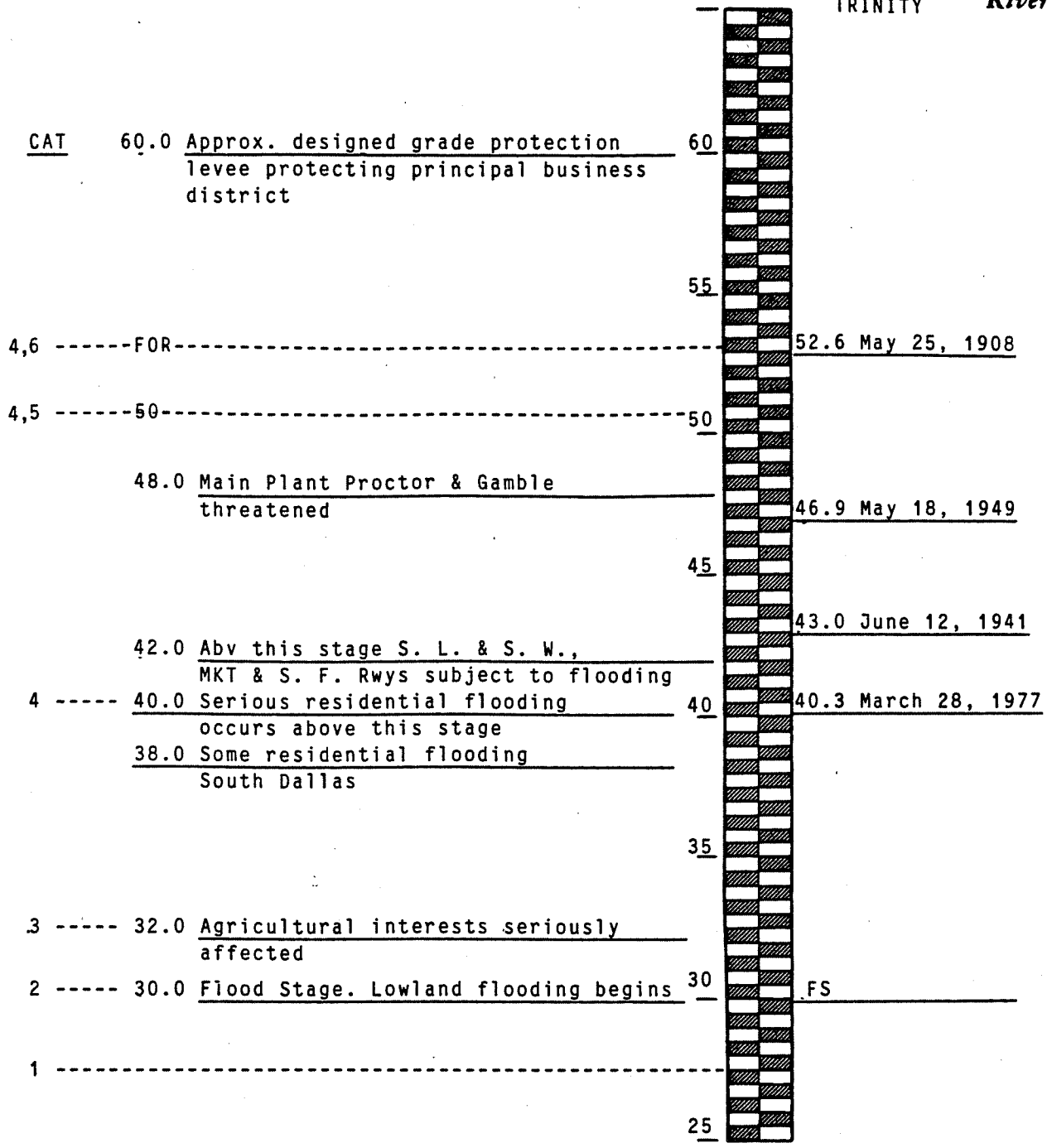
Figure 2 illustrates flood categories marked on a typical, parabolic shaped rating curve. However, whether the curve is single valued, looped, reverses at high elevation, or shifts like crazy during flood is of no consequence. The stage/categories remain as assigned, all of which is obvious to the practicing hydrologist. Such an illustration is not mandatory for a forecast point, of course. But it serves the purpose to perhaps more clearly focus on certain things of interest to the neophyte hydrologist, and the non-hydrologist who has been exposed to stage-discharge relationships somewhere. Bankfull (BF) and warning stages (WS) are also noted, and are commonly used to initiate a river forecast. The verification plan ignores this fact, as it is trivial to what we are looking

RIVER STAGE DATA

RIVER DISTRICT OFFICE
Fort Worth WSFO

DALLAS *Station*

TRINITY *River*



FS: Flood Stage
FOR: Flood of Record

Elevation Zero 368.02

REACH

Data Credits NWS, USGS

Date 6/30/80

FIG. 1. TRINITY RIVER AT DALLAS, NWS FORM E-19, STAFF-GRAPH EXAMPLE

for. When or why a forecast is issued is not a question we need be concerned with herein. Any event below flood stage (FS) is a "no flood." Any crest forecast below FS is a no flood forecast. For verification purposes, we will not concern ourselves with the occurrence of below flood stage rises unless a forecast is issued on the rise. We will go on the assumption that such rises are of little consequence. However, any crest forecast must get our attention, even for below FS, since "by chance," the observed crest may turn out to be something much higher, and we would be remiss in not giving the office credit for a bust.

3.4 SINGLE FORECAST, SINGLE CREST FLOODS

Figure 3 is a stage hydrograph, a single-crested flood rise Anywhere, USA, showing what the hydrologist sees every day. It is the central object of our grandest computer programs, but this example also indicates our flood categories, for purely illustrative purposes. However, aside from report illustration, the demarcation of flood catgories on any forecast hydrograph output is obviously of value to the hydrologist concerned with Agency verification of flood warnings, not to mention the otherwise inherent educational value of seeing the rise in a perspective of relative magnitude at the time the forecast is being prepared. The stage or crest, be it forecast or observed, is where the tag for event classification comes from (forecast event or observed event). In this report, crest will be considered as maximum stage on a rise, as opposed to a segment of the hydrograph. Bracket forecasts are covered in a separate section. Figure 3 makes it all the way to record flood (Cat 6). It is an observed record flood event. Forecast a stage anywhere above a Cat 3 rise, and you get credit for forecasting a major flood (Cat 4), because it happened. Forecast a near record crest (Cat 5), you miss because it didn't happen, but you still get a hit for Cat 4, as we just stated. Forecast a record flood (Cat 6) and you score another hit. Nice-good job. You did forecast a record major flood, and we give you due credit for both the major flood forecast and the record stage forecast, because both, in fact, did occur, so it was "your day." Forecast a, say, moderate flood, and you score a miss, period. Don't issue any forecast, for a rise above FS, and you score a "no forecast", so you cannot beat the system by saying nothing. Now, all along we have assumed a single forecast on a single event. We'll deal with multiple forecasts on a rise shortly, but first, what about stage error (how close we got), something other than "calling the event," that hit or miss thing we just talked about. Here's the philosophy, in the form of a dissertation: The Agency, for the purpose of verification, cares whether or not you called the event correctly. It cheers

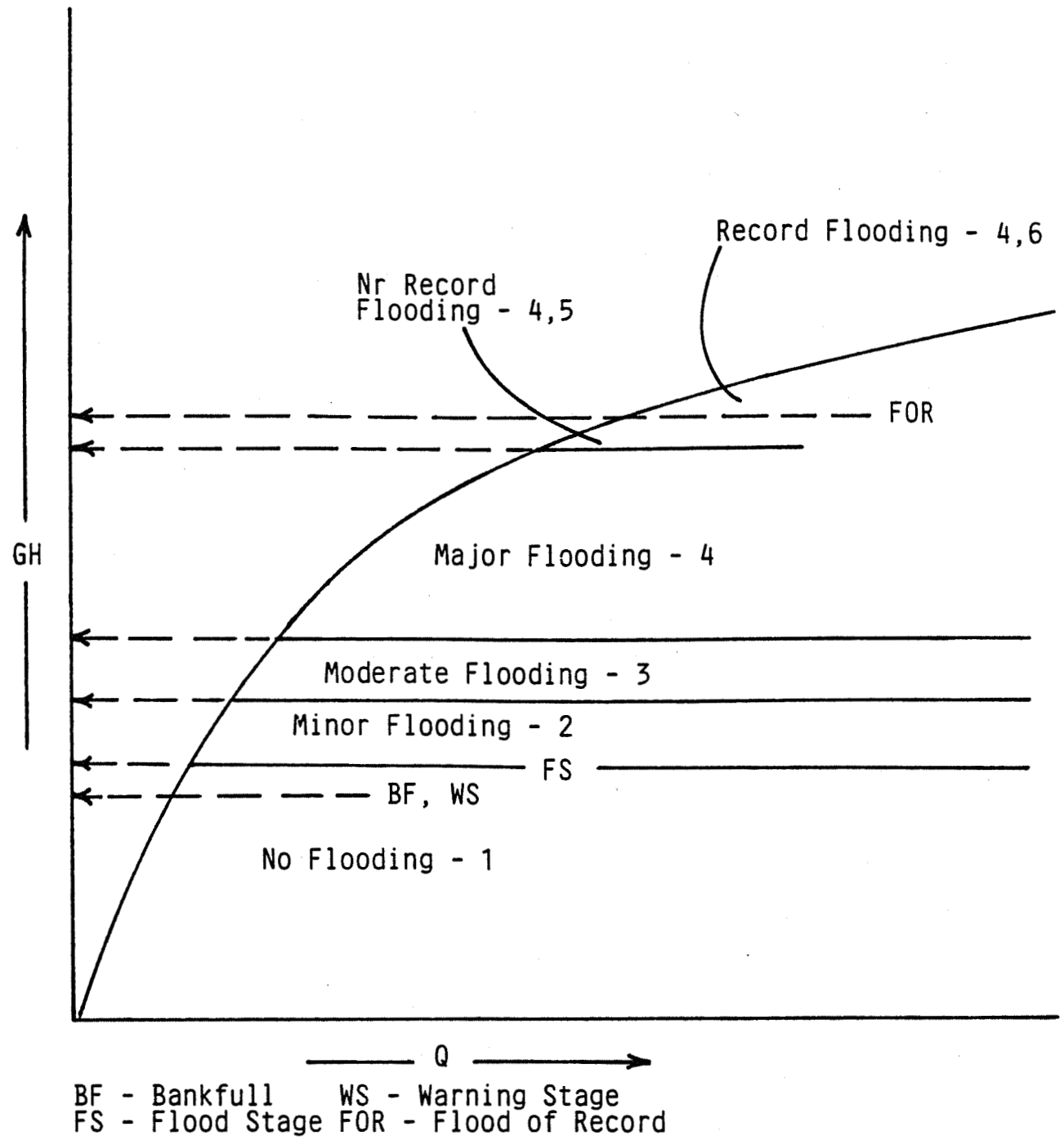
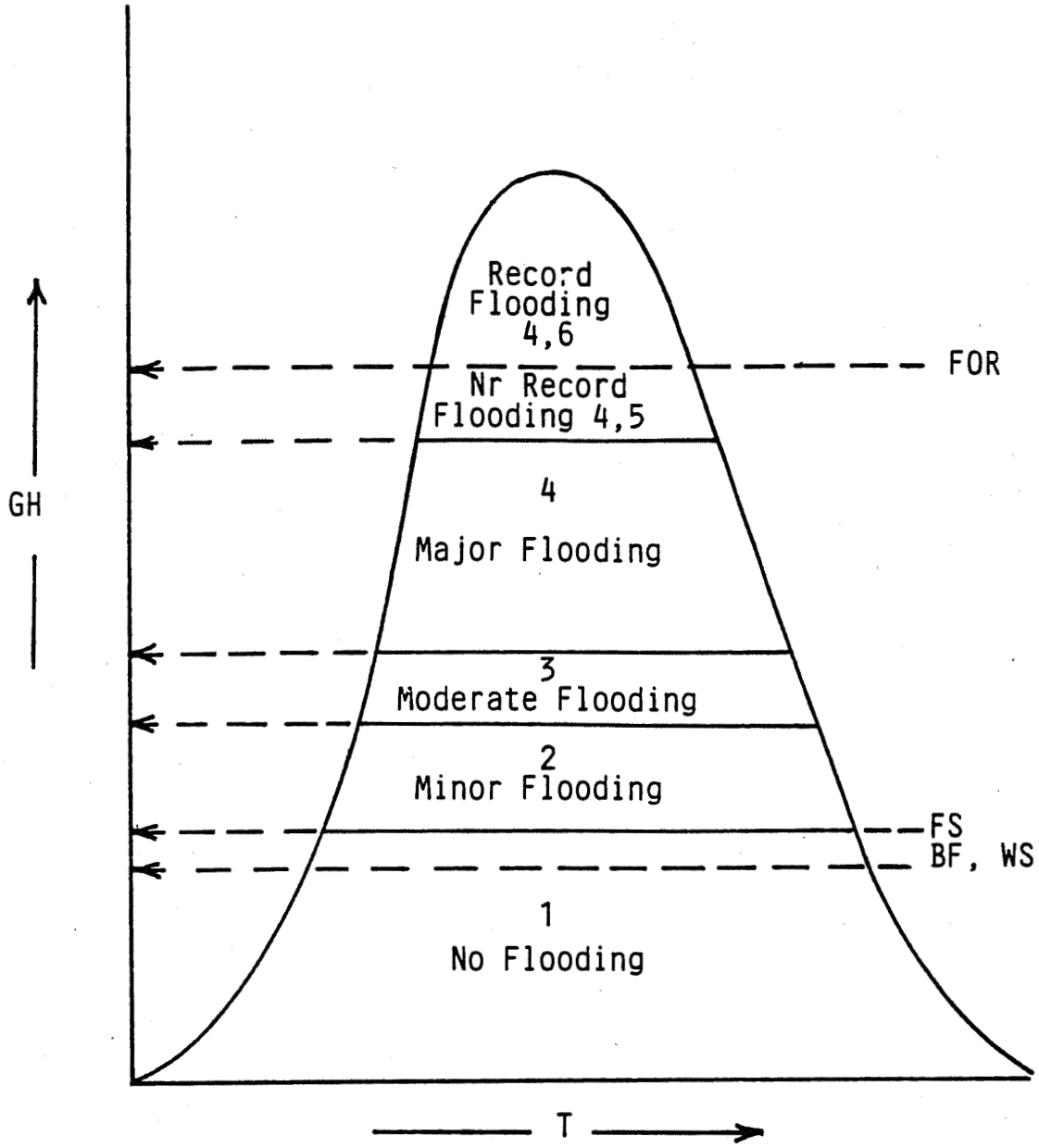


FIG 2. STAGE-DISCHARGE RELATION (RATING CURVE) INDICATING FLOOD CATEGORIES 1-6

STAGE (GH) VS TIME (T)



BF - Bankfull
FS - Flood Stage
WS - Warning Stage
FOR - Flood of Record

FIG 3. FLOOD HYDROGRAPH INDICATING FLOOD CATEGORIES 1-6

a hit, like in baseball. Call the right flood, and there is no verification error. It's Bingo!, and you get the pot. A crest forecast anywhere in the right category, whether above or below observed, whether off a half foot or much more, gets you gold. Your crest error may disturb someone in the flood plain, but it is possible that many more folks in that same basin are pleased. You called the event, and the Agency can at least defend the fact that the magnitude of flood (a most important thing) was correctly predicted. Miss calling the event, however, and we must look at the error - event error; not crest error. Crest error is an RFC problem. Event error is an Agency problem. Tell me there is always more science in your office than simply forecasting correctly flood events (and I obviously mean much more than flood vs no flood), and I'll then point out how you will just love this verification plan, since your skill is well within its tolerance for error.....and the Agency will watch your numbers for years to come to see how good good is. Point out to me that you can be off in a crest forecast by a half-foot, likely a super, super forecast anywhere, but still fall "outside" the observed event by an inch or two, thus getting nailed with a miss in categories, and I promise to sympathize with you. Then I'll point out how it is that any threshold based verification system can bite that way, and explain that the meteorologist faces similar verification frustration. There is nothing left but to apologize for missing the flood event by two inches. Undoubtedly the best advertisement an office ever got. If asked, nothing wrong with also explaining that the crest error was six inches - the question constitutes administrative investigation. The verification system will store in its data file your misfortune to the tune of two inches. I honestly would love for my office to post end-of-year stats that show we hit but few floods, and all those we missed we did so by an error averaging under two feet. All this strikes me as being preferable to a bunch of crest error numbers for river rises "zero to 100 feet" that nobody can make sense out of in terms of the service.

So we will state, as a matter of definition, that event error is the difference in feet between forecast crest and the observed event. Any stage prediction of "rise to" will be treated the same as a crest forecast. The algebraic sign is important: Positive is "overforecasting"; a forecast crest higher than the observed event, and negative is "underforecasting"; a forecast crest lower than the observed event. We always use forecast stage or crest values to, first of all, determine for the event the flood category predicted, and secondly, to compute for the event a stage error if the flood category is not correctly predicted. The only

significance of the observed stage or crest value is to determine the observed event category, but the observed category stage limits, at least one of those limits, will be used to compute forecast error if the forecast stage or crest is for an event of different category. All this is best made clear by example. Timing error, in reference to time of forecast stage vs time of the equivalent observed stage, will not be addressed in this report, as it is not (as will be seen) relevant to flood verification by categories. Forecast lead time, however, is addressed later in Section 4, and will not be elaborated on until then. Also, suffice it to say for now, and all discussion through this Section 3, a forecast is always viewed as occurring at the specified time or period of time so stated in the public release, and, categorically speaking, we view observed data to verify against, for the same time or period of time. This matter is discussed in detail in Section 4. Also, in this section, a "zero lead time warning" is not considered. This occurs when a river is already at a level within a category of flood, that was not predicted, and the subsequent crest forecast falls within the same category. The zero lead time warning will be discussed elsewhere in this report. Otherwise, the system would work like this, using Dallas to illustrate (Table 2):

Example 1. Observed moderate flood event: stages are 32.0 ft. to 39.9 ft. in this category. Forecast a crest anywhere in this range, and observe a stage in same range for the same period, and the forecast event error is zero. You did, in fact, forecast a moderate flood, and one was observed. That's important information to the Agency.

Example 2. Same observed event, but the forecast crest was 46 ft., a major flood. No "bingo" here this time. You failed to forecast the moderate flood, which only goes as high as 39.9 feet. Event forecast error is 46.0 - 39.9 = 6.1 feet. You overforecast the event by 6.1 feet. That is true, and the Agency cares.

Example 3. Same observed event, but the forecast crest was 31 ft., a minor flood. You failed to forecast the moderate flood, which begins as low as 32.0 feet. Event forecast error is 31.0 - 32.0 = -1.0 foot. You underforecast the event by one foot, which is fact.

Example 4. Suppose, for this example only, that Dallas had no criteria for moderate flood (Cat 3). It's minor then major flooding. The minor flood category must then be 30.0 to 39.9 feet. No problem. Observe a crest of 47 ft., a major flood, and assume a forecast of 35 ft., a minor flood. It's a

miss, of course. You failed to forecast a major flood, which begins at 40 feet. Event forecast error is 35.0 - 40.0 = - 5.0 feet. You underforecast the event by 5 feet, which again is fact.

Example 5. Observed crest was, say, 48 ft., a major flood. Forecast was 31 ft., a minor flood. You failed to forecast the major flood, which begins at 40 feet. Event forecast error is 31.0 - 40.0 = -9.0 feet. You underforecast this event by 9 feet. Now we really care.

Example 6. Observed crest was 53 ft., a major flood and a record flood. Forecast was 45 ft., a major flood. You get a hit for major flood (zero error for that event). You did in fact forecast a major flood, and that is very important. You did not forecast a record flood, any crest 52.6 ft. or more. Error for this event then is 45.0 - 52.6 = -7.6 feet. You missed the record flood event by underforecasting 7.6 ft., which is true.

Example 7. Same observed event. Forecast was 51 ft., a major flood and a near record flood. Again, you hit the major flood, which is always nice. You did not forecast the record flood. Error for this event is then 51.0 - 52.6 = -1.6 feet. You missed the record flood event by underforecasting it 1.6 feet.

Example 8. Same observed event. Forecast crest was 54 ft., a major flood and a record flood. Two hits here - double credit, and zero error. One may legitimately claim credit for forecasting a major flood, which occurred, and a record flood, which also occurred. The Agency is happy, as the flood event was correctly forecast, any way you look at it. Event forecast error was zero. Incidental is the fact that crest error was 1 foot, should anyone ask - and they will have to ask. No sense setting up a possible Agency internal debate over whether or not a one foot miss for a particular record breaking, frog strangling, people threatening, property destroying, monster flood was a "good" forecast, when it can be shown to the Agency that the event was correctly predicted according to pre-determined criteria used nationwide. Also, a chorus of public complaint, not unusual after a record breaking flood, is likely better handled by an office when a verification system such as suggested herein is utilized and recognized by the profession. However, regardless of one's opinion on such matters, the Agency had certain key questions answered in all the above examples, and that is what verification is all about.

It should now be clear from these examples what the concept is supporting our categorical forecast verification plan. Simply stated, we say that at any given time, observed river stage reflects an observed rise of some established order of magnitude, and, similarly, forecast stage reflects a forecast rise of some established order of magnitude. So long as the magnitude (category) of river rise, forecast versus observed, is identical, there is obviously no categorical error. If the rises are not identical in magnitude, the error involved is addressed by answering the question "what is the minimum stage, plus or minus, required as a change to the forecast such that the event would have been correctly predicted." We thus compute a difference in river height that is a reasonable measure of our failure to warn the public of the magnitude of the flood event. If the hydrologist can be comfortable with computing a stage or crest error - and he always has been - the hydrologist should be comfortable with computing an event error. Same type of thing, we simply look at a different thing - one that lends itself to verification interpretation. It is a thoroughly valid concept, I believe, and one that fits nicely into the existing framework of Agency procedure for weather event verificaton.

Example 9. This last example is a really bad bust: Forecast was 33 feet, a moderate flood. Observed crest turned out to be 53 feet, a record flood. River stage at the time of forecast issuance was 30 feet, a minor flood. You forecast a moderate flood, and it happened - enroute to the record. You score a hit for the moderate flood event. But, unfortunately, the record flood was not predicted. A record flood event at Dallas begins at 52.6 feet. Forecast error would be 33.0 - 52.6 = -19.6 feet. You underforecast the record flood event by 19.6 feet.

One final consideration, or two. Suppose a flood of some classification is observed, and no forecast is issued - not even one. It's a "miss", to state the obvious, but we cannot compute a forecast error. In the verification summary for that office, the number and kind of flood events observed with no forecast issued will simply be tabulated for the record. There may well be an explanation for the "non-forecasts", but I don't think I would care to have a bunch of these showing up for my office. Also, for the record, we state that qualifiers, like "near" and "around", will be ignored. For example, forecasts of "rise to near 20 feet Tuesday" or "crest 30 feet around noon tomorrow" would be viewed, for verification purposes, at face value minus the qualifiers.

3.5 MULTIPLE FORECAST, SINGLE CREST FLOODS

Now, how about multiple forecasts for a developing rise, a more typical scenario along the larger rivers? We will again not consider forecast lead time until later in Section 4, and take the same approach to verification as in Section 3.4, but work up the rising limb of the hydrograph until crest occurs. There really is no difference in what we do. The following so illustrates this, again using Dallas (Table 2) as an example. We will assume a steady rise and a record flood for "good exercise".

First forecast: Crest 30 ft. (Cat 2), or rise to 30 ft. (Cat 2) tonight, tomorrow, or whenever. From a categorical viewpoint, there is no difference between a forecast of "crest at" or a forecast of "rise to". We look only at the forecast river level because that determines the forecast flood category, and both terms provide the necessary stage information to the public.

Second forecast: The river is now 33 ft., a Cat 3 moderate flood, and rising. The event is not over yet (crest has not occurred), and so you get credit for the minor flood forecast. It did happen. But should the river now "roll over" in Cat 3, you busted it - you forecast a minor flood and observed a moderate flood, and an error will be computed. No credit given for the minor flood forecast under these circumstances, as it is a single forecast, single crest event as discussed in Section D. However, all this is not the case and the continued rise affords one the opportunity for another forecast. It is not necessary that one forecast every degree of flooding on a rise for a multiple forecast event. In fact, the hydrologist may not have the opportunity to do so. The rain continues, and your second forecast is "rise to" 45 ft., a Cat 4 major flood.

Third forecast: This is getting serious. The river is now 43 ft., a Cat 4 flood, and due to more rain, you wish to forecast a flood of record. Well, fine, but first you score a hit for the major flood in progress. You said it was going to happen, and it did. From here on in the sequence of forecasts you cannot again claim credit for major flood prediction, so long as the river remains above 40 ft. (Table 2), the major flood threshold stage. You now issue a third forecast for 53 ft. That's Cat 6 degree of flooding according to Table 2; a record flood.

Fourth forecast: Rate of rise has slowed, you have discovered some "exaggerated" rainfall reports or bad stage data, and you change your mind. Right now the river is, say, 49 ft., still Cat 4 level, and you decide to downgrade the projected crest to 52 ft., a Cat 5 near record flood for tomorrow afternoon.

Fifth forecast. The river crests at 52.6 ft., to tie the flood of record. That's a Cat 6 event. No cause for tears, but you did miss calling it in the end. Now how do we score all this? Well, here is what happened for the storm:

1. Forecast minor flood. Verified.
2. Forecast major flood. Verified.
3. Forecast a record flood. Verified.
4. Forecast a near record flood. Did not verify. Error is 52.0 - 52.6 = -0.6 ft.

Perhaps it is not clear why you even get credit for a record flood forecast after you changed the prediction to something less. Well, note that the change cost you. Here, again, is what we seek to learn: the Agency, for verification purposes, needs to know what your forecasts were during a flood period, and how it turned out. Makes no difference that you changed your mind somewhere along the line - you don't retract a forecast. You can amend a forecast, but that's a new forecast. By good fortune, you still scored when the river crested, and it's a shame circumstances dictated a final crest forecast that proved to be a "catgegory off". The record so notes this, but obviously it was hardly a serious matter. And what is the public reaction to this chain of prediction? I think it is fair to state that forecast number 3 is what really got that public "moving" in the flood plain, and the later downgrade by a foot was of little consequence - to your public reputation (hopefully) or your Agency reputation (for sure). The only reason I mention such things again is to demonstrate that this verification system is not detached from the realites of river forecasting and our dealings with the "outside world". I believe it accurately reflects what we do in an office on a day-to-day basis, is a valid measure of the service rendered, and does not penalize the hydrologist for the sake of a score card. There is no arbitrary, off-the-wall stuff going on here.

A final word: Here is a circumstance not mentioned in these examples. Suppose we issue a forecast for <u>Cat 3 moderate flood event.</u> and the river rises to Cat 4 major flood. If the river crests, the bust is obvious, and an event forecast error would be computed. But suppose you think the river will continue to rise to a Cat 5 near record flood event, and you so

state. Say the river does, in fact, crest in the near record flood range, so you do get credit for the Cat 5 hit, of course. Do you also get credit for a Cat 4 major flood forecast, since a Cat 5 event is obviously a major flood? The answer is no. The river was already in Cat 4 range when the Cat 5 prediction was issued. You do not score a Cat 4 miss either. But should you be so unfortunate as to witness the river crest in the Cat 4 range after the Cat 5 prediction, you have a double bust on your hands: The river crested major flood, which you did not predict, and the Cat 5 forecast also did not verify. However, depending on the forecast crest numbers, the computed event errors could be small. Once a river moves into a category of flood not forecast, during the period in question, the only way to escape a verification "ding" is for the river to rise into a higher category of flood that "gets" forecast.

3.6 MULTIPLE FORECAST, MULTIPLE CREST FLOODS

Same approach to business as in Sections 3.4 and 3.5. Here is the guiding rule: a verification event always occurs whenever the forecast or observed stage, on a river rise, changes categories. We do not verify recessions - forecast or not forecast. Once a river is in a certain degree of flood, you can forecast one crest after another for flood in the same category, but we only score the first. The rationale is simple. We verify your ability, or lack of it, to forecast a category of flood along a river. Once the river reaches that flood level, we have no reason, for verification purposes, to monitor the subsequent rise and fall, or fall and rise if you will, of stage within that category, except for the purpose of determining when the river leaves a given category of flood. Multiple crests, especially a low stage crest followed by a high stage crest, all within a given category of flood, can obviously be of great importance to the public. But such a series of predictions cannot be of consequence to the Agency's categorical verification program. Major rivers can, of course, go into protracted flood lasting weeks, and it is grinding hard work - stress, problems galore, and one forecast after another. However, we only want to know did you forecast each level (category) of flood (Table 1) during the storm, so it is that initial forecast for each category that counts. In other words, we ask, did you "forecast that flood?" We do not ask, "how long did the flood last?" If a river comes off a crest, recedes temporarily to a lower category of flood, and then rises again to another flood catgegory, the rise is another verification event, and will be scored. So, folks, watch your recessions. Hydrologists may suffer some indigestion over the prospect of laboring over a few or more important forecasts for which no

verification is performed, but one should not expect a verification plan to faithfully document the amount of sweat expended any more than one should expect the plan to fully support Agency or public appreciation for services rendered. Verification may or may not reflect the total forecast effort, and Agency awards should be used to recognize an office for extraordinary service largely independent of any internal forecast "scoring" that is being maintained. Now that the preamble is over with, we can afford to look at the next Figure.

Figure 4 is a near eight-day hydrograph for Dallas that is a product of several runoff producing storms. For this flood a total of eleven forecasts were issued. It is a multiple forecast, multiple crest flood event, like are so common along rivers during wet periods. We will ignore lead time computation at this time, and concern ourselves with only the occurrence or non-occurrence of the forecast event for the forecast period. Here is what happened:

Forecast (1) - crest 33 ft. (Cat 3 flood) for time period "x". We observe 28.5 ft. for the crest (Cat 1 no flood) during the same period. The forecast is a "bust". You forecast a moderate flood (any stage 32.0 - 39.9) and observed no flood (any rise up to 29.9 ft.). Event forecast error is 33.0 - 29.9 = 3.1 feet. You over-forecast the "no flood event" by 3.1 feet - 'tis true.

Forecast (2) - crest 31 ft. (Cat 2 flood) tomorrow evening, but due to overnight rains, the river rises through 31 ft. and "keeps on going". You forecast a minor flood, and it was observed. You score a hit - and it's time for another forecast.

Forecast (3) - crest 37 ft. (Cat 3 flood) tonight. We observe 36 ft. (Cat 3 flood) as the river rolls over. You forecast and observed a moderate flood. Nice job, and a "hit". But it keeps raining, and the river again rises, to your surprise, cresting out again at 39.8 ft. That's another Cat 3 crest, for which no forecast was issued. But it doesn't hurt you because it is not scored. The river never left a Cat 3 level of flooding. Had you issued a forecast on the rise calling for another Cat 3 crest, it still would not have counted in the scoring. Remember, it's only the first forecast for a flood category that counts. However, had the river crested at some higher level of flood, with no forecast issued, the miss would have been noted for the record. It starts raining again, and so:

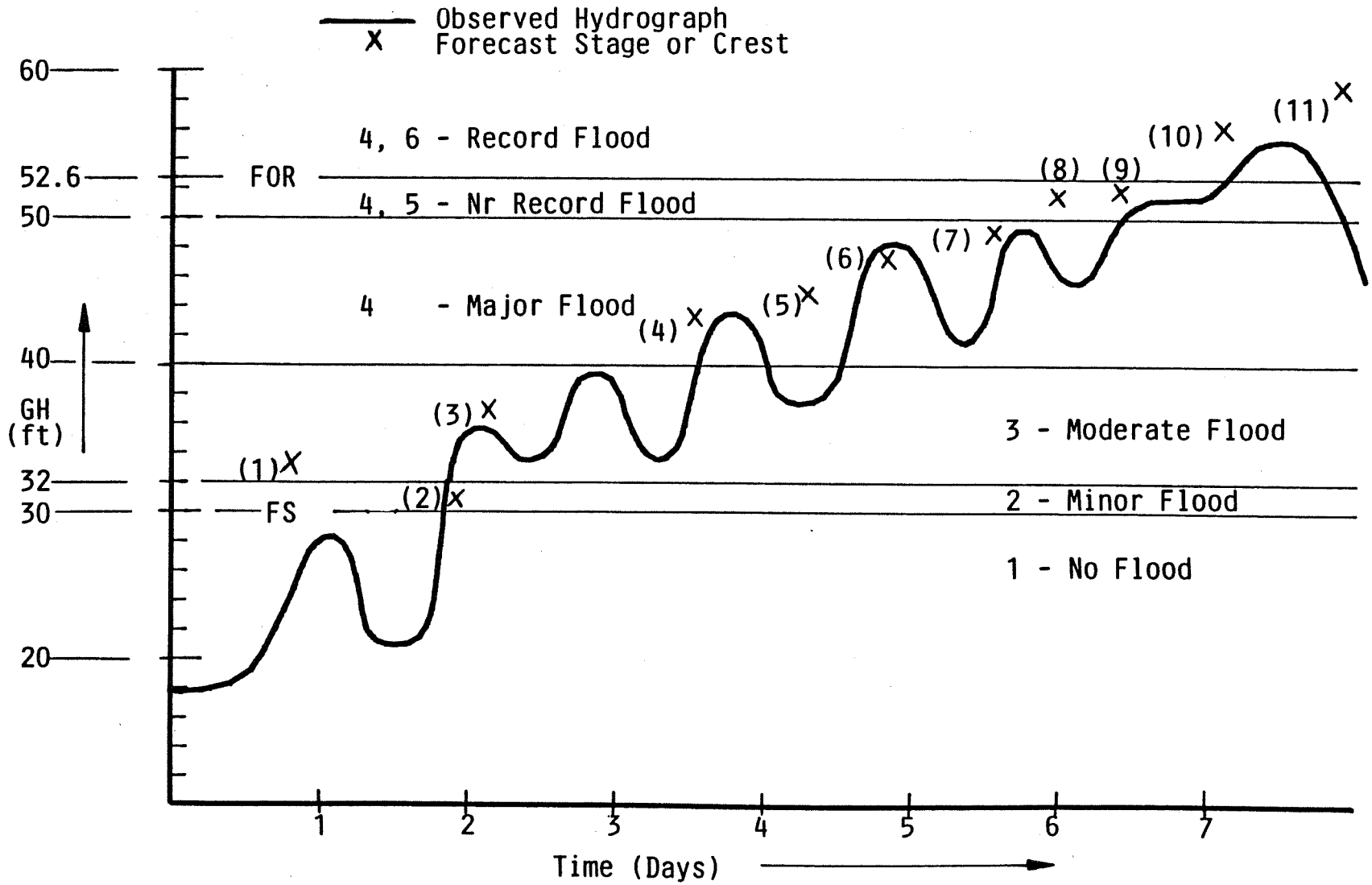


Fig. 4. COMPLEX STORM HYDROGRAPH FOR TRINITY R. AT DALLAS

Forecast (4) - rise to 43 ft. (Cat 4 flood) tomorrow late morning, and we observe a stage of 44 ft. (Cat 4), so a "hit" is credited for this major flood. It verified. We have reason to think the river will continue to rise after a pause at the 44 foot level, so we issue another prediction:

Forecast (5) - crest 45 ft. (Cat 4) AM tomorrow. However, the river does not continue to rise, but instead recedes below 40 ft. into the Cat 3 moderate flood range. How embarrassing, but it happens. You forecast the river to be in major flood the following morning and it receded to moderate flood. Event error is 45.0 - 39.9 = 5.1 ft. You overforecast the event by 5.1 feet. Note: If no forecast had been issued, no error would have occurred because we are not concerned with recessions. However, should the river "turn around" after receding to a lower category, and then rise to a higher level (category) of flood, this represents a new event, and would be scored - either as a "no forecast" event, or as an event for which prediction was made, good or bad. But we are in a wet cycle, it rains some more, the river is now at 37 ft. (moderate flood level), and you must forecast an impending rise in the river.

Forecast (6) - crest 46 feet this evening. You have called for the river to rise again into major flood category. The river crests at 48 ft., a major flood. Nice. You did forecast another Cat 4 flood, it happened, and a "hit" is credited to your account. The river again recedes, but not below major flood level; it rains, and time for another forecast.

Forecast (7) - crest 48 ft. (Cat 4) early afternoon, and the river crests then at 49.5 ft., Cat 4. Good, but you were already dealing with a river in major flood, so the forecast is not scored. However, you think the river will shortly rise again to near record flood level, so the public is advised of this.

Forecast (8) - crest 52 ft. (Cat 5) late evening. However it doesn't happen, at least not then. In fact, the river is receding (still in Cat 4 range). You forecast a Cat 5 flood, but observed a continued Cat 4 event. Error is 52.0 - 49.9 = 2.6 ft. for the event. But the river turns around, and you try again. The flood is not over yet.

Forecast (9) - crest 51.5 ft. (Cat 5) AM tomorrow., and the river rises to 52 ft. and holds. Fine job. Forecast (9) verifies.

The river is not receding yet; a sister agency advises you that flood control projects must now pass inflow, and it commences to rain again, to make things even worse. Time to forecast a record flood.

Forecast (10) - crest 57 ft. (Cat 6) late afternoon. The river slows at 54 ft. (Cat 6). We have successfully forecast a new flood of record, and score accordingly. We re-examine our routings and decide the river will climb even higher, and so:

Forecast (11) - crest 58 ft. (Cat 6) late afternoon the following day. Instead the river recedes dramatically from 54 feet to near record flood level (Cat 5). Your attempt to forecast an even higher record stage failed. Event error is 58.0 - 52.5 = 5.5 ft.

For this multiple forecast, multiple crest flood, the verification record looks like this:

1. Forecast moderate flood. Did not verify. Error is 3.1 feet.
2. Forecast minor flood. Verified.
3. Forecast moderate flood. Verified.
4. Forecast major flood. Verified.
5 Forecast major flood. Did not verify. Error is 5.1 ft.
6. Forecast major flood. Verified.
7. Not scored - continuation of same flood.
8. Forecast near record flood. Did not verify. Error is 2.6 ft.
9. Forecast near record flood. Verified.
10. Forecast record flood. Verified.
11. Forecast record flood. Did not verify. Error is 5.5 ft.

Pretty good forecasting here, and, I believe, honestly evaluated. It's what goes on in a River Forecast Center, and some Met forecast offices as well. The Agency has a valid measure of "how we did". I suspect the typical forecast effort would look very much like this hypothetical flood.

Undoubtedly by now someone is picturing in his mind a forecast point with a potential stage range from, say, 10 feet (low flow) to 70 feet (flood of record), with normal stage near 15 feet and major flood designated 25 feet and up. The consequence of this during a wet period could very well be a

quick one-shot forecast for 25 feet or more to hit the major flood category, followed by a possible half-dozen or so forecast major flood crests under, say, 50 feet, none of which would count. And the conclusion would thus be that the proposed verification system is therefore nonsense. Not so. The nonsense is how the flood categories were assigned at that forecast point in the first place. The flood definitions are not a substitute for hydrologic intelligence, and it seems to me that flood levels would naturally be assigned by the hydrologist so as to also reflect flood frequency, normal, and likely range of stage a river is apt to experience. If, by way of example, there is a population somewhere suffering "major flood" at 25 feet for a river normally at 15 feet that frequently reaches 30 feet, then I must conclude that a bunch of folks are using natural overflow to irrigate their lawns, and the National Weather Service is well within its right to raise the Cat 4 level to some value above 25.

3.7 THE BRACKET STAGE FORECAST

It is common practice for River Forecast Centers to predict rivers to rise to some range of stage (a bracket). Such a forecast reflects the uncertainty inherent in river prediction under some circumstances. Typically, the bracket forecast is issued early on in a developing rise, and then narrowed to a specific value during later public updates. The Categorical, Event Oriented, Flood Forecast Verification system accommodates the bracket forecast nicely, and with no change in logic. Again, we will use Dallas (Table 2) to illustrate verification procedure.

Suppose a forecast is issued "crest 30 to 31 feet Monday". A minor flood event at Dallas occurs 30 to 31.9 feet, by earlier definition (Table 2). Consequently, the bracket forecast falls entirely within the minor flood (Cat 2) range. If a minor flood crest is observed (30 to 31.9 ft.), the forecast is a hit (it verified), and nothing more need be said. But, should the river crest out at, say, 33 feet, a moderate flood (Cat 3) was observed, and an error must be computed. In this case, event forecast error is 31.0 - 32.0 = -1.0 foot. We underforecast the event by one foot. The rationale should be clear: We predicted that the river could rise to a stage as high as 31 feet (a minor flood), but it crested out in the moderate flood category, which begins at 32 feet, so we missed predicting the actual event by one foot.

Now suppose the forecast was "crest 38 to 40 feet Monday". That is a forecast of Cat 3 (moderate) to Cat 4 (major) flood,

according to Table 2. If the river crests in the moderate category, you score a hit for Cat 3, and a miss for Cat 4. If the river crests in the major flood category, you score a hit for both the Cat 3 and Cat 4 events. If the river reaches Cat 5 (near record flood), you, of course, still verify nicely for the Cat 3 and Cat 4 predictions, but an error must be computed for the Cat 5 miss. Example: The observed crest was 50.5 feet (Cat 5 flood). Your highest crest prediction was 40 feet, for a major flood event. The observed near record flood event begins at 50 feet. Event forecast error would be 40.0 - 50.0 = -10 feet. We missed predicting the event by 10 feet, which is true. The verification score card would look like this for the bracket forecast of 38 to 40 feet and an observed crest of 50.5 feet.

1. Forecast moderate flood. Verified.
2. Forecast major flood. Verified.
3. Observed near record flood. Did not verify. Error is 10.0 feet.

Let's take a look at a bracket forecast of "crest 38 to 41 feet", which is a forecast of Cat 3 to Cat 4 flooding. We stated that if the river crests in the moderate flood category, you score a hit for Cat 3, and a miss for Cat 4. How would the Cat 4 major flood event error be computed in this case? For an observed crest falling within the lower category of a forecast bracket that includes two categories of flood, we really do not have a peculiar situation. By definition, event error is the difference in feet between forecast crest and the observed event, and the error represents the minimum stage, plus or minus, required to change the forecast such that the event would have been correctly predicted. Now, what is the forecast crest within a bracket? We do not have an explicit single number to compute an error from. So we have a computational problem, or do we? In the case of Dallas, the 38 to 41 ft. bracket is (Table 2) moderate flood, 38-39.9 ft., and major flood, 40-41 feet. Say the river rises from minor flood to crest at 39 feet. The moderate flood event was correctly forecast. But what is the forecast crest for major flood - 40 ft.?, 40.5 ft.?, 41 ft.? What should we use for error computation? Answer: We will always use the highest or lowest stage in the forecast bracket to compute event error, regardless of circumstance. We will reason here as always: The observed event, moderate flood, which was forecast, "tops out" at 39.9 feet. The bracket forecast also warned the public of a possible major flood to 41 feet. No major flood event was observed. What is the stage reduction in the bracket forecast necessary to eliminate the erroneous forecast of major flood? Answer: 41.0 - 39.9 = 1.1 foot. You overforecast the event by 1.1 foot, which is true.

There is no computational problem. Notice how this encourages the hydrologist to keep the bracket as small as possible? This is good.

There is one more situation to ponder. What if the river never rises out of minor flood? Real nice bust going now. Forecast a moderate to major flood (two degrees of "anticipated wet"), and observe a minor flood, and we have lots of error to compute - two misses for two events. You did, in fact, forecast two events and neither was observed. Reasoning stays the same. Say the river crests at 31 feet, Cat 2 minor flood. The minor flood observed event goes as high as 31.9 foot stage. The bracket forecast alerted the public to a stage of at least 38 feet, a moderate flood. A flood of this magnitude did not occur. Event error is 38.0 - 31.9 = 6.1 feet. By forecasting a 38 foot stage moderate flood, you overforecast the minor flood by 6.1 feet. The bracket forecast also warned the public of a stage as high as 41 feet, a major flood. Event error is 41.0 - 31.9 = 9.1 feet. By forecasting a 41 foot stage major flood, you overforecast the minor flood by 9.1 feet. Two "misses" and two errors credited to your verification account ..... encourages the hydrologist to think carefully before issuing a forecast for a river to rise into two higher flood categories via the "safety" of bracket forecasting. This is good, too. If done with skill and confidence, the verificaion payoff is high, but failure tends to ruin your day. No complaint is due here. The stage or crest error, had we looked at that instead, would not have made one proud. At least with event error the Agency can draw some intelligent conclusions.

3.8 THE NON-STAGE SPECIFIC CATEGORICAL FORECAST

From time to time an office may issue to the public a flood warning for a given reach or site along a stream that calls for some degree of flood, but no specific stage or range of stage is mentioned. One such example would be "a major flood is developing along Dung River due to continued heavy rains. All interests within the Dung flood plain should immediately take precautions....", and so on. Typically, the flood warning further states that a stage or crest forecast will be issued shortly (within hours). For verification purposes, we will wait until the stage specific forecast is issued, because it would be awkward, if not impossible, to attempt earlier verification. Bear in mind, <u>we always use stage or crest forecast numbers to accomplish two things in our verification plan:</u> (a) Determine for the event the flood category predicted, and, (b) compute for the event the stage error if the flood category is not corretly predicted. A categorical flood warning without stage causes no

problem for item (a), but item (b) cannot be accomplished if the forecast is a bust. One could argue that the mere mention of a flood category in the public message (like moderate flooding) implies a bracket of stages (in the case of Dallas, 32 to 39.9 feet), but the range of stage is typically quite large, is likely not known to the public, and thus would not be a suitable substitute for a hydrologist determined re•sonable span of river rise. Regardless of one's opinion on this, a non-stage specific hydrologic forecast is sufficiently vague so as to not have much value in any verification plan that attempts to examine the primary river forecasting efforts of the Agency, particularly if a numerical forecast error is a factor to be computed.

I believe we can legitimately say that pure categorical forecasts are relatively few in number, are primarily public "heads up" warnings to take precautions, and serve to cause those affected by rising water to listen for follow on crest predictions that will, in fact, be verified. The Agency is not overlooking a vital piece of it's hydrologic service by excluding non-stage specific flood advisories in its verification program. The Weather Service Forecast Office (WSFO) based county flash flood (generalized) warnings are, of course, a different type of warning, and are already handled by another verification program.

3.9 MINOR CHANGES IN STAGE ABOUT A FLOOD LEVEL

What we have in mind is the river that fluctuates or oscillates around a particular level for some period of time, i.e., there is a continuing rise and fall of the stream without substantial change in stage. Some rivers, particularly the larger streams swollen by flood, can exhibit this kind of behavior, and if the oscillation in stage should occur about a given verification level, the consequence would be a river forecaster uttering unprintable things about verification madness. This can be prevented.

The question to be raised here because of the threshold nature of our verification plan, is what constitutes a significant change in river level when a stream is running close to a flood category? Within any category, a change of stage, large or small, is not important to verification; however, not so at the flood threshold levels. So what should we do? Well, let's review what the river forecaster does.

Case 1. The hydrologist feels that the change in river level will not be significant (plus or minus a foot or less?), and need not be spelled out in the forecast, so the

public statement takes the form "LCRVR" (little change in the river next few days). This is another non-stage specific forecast, and will not be verified.

Case 2. The hydrologist feels the change in river level is sufficient that a range in stage should be mentioned, and he decides what the range will be. This is a bracket forecast, as discussed already in Section 3.7 of this report. For such a river holding near flood level, let's look again to the Trinity River at Dallas for illustration. Say the river has risen to 39 feet (moderate flood), and is expected to run 39 to 41 feet for a few days. Major flood starts at 40 feet. We will not permit a "few days" time specfication. It is acceptable to so state this for public consumption, but we cannot live with such vagueness for any kind of verification. We require a set forecast time, like, say, three days, for internal purposes. So the forecast is "39 to 41 feet for next three days", which is moderate to major flood. This implies that, for the stated period, the river at Dallas will experience stage in both the moderate and major flood categories. Since the river is already in moderate flood, no "hit" possible here. If within three days the river rises at least once (and only the first counts, remember?) to any stage 40 feet or higher, a hit is scored - major flood was observed. If the river rises to, say only 39.9 feet, that is still only moderate flooding, and the forecast is a bust. Consequently, the event forecast error would be 41.0 - 39.9 = 1.1 foot. We predicted the river to be within the range of moderate flood (39 to 39.9 feet) to major flood (40 to 41 feet), for a bracket total of 39 to 41 feet. We observed only a moderate flood event, which can go up to the 39.9 foot stage. But we alerted the public to a possible 41 foot stage major flood, which did not happen. You overforecast the observed moderate flood event by 1.1 feet.

Case 3. The hydrologist does not wish to use the "brackets" to predict river conditions for three days, and chooses instead to forecast a specific maximum stage each day. This is fine. When verification is performed on the forecast, he will likely find he "won a few and lost a few". The river forecaster has always worried about a stream moving in and out of banks, or in and out of flood stage, so there is nothing dramatic about a worry over some other flood level that flags the degree of flood.

It is Case 3 that focuses attention on the question of specifying a "significant change of stage" required before verifying a river holding near a category boundary. Now, it is true that a significant change in the stream likely varies with

river level due to the nature of the stage-discharge relationship and/or development within the flood plain. We could get real fancy about all this - think in terms of a specified range of stage (plus or minus X feet at each category boundary), and require a change in stage exceeding that range before verification counts, etc., etc., and complicate the verification plan something awful. Let's not do that. It is not worth the effort. First of all, a river "bouncing up and down" between flood categories through a small range in stage should be rare, and I see no reason to complicate the "frequent simple" to accommodate the "infrequent odd". Secondly, the "odd river" can be handled nicely, for verification purposes, and generally for public warning purposes as well, if the forecaster will just do like noted in Cases 1 and 2. This he would likely do even in the absence of any verification program. But if the forecaster insists on conducting business, makes no difference why, like discussed in Case 3, no problem. It is just another event forecast. So, not finding good reason to alter the verifcation plan for a river running close to flood level, we state for the record that any change in stage, a few inches to a few feet, about a category of flood, is significant. Believe it.

3.10 THE FLASH FLOOD

A flash flood is defined by Weather Service Operations Manual E-13 as "a flood which follows within a few hours of heavy or excessive rainfall, dam or levee failure, or a sudden release of water impounded by an ice jam" (NWS, 1981). More specifically, amongst field personnel, the flash flood is generally considered to be a rapid rise in water caused by intense rainfall over a relatively small watershed during a period of three hours (roughly) or less, and I doubt that anyone anywhere would consider a stream "flashy" if the cresting time exceeded six hours. There is no clear delineation for the "line" between flood and flash flood. I personally do not care for the use of the term flash flood for any basin outside mountainous regions, but somehow it has become vogue to apply the term to urban catchments, small town bogs, and country bayous. Permit me to say that we have been swamped by the overuse of flash flood warnings because it is a convenient word that conjures up in the mind just what we "warners" want conjured - a near panic reaction to rapidly rising water. Well, it is not the intent here to debate the issue of terminology. The report addresses only the problem of verifying <u>site-specific</u> flood warnings, including the "flashy kind".

We will view, as we must, the flash flood as a different

kind of event, because it has become standard practice to do so. It is obviously an event classified by time. It is an event distinguished from flood, a more generic term, by lack of time. Can we also think of further classification according to degrees of flash flooding? We occasionally hear the expression "severe or extreme flash flooding expected....," which implies something worse than a flash flood of more common variety. Is it sensible then to think in terms of minor flash flood, moderate flash flood, and major flash flood, so as to separate out the common from the less common? I don't think so. But since the flash flood event has magnitude, and clearly, for verification purposes at least, the Agency would like to develop a measure of our ability to predict the magnitude of flash flooding, we are compelled to consider flash flood categories. The following definitions are suggested for Agency adoption to use to verify site-specific flash flood warnings and distinguish flood from flash flood:

1. A flood is the inundation of normally dry area to the extent that property damage, personal injury, or economic loss takes place.

2. A flash flood is a flood in which the inundation follows the observable causative event by less than six hours. (Sittner, 1987, with modification.) By "observable causative event," for example, we distinguish between, say, a dam failure causing flash flood (a less than six hour event) and the copious rain leading to failure, which may occur for many hours prior to failure of the structure.

Now let's talk about what was just said. We clearly time specify the flash flood as an event with a cresting time less than six hours. Any flood cresting less than six hours may be classifed (hydrologists' choice) as a flash flood, bearing in mind that massive flood damage can occur long before crest on many rivers, and this is particularly worrisome for a stream with a spiked hydrograph. Should the forecaster deem a river site "flash flood", a different set of verification categories will apply: Category 1, No flood; Category 2, Flash Flood; Category 3, Severe Flash Flood, and Category 4, Extreme Flash Flood.

Table 3 is a listing of these suggested flash flood categories. Hopefully, the flash flood definitions, as was the case for flood in Table 1, will be largely acceptable to the profession. Figure 5 illustates the categorical flash flood. The rationale is the same as that behind Figure 3. The hydrologist for the flash flood site must establish both a Flood

Stage (FS), and, what we will call, Severe Flood Stage (SFS), and Extreme Flood Stage (EFS). What we are doing here, basically, is recognizing that a flash flood event does not lend itself to the detailed, established, and conventional descriptors of flood (Table 1), but can be adequately judged as to magnitude by the above Categories 1-4. These categories of flash flooding have, or should have, reasonably clear meaning to everyone. Thus we can understand what someone else is saying about the flood event when verification data are compiled, and the Agency acquires the means to judge our ability to predict flash flooding.

The establishment of FS requires no comment in this report. The establishing of SFS, however, is something new. What SFS is, as I see it, is a rather arbitrary, but realistic stage, above which a flash flood could be termed severe. This should be an easy thing for the hydrologist to determine, even if done as a reasonable guess, in the absence of a flood history. We would similarly establish the EFS, where EFS represents some very high flood level for rare events. The stage level for each category of flash flood, as in the case for flood, may be changed at any future time without injury to the verification database. Prior discussion in this report regarding document E-19 supporting data similarly applies to the flash flood forecast point, as does the rationale behind event error computation. For the flash flood we really do nothing different, conceptually, in verification. Verification error is still forecast stage or crest minus observed event. The following examples should demonstrate this fact well.

Table 4 and Figure 5 illustrates flash flood category stage assignments for a hypothetical river. Any rise below 10 feet is "no flood" (Cat 1); a rise above 10 feet and below 15 feet is a flash flood (Cat 2); and a rise of stage to 15 feet or more is, of course, a flash flood (Cat 2), but it is also a severe flash flood (Cat 3). A rise above 25 feet is a Cat 2 and Cat 4 event. All this look familiar?

Example 1. Observed a severe flash flood of 17 feet. Forecast crest at Marsville was 8 feet, no flood. Severe flash flooding commenced at 15 feet. You did not forecast a severe flash flood, but one was observed. Event error is 8.0 - 15.0 = -7.0 feet. You underforecast the severe flash flood event by 7 feet. Happens all the time.

Example 2. Observed flash flood of 14 feet; forecast crest was 9 feet. A Cat 2 observed event, and a Cat 2 forecast event; no error, and a hit is scored for the rise.

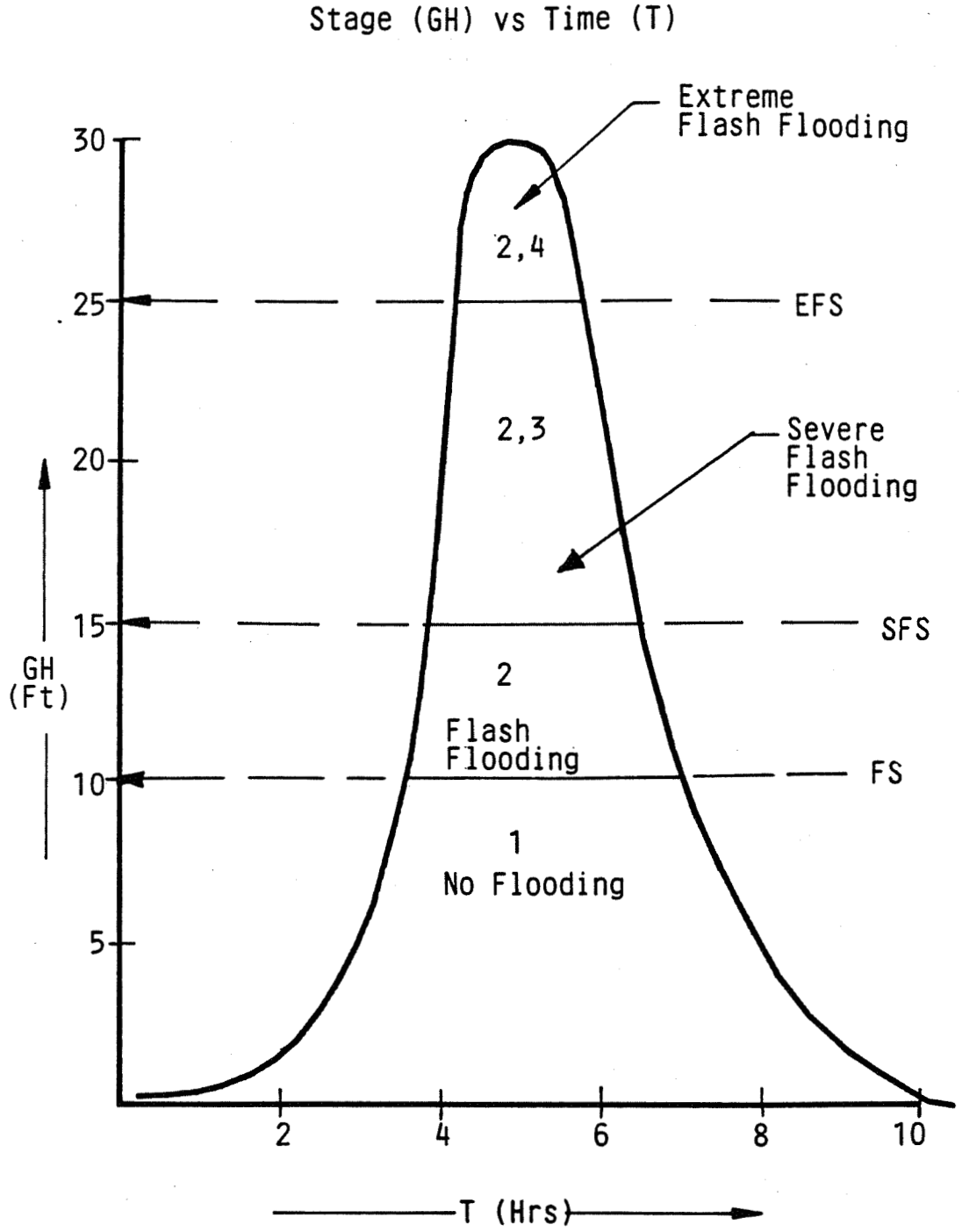
TABLE 3. CATEGORICAL FLASH FLOOD EVENTS

CAT	DEGREE	GENERAL DESCRIPTION
1	NO FLOODING	NO FLOODING EXPECTED
2	FLASH FLOODING	SOME INUNDATION AND DAMAGE. IMMEDIATE EVACUATION MAY BE NECESSARY.
3	SEVERE FLASH FLOODING	VERY DANGEROUS FLASH FLOODING. EXTENSIVE INUNDATION AND DAMAGE. IMMEDIATE EVACUATION IS NECESSARY.
4	EXTREME FLASH FLOODING	FLASH FLOODING OF UNUSUAL OR UNPRECEDENTED MAGNITUDE. EXTREMELY DANGEROUS. MAY APPROACH OR EQUAL THE "PROBABLE MAXIMUM FLOOD".

TABLE 4. EXAMPLE OF STAGE ASSIGNMENTS FOR FLASH FLOOD CATEGORIES, THE DOONES RIVER AT MARSVILLE

STAGE (FT)	FLASH FLOOD EVENT	FLOOD CATEGORY	STAGE RANGE
	NO FLOOD	1	00.0-9.9
10 FS	FLASH FLOOD	2	10.0+
15 SFS	SEVERE FLASH FLOOD	2,3	15.0+
25 EFS	EXTREME FLASH FLOOD	2,4	25.0+

FS - FLOOD STAGE
SFS - SEVERE FLOOD STAGE
EFS - EXTREME FLOOD STAGE



FS-Flood Stage SFS-Severe Flood Stage EFS-Extreme Flood Stage

Fig. 5. FLOOD HYDROGRAPH INDICATING FLASH FLOOD CATEGORIES 1-4.

Example 3. Observed flash flood of 28 feet; forecast crest was 16 feet. A Cat 4 observed event (an extreme flood), and a Cat 3 forecast event. You score a hit for Cat 3 - you forecast a severe flash flood and one was observed. You did not, however, forecast the extreme flash flood, so an error must be computed. The extreme flash flood for Marsville begins at 25 feet. Event error is 16.0 - 25.0 = -9.0 feet. You underforecast the extreme flash flood event by 9 feet, which is true.

Example 4. Observed flash flood of 13 feet (Cat 2); forecast crest was 27 feet (Cat 4). A flash flood did occur, but the forecast extreme event did not occur. Credit a hit for the correct forecast of a flash flood; score a miss for the incorrect forecast of an extreme flash flood. One verified and the other did not. Error for the miss is 27.0 - 14.9 = 12.1 feet. You were 12.1 feet too high in stage in predicting an extreme flash flood event, when in fact a flash flood of lesser magnitude was observed.

Example 5. This last example is the really bad bust: Forecast was 8 feet, no flash flooding. Observed crest turned out to be 26 feet, an extreme flash flood event, which begins at 25 feet. Forecast error would be 8.0 - 25.0 = -17.0 feet. You underforecast the extreme flash flood event by 17 feet.

In all of the above examples verification data were derived that would provide the Agency with factual information on the flash flood program, as pertains to site-specific forecasts. Given the vast increase in the number of radio-based, small basin flood warning networks that report in real time, such a verification system takes on increasing importance. It should be emphasized that the classification of a forecast site as "flood" or "flash flood" is entirely up to the hydrologist, subject to any guidelines drawn up by the Agency. If an office is dealing with a real time reporting hydrologic network, such as IFLOWS or ALERT, for a watershed with a long flood history for which a more detailed scale of flood categories is proper (like in Table 1 for flood), there is no requirement here that flash flood categories be used just because stream cresting time is less than 6 hours. Could both flood and flash flood classification "systems" be used for the same forecast point, say, for a river gage that can crest under 6 hours on local runoff, and thus be "flashy", while for different storms reach flood crest much later due to more upstream runoff? Sure, if the Agency so desires, but I think the need for doing this at any given forecast point should be weighed carefully before a

decison is made. It is reasonable that a reach along a river could begin as a flash flood and then move into a larger scale flood at time 6 hours or later. The proposed verification system would handle this nicely - we simply move out of a flash flood category into the appropriate flood category, as dictated by observed stage, and keep on verifying.

A final consideration is this: How would these verifiable site-specific flash flood warnings relate to the non-site-specific flash flood warnings (FFW) issued for broad areas (counties, etc.) by meteorological offices, which have long been verified by another system? My recommendation is as nearly the same as that suggested by Sittner (1973), but applied to event verification.

1. A FFW for an area applies to any forecast point in that area which is specifically mentioned in the warning. This means that credit would be given for a Cat 2 event at the site (forecast point) in question, in the absence of a site-specific stage or crest prediction. But we would not allow credit for a Cat 3 flash flood event unless a crest forecast was so issued to support this.

2. If any forecast points are specifically mentioned in a FFW, it does not apply to any point not so mentioned.

3. If no forecast points are specifically mentioned in a FFW, it applies to all flash flood forecast points in the designated area. Again, Rule 1 above would apply for categorical, site-specific verification purposes.

4. A FFW is applicable to an event only if it is issued less than 12 hours prior to the occurrence of flood stage.

5. Only one FFW (the earliest) is applicable to any event. However, if the FFW is cancelled, and the event subsequently occurs, no verification credit is allowed for the earlier warning.

6. The FFW verification program maintained by meteorology would continue even with Agency implementation of a categorical, site-specific verification plan, due to the fact that it is not realistic to assume that all potential flash flood prone streams could eventually be covered by individual flood warnings.

None of the above rules for FFWs are cast in concrete.

Should the Agency decide to keep site-specific categorical verification fully independent of FFW verification (the easiest course of action), this would be just fine.

4. FORECAST/OBSERVED LEAD TIMES

The National Weather Service defines forecast lead time as "the time of issuance to the time an event occurs" (Campbell, 1985). NWS also classifies it's public service products as being routine and non-routine (NWS, 1982). By way of example, most weather forecasts are routine, whereas severe weather watches and warnings would be considered non-routine products. For the purposes of verification, we will consider the hydrology program site-specific flood forecasts (all forecasts) as being routine. While it is true that floods are unscheduled events, obviously, the RFCs exist to handle them as a matter of routine, and upon the onset of a wet cycle and flood, the stage and crest predictions, as well as other products, do become a routine office service, and often with a reasonably well set release schedule.

4.1 THE PROBLEM

Determining the lead time for a flood warning is not the simple thing it appears to be. First of all, a single time of issuance may be debated, as warnings to the public sector may take place at different times, once the warning is formulated, from different offices in the Agency. Sittner (1973), computed MFLT based on forecasts released by a meteorological (met) office, rather than on the time the forecasts left the RFC, the thought being that it is the met office that is the primary interface with the public. This is reasonable. But close examination of met offices reveals that flood warnings (as well as other types of warnings) take different paths (telephone, AFOS, weather wire, weather radio, etc.) and involve different times, depending upon the degree of emergency and other factors. By way of example, an emergency flood warning (or severe weather warning) will likely reach public officials via telephone or NAWAS, person-to-person, then subsequently travel mass media via NOAA Weather Wire as a follow up. There can be substantial differences in these release times (30 minutes, plus), so that the public warning frequently involves more than one "the". It is also fact that RFCs commonly deal with state and local agencies directly during serious flood while the river prediction is being formulated, and warning action by the public may commence immediately due to these communications alone. Of course, subsequent public bulletins may then follow as a result of RFC to met office/Service Hydrologist communication, thus

generating another "release time" for the record. It is not unusual, I suspect, for flood warning dissemination to include the efforts of one RFC, two met offices, and three "publics". Also, more and more we are seeing local computer driven warnings (the computers being both internal and external to NWS) via flood warning systems like ALERT. Here NWS may very well be involved in the warning formulation via quantitative precip forecasts or some other interface, but not be a direct part of the warning dissemination in the usual sense. We can only conclude from all the above that warning information to the public may or may not entail a single lead time number, and it would be or could be misleading to formulate a national verification system based on a single "release point" without the understanding that the forecast release time is only approximate.

Another facet of the forecast lead time problem is that the stage or crest prediction implies certain important information beyond the fact that a river will reach a specified elevation at a specified time. A forecast of "crest 35 ft. Saturday", issued Thursday, for a river at 10 feet with a flood stage of 20 feet, does not give everyone in the flood plain below 35 ft. stage two days' lead time for action. Suppose the forecast is "crest 22 ft. Saturday". It is true in this case that fewer people have more time to prepare - right? However, for verification purposes, <u>assuming the river rises as forecast,</u> both warnings would show a 48-hour lead time, which strikes me as being a number of little more than curiosity value, given no additional information. Perhaps knowing the magnitude of the rises would be valuable for interpreting the significance of lead time, <u>both</u> forecast and observed, and you can see where I am heading. Still another point to ponder, is that the traditional (?) method of computing forecast lead time is based on time of issuance to time the event (weather or flood) <u>actually occurs.</u> If the event does, in fact, take place before or after the forecast time, the computed lead time number is reduced or increased accordingly. This bothers me. Let's examine it. Seems to me that what we have been doing is computing observed lead time, not forecast lead time. We have a true "forecast" lead time here, if event is to be defined solely in terms of a single stage, only when the event occurs exactly as predicted. Otherwise, not so. For example, if the forecast is for flood (X feet) in 20 hours, and X feet is reached in 10 hours, we say the public had 10 hours to prepare for a flood level when the public expected 20. Unless you are one of those folks who decided to take action 10 hours earlier than advised, the 10 hours "actual" computed lead time figure is phony. Suppose instead, same forecast, that X feet is reached in 30 hours. We say the public

had 30 hours to prepare for the forecast flood level when the public actually prepared 20 hours. I suggest that in both cases forecast lead time was 20; not 10 or 30. It's just that a 10 hour error one way hurts a lot more, potentially, than the other way. Let's not state for the record a forecast lead time, observe a different lead time, and store in the verification database observed as forecast. Yes, I know, it is all a "matter of definition". Well, I prefer a different definition of forecast lead time for categorical flood forecast verification that I believe makes more sense for what the Agency is trying to learn from all the numbers hydrology generates.

At this point in our discussion of lead time one could point out that the issues raised above, as was true for certain defined flood levels, also involve user response/customer satisfaction, and one might question relevance to forecast verification, as opposed to forecast or service evaluation. It is my feeling that these questions and issues are more a matter of semantics and definition than substance. The facts are, and it should be clear, we are dealing with the flood as an event, something that has both time and magnitude, and that has both service and historical significance - always . From these facts we have drawn justification to classify floods, and now justify a reasonabale measure of "time to predicted event", which we call forecast lead time for the flood in question. We similarly look at the observed lead time

It is fair to say, I believe, that flood forecast lead time is a verification parameter that involves uncertainty at both ends of the measure. There is, or can be, a significant span of time on the release end, and a time span of user value, within a category of flood, on the forecast end. It is not operationally practical to track, as a matter of office verification routine, each dissemination path, time-wise, nor is it smart to assume that the time of an observed stage or crest reliably reflects flood preparation time. Should there occasionally be Agency interest in exact warning lead times in the flood forecast service, which may vary dramatically from site to site and rise to rise, such information should be available on a case by case basis through administrative investigation. For any given serious flood, the Agency can determine both stage or crest error and precise lead time to whomever as additional information required for report and assessment. For Agency verification purposes, however, we must be satisfied with less detail, and instead direct our attention to compiling data that permits a comprehensive determination of the ability of the forecaster to predict floods of specified magnitude, to include information on when the predicted event occurred.

4.2 A SOLUTION

We now conclude that it is both reasonable and sufficient to view forecast lead time as a value in hours computed from some release time span indexed by one office, and the prediction of a category of flood indexed by forecast stage or crest. This is different from how flood forecast lead time has been viewed in the past, but this report deals with different things, like floods as events. A verified flood event is one in which there occurs, during the forecast period, an observed stage or crest in the same category of flood as was predicted. If the forecast event proves to be a bust (does not verify) there is, of course, no lead time to compute. If a forecast category proves to be a "hit" (does verify), there are two lead times that can be computed: (1) Forecast lead time, and (2) Observed lead time.

"Time of Issuance" (TI) is the time the forecast leaves the primary office to the primary recipient (public). Pick an office - it may be the RFC or some met office, and can vary from forecast point to forecast point, depending upon local dissemination practice. All we are doing here is allowing the latitude to field select forecast release point for verification purposes. However, if the Agency prefers that, for the sake of uniformity, a single office like the WSFO be deemed the "primary" office for all public releases of flood, as noted by AFOS statement, insofar as logged verification time goes, this would be acceptable. There is, however, some loss of actual lead time involved with this practice, but one could argue it is trivial, in most cases, considering our method of verifying flood predictions.

"Forecast Event Time" (FET) is the time of the forecast stage or crest. Forecast lead time is computed to this time. Using this single forecast stage as the index to flood as a predicted event may be largely justified as follows:

1. A category of flood represents a river in a specified range of stage. Any stage, forecast or observed, within this range is equally significant for categorical verification purposes.

2. A category of flood represents a range of stage over a span of time. In other words, a flood event is not a stage at a time. It is many stages over some time. Any time in the event, forecast or observed is equally significant for categorical verification purposes.

Items 1 and 2 above, plus the earlier thoughts in Section 4.1 on "time of actual occurrence", form a framework of justification that should serve as a solid basis to tie forecast lead time directly into the time of forecast stage, for the purpose of categorical verification. Sure, at the time of the forecast stage, the river may very well be higher or lower, but there is likely an equal chance of either, and it is really the forecast stage that warns of the event and drives resulting public action.

The following rule and definition apply to lead time computation, forecast or observed, as appropriate:

1. Any forecast, obviously, must state a time the stage (event) is expected to occur, and this forecast time may be exact, or cover a period. If a period (block time) is used, a specific time for verification will be assigned somewhat as folows: (a) a day (24 hours) - 1200 hours assigned. Commonly used as "crest xx ft. tomorrow", (b) AM (12 hours) - 0600 hours assigned. Commonly used as "rise to xx ft. AM Tuesday", (c) PM (12 hours) - 1800 hours, (d) early morning (4 hours) - 0600, (e) mid-morning (4 hours) - 0900, etc., etc. In all cases for implied forecast periods, a convention must be established whereby the "block time" is bounded by generally accepted clock times. For any block, stated or implied, the forecast stage time, for categorical verification purposes, will be the average time for the block. This is FET.

While a specific time for stage will always be used to anchor the forecast event, (FET), it is still the forecast period, as issued to the public, that will be used to verify against via observed hydrograph data. Should, however, the public forecast be for a stage to occur at some exact time, e.g., "crest 25 ft. (say, a Cat 3 event) 11 PM tonight", it is only the observed stage at 11 PM that will be used to determine the observed category. I have a hunch that there are precious few such precise forecasts being issued nationwide.

2. Forecast Lead Time (FLT) for a verified flood event is the difference, in hours, between time of issuance and the time of forecast stage or crest, i.e., FLT = FET - TI.

3. The "zero lead time warning" occurs when a river rises to some degree (category) of flood not predicted, but a public statement (warning) is still issued regarding current or forecast river stage conditions, a common situation. If the river continues to rise, and does rise to a higher degree of flood, only this "greater" flood is subject to verification. If

this does not occur, the "zero lead time warning" will be noted in the database as a "non-forecast" event. As a statistic, per se, we make no distinction between "zero warning" and "no forecast".

4. Observed Lead Time (OLT) for a verified flood event is the difference, in hours, between time of issuance and the time the event occurred, i.e., OLT = OET - TI, where OET is Observed Event Time, to be discussed next. The observed event is determined as follows:

For a river on the rise, the forecast period starts at time TS and ends at time TE. Let TL be the time the river reaches the lower threshold (stage) of the predicted category of flood, and let TH be the time the river reaches crest within the predicted category, or the higher threshold of the predicted category, whichever occurs. TI is the time of forecast issue. Obviously, if TL < TI, we have the "no forecast" or "zero lead time" event discussed earlier. The observed event, for lead time computation, starts at time TL, ends at TH, and includes observed or estimated observed stage in the predictged category for the forecast period TS to TE (See Figure 6). Neither TL nor TH are required to occur within forecast period TS to TE. Let OET = (TL + TH)/2. TS and TE define the "window" for verification. The following two rules, first off, address the problem of a river receding through forecast level when the forecast was otherwise:

Rule 1. If the TH < TL, and the crest occurred at a higher than predicted flood category, the river is falling through category, and OET is automatically set to zero. Yes, you did correctly predict the river to be at a certain flood level (category) during period TS to TE, but, unfortunately, the crest occurred earlier and higher (categorically speaking) than forecast, and we are, for flood prediction purposes, in the business of forecasting river rises. The verification record notes one category of flood predicted, zero forecast lead time (FLT), and zero observed lead time (OLT). This is probably the best way to handle this (unusual?) case, since we cannot compute event error - the forecast and observed flood categories "match". However, should the river be in some higher or lower flood category during the forecast period, event error for the "busted" forecast can be determined.

Rule 2. If the river is below the predicted flood category, and recedes through more than one lower category by time TE, it is the lowest category that will be used to compute event error.

From here on, it gets easy - we forecast a rising river, and, thank goodness, it does rise. Now we stand a good chance of predicting the flood event.

Rule 3. The crest of the observed hydrograph, for verification (not forecast) purposes, will be considered the entire segment of the observed hydrograph above level TL during period TS to TE, which includes rising and/or falling stages above TL. Thus, if the river crests within the predicted category, after TI, so long as at least the falling limb of the crest segment occurs within forecast period TS-TE, the prediction verifies and lead time will be computed. This is the only case where a receding hydrograph verifies prediction and lead time is computed. What we are doing here is acknowledging the fact that rivers do crest as forecast, but often earlier than forecast. There is no similar problem for a river cresting after the forecast period, as will be evident from Figure 6.

Figure 6 illustrates a few FLT and OLT computations. In all four examples (A-D), FLT has the same value, but obsrved rise segments vary. The flood category is whatever is being verified. Case A brings the rise into flood before TS, and into a higher degree of flood after TE. OLT computes to be something slightly less than FLT. Case B indicates rise into flood after TS, with crest occurring within the flood category before TE. OLT computes to a number substantially less than FLT. If we were to shift the crest, TH, to the left so that TI < TH < TS, still within the forecast category of flood and with the falling limb passing through period TS-TE above level TL, we would have the Rule 3 circumstance. Shift TH to the right so that TH > TE, if the rising limb passes through TS-TE, you verify the prediction. If crest occurs after TE but at a higher level (degree) of flood, we hope that the hydrologist has a later forecast out to cover it. If not, a "no forecast" flood gets credited to your verification account for that later and "higher" flood event. Case C rises into flood before TS and to a higher degree of flood before TE. Again, OLT turns out to be some number considerable less than FLT. Case D is, categorically speaking, the "perfect" forecast: TL equals TS, TH equals TE, and OLT therefore equals FLT.

It should be evident from Figure 6 that the observed lead time, as we define and compute it, is a function of the shape and slope of the rising limb of the hydrograph, all of which makes the kind of sense it should. Given a reasonably sharp rise, time TS - TE is likely short, and FLT and OLT would be brief and undoubtedly close in value. That's the way it is in

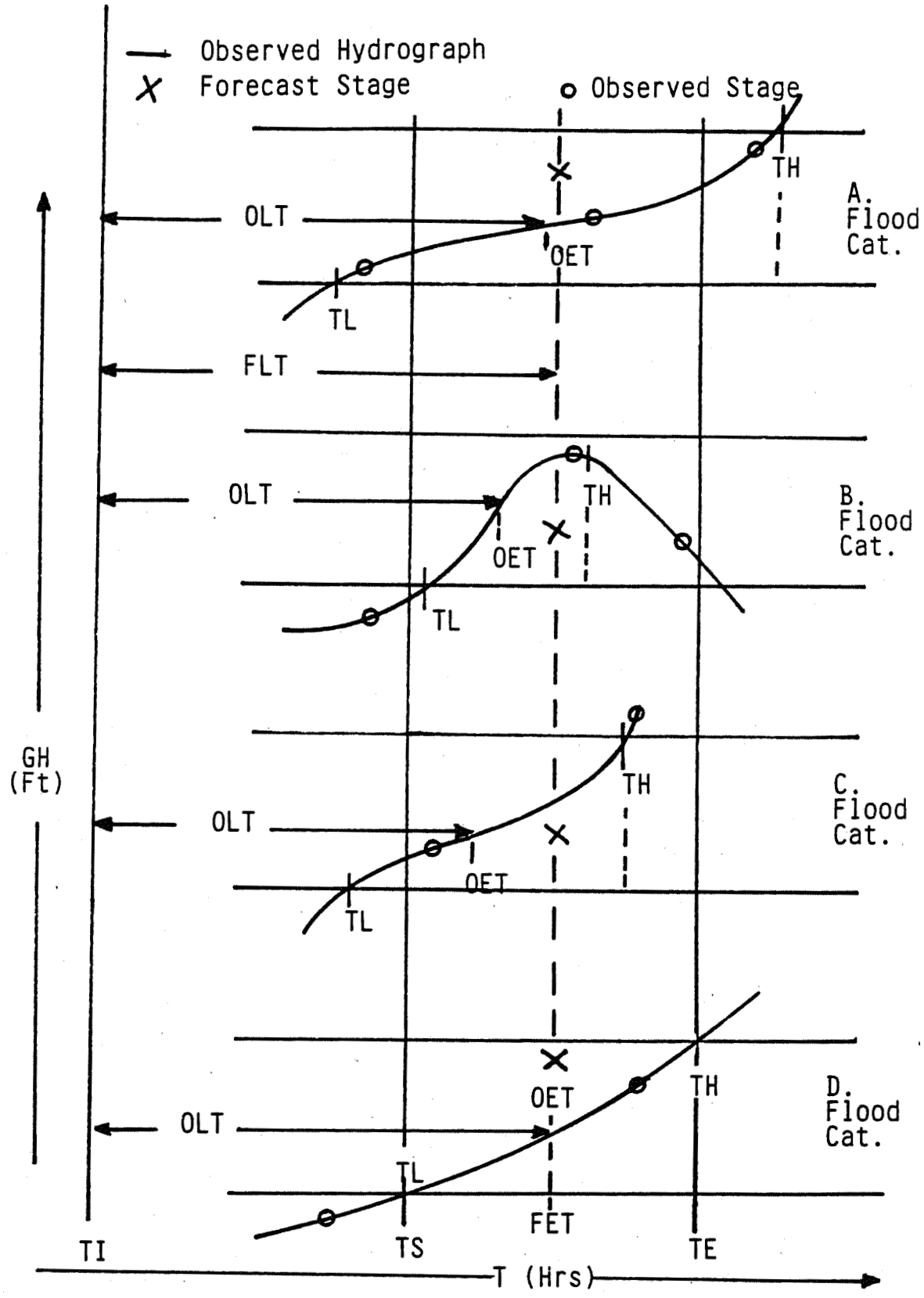


Fig. 6. EXAMPLES A-D OF LEAD TIME COMPUTATION

forecasting sudden river rises - little time, and it's "hit or miss" real quick, whether one is viewing specific flood stage or specific flood category. However, slower rising streams could very well exhibit substantially different values of FLT and OLT, which is important information to the Agency. A rejected thought: Why not require that TL fall within the TS to TE verification window, and any exit to a higher flood category, TH, be later than TE? This would be far too restrictive, as we would be in essence stipulating that, in order to "hit a forecast," one must predict the time flood category is reached, and also require that the flood category not be exceeded during forecast period. This is not what our verification is about. Line hydrologists are not in the business of predicting flood categories, per se - we simply use their stage and crest forecasts to verify our ability to predict the time and magnitude of flood events, for reasons thoroughly covered earlier in this report. Also, such a tight verification requirement could result in truly excellent forecasts consistently being scored as "misses," even if justified, just because the hydrologist was off by a few minutes or so in stage prediction time.

So what all this boils down to then is that we have decided to first determine whether or not a flood of specified degree, as dictated by the forecast stage, did occur during the forecast period. If so, we view the predicted time of event as being the stage forecast time, which is always average time for the forecast period. The observed time of event is similarly an average of the time for which the river was rising within the predicted flood category. Stated differently, the forecast stage is an index to a flood of predetermined degree, and the time of the forecast stage is the index to approximately when the flood should occur. We then observe the rise as an event, and time index its occurrence. Consequenly, we can say "yes, no" as to whether or not we forecast the flood event, and also say something roughly about how much warning time was given to, and then actually available for, preparation activites in the flood plain. I suggest this is a reasonable approach to "framing out" the flood prediction service of the Agency. Categorically speaking, it should nicely answer certain questions.

Figure 7 illustrates lead time computations for a hypothetical river. Four predictions were issued during the Cat 6 record flood. The hydrograph is defined by 6-hour ordinates for a rise to near 51 feet. Forecast (1) was issued at time 24 hours (TI) for a rise to 18 feet (a Cat 2 event) during the period 30-54 hours (TS-TE) so that FET = 42. It is clear from

THE SWEET RIVER AT DUMP

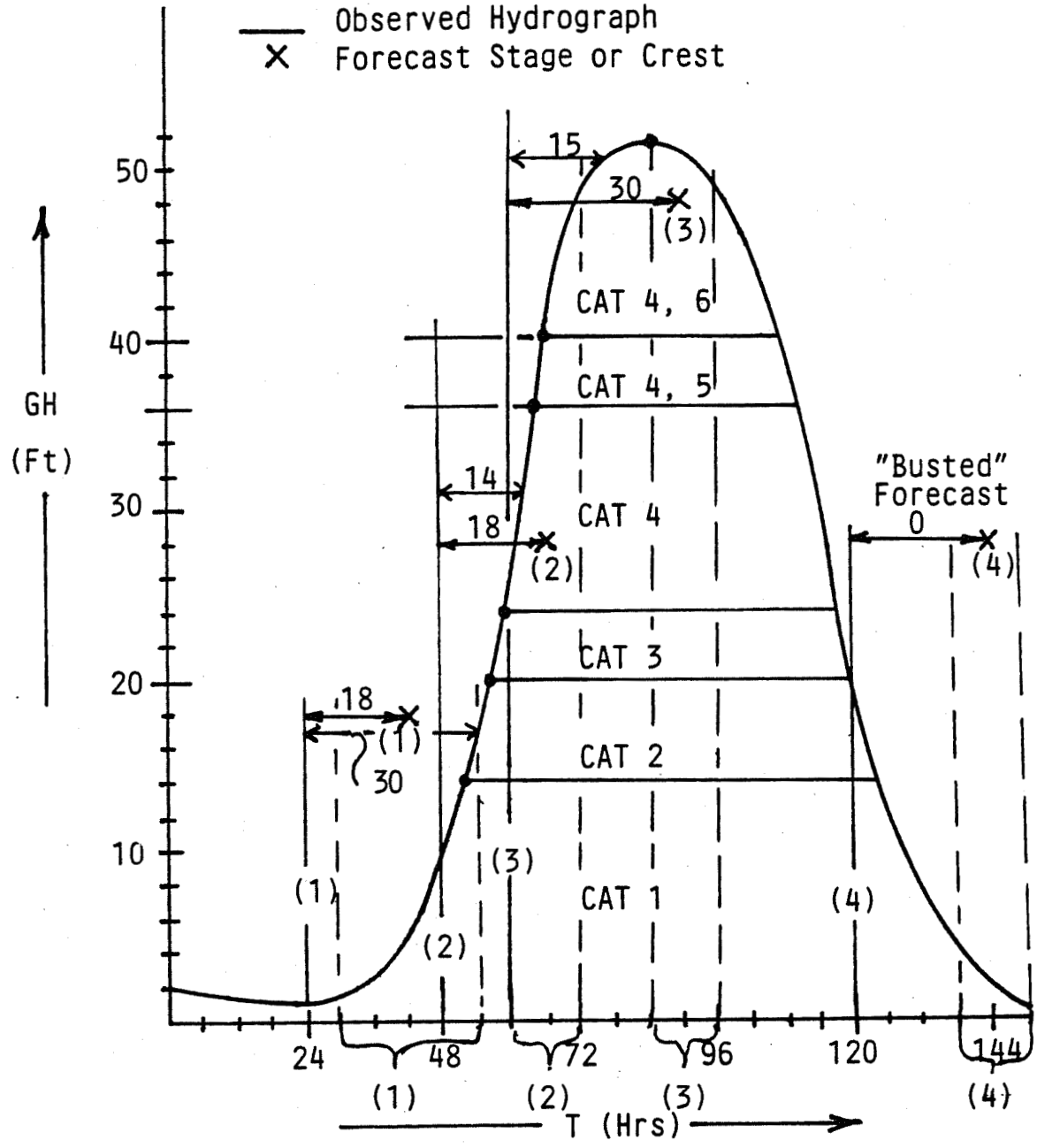


FIG. 7. FLOOD HYDROGRAPH INDICATING LEAD TIMES FOR PREDICTIONS (1) THROUGH (4)

the observed hydrograph that Cat 2 flood did occur during the forecast period, with TL ccurring at 51 hours and TH at 57 hours. Therefore, OET = 54 hours. FLT = FET - TI = 42 - 24 = 18 hours. OLT = OET - TI = 54 - 24 = 30 hours. The verification record thus notes one minor flood predicted for the Sweet River at Dump with a forecast lead time of 18 hours and an observed lead time of 30 hours. It keeps raining, and Forecast (2) is issued at time 48 hours (TI) for a rise to 28 feet (a Cat 4 event) between time 60 and 72 hours (TS-TE), so FET = 66 hours. Indeed, the river did rise to the Cat 4 level - and kept going. We see from the hydrograph that TL = 60 hours and TH = 64 hours, so OET = 62 hurs. FLT = FET - TI = 66 - 48 = 18 hours. OLT = OET - TI = 62 - 48 = 14 hours. We have predicted one major flood with a forecast lead time of 18 hours and observed lead time of 14 hours. The storm continues and at time 60 hours (TI) the hydrologist issues Forecast (3), a prediction of crest 48 feet (a Cat 6 event) duing the period of 84 to 96 hours (TS-TE), so FET = 90 hours. We see from the hydrograph that the river did crest at record level, with TL = 66 hours and TH = 84 hours. So OET = 75 hours, and thus FLT = 90 - 60 = 30 hours, and OLT = 75 - 60 = 15 hours. One record flood predicted with forecast lead time of 30 hours, an observed lead time half that figure. Suppose Forecast (2) called instead for record flood of 48 feet at time 90 hours. We said earlier in the report, that you also get credit for predicting major flood - two "hits". The Cat 4 threshold level starts at 24 feet (TL) and tops, for our purposes now, at a stage of 51 feet (TH). Compute OET from this set of numbers and then OLT for the Cat 4 flood. The OET and OLT for Cat 6 flood would be computed as in Forecast (3) above. FLT is the same for both Cat 4 and Cat 6 events - that prediction, after all, called for both major flood and record flood, and it was the only such forecast issued for major flood and above. It would turn out that OLT (Cat 4) < OLT (Cat 6) < FLT, which makes sense. Now, if the forecaster would like for the OLT (CAt 4) to come in closer to FLT, I would suggest that he issue an earlier prediction for Cat 4 level flooding (if possible). If the river is already in Cat 4 flood, not predicted, when the Cat 6 forecast is issued, then in this case, credit would not be given for a Cat 4 prediction along with the Cat 6 "hit". Notice how all this encourages the hydrologist to issue stage forecasts for each level (category) of flood in order to obtain maximum credit in lead time? This is good. Results, I should think, in a better public service. However, let us emphasize the fact that the object of river forecasting is to provide public interests with as much lead time to flood as our science and circumstance permit; not see how close we can get warning times to compare. There should be little satisfaction in, say, a major flood prediction with a one

hour forecast lead time and a one hour observed lead time if a much longer warning time was possible. But the matter of ideal or maximum possible lead time is beyond the scope and intent of this report. We are only concerned with "what was forecast versus what was observed," within the context of flood categories, as defined. This, so it seems to me, is a logical first step in the design of Agency verificaion.

Figure 7 also indicates a Forecast (4). What happened is this: During recession, as the stream was going down through Cat 2 flood level, word came of a dam break. So the hydrologist at time 120 hours jumped with a forecast sharp turnaround in the river to 28 feet (Cat 4 flood) at time 144 hours. However, the dam did not fail. A major flood event error gets computed, based on observed recession into the no flood category (ouch!), and, of course, there is no lead time to compute.

What about the bracket forecast for, say, two categories of flood? Any problem computing FLT and OLT? No, not at all. FLT is, of course, the same for both categories - if both verify - and OET is computed for each event, resulting in two OLT values.

Could it be that this verification system encourages the hydrologist to use as lengthy a forecast period as possible? I doubt it. Business would continue as always. Neither the public nor the Agency would accept broad forecast periods that render the service worthless. There are always pressures (that increase with magnitude of flood) to narrow down the time of predicted stage. Also, with this verification approach, "missing the flood" in time by just a few hours also means that the catgegorical stage error is apt to be smaller, so the river forecaster is not stung by some unforgivable prediction error just because he failed to "add an hour or two" to what he really thought would be the time of flood.

It would be easy, at first glance, to conclude that the proposed method of computing flood lead times is rather crude. I do not think so. It certainly is more precise than the meteorologist's verification of categorical weather events, and clearly is compatible in verification approach, which is advantageous to the Agency. To emphasize this argument, I would point out that certain forecast weather events are verified by periods: "Today" (12Z-00Z), "Tonight" (00Z-12Z), and "Tomorrow" (12Z-00Z). If the predicted event, like rain or snow, occurs at selected points any time within the forecast period for the area (zone) in question, the forecast is a "hit" and lead time is an inferred value from the forecast block time.

I believe that categorical forecast/observed lead time is as good a time measure as we have in hydrology considering the fact that a flood is an event transpiring over a substantial length of time. I suggest the proposed definitions of forecast lead time and observed lead time for hydrology are "both reasonable and sufficient" when dealing with floods as event phenomena.

4.3 DATA REQUIREMENTS

It is likely true that any meaningful verification in hydrology requires a rather detailed description of the observed hydrograph. With more and more stream gages becoming automated, and real-time interrogatable, the hydrologist is obtaining river level data in quanitities unimagined years ago. Makes for better predictions, and verification. Of course, stream gages malfunction, and sometime down-right quit functioning, leaving one with a stage estimating problem on his hands - for prediction and verification. Sittner (1979) developed an algorithm that should nicely estimate missing data, to the extent that limited observations permit, to include peak stages, if necessary. It is a rather sophisticated procedure that has the potential to make verification possible at any forecast point, if just "sketchy" data are available. It cannot, however, manufacture data. Sittner's algorithm would be a necessary part of the verification software. If reasonable stage estimates are not possible, then verification is not possible.

If verification is not possible for a rise at some forecast site, then tabulation must simply note this fact, and record made for "history". Perhaps such a history would prove valuable for documenting gage non-performance, an added bonus to maintaining verification record.

5. A REVIEW

The issue of forecast accuracy is addressed by examining our ability to predict river rises of specified magnitude. We also compute measures of how far in advance the rise was predicted and when the rise occurred. If this report so far is stripped of all the verbage necessary to review, argue the categorical/event case, and illustrate procedure, what is left are relatively few words outlining a fairly simple approach to a very complex problem in hydrology. By way of definition and rule, for review, the flood forecast verification program gets driven this way:

1. A flood is an event classified by stage, with a significance categorized by the magnitude of the rise.

2. A flood forecast event error is the difference in feet between forecast crest and the observed event, whereby the error represents the minimum stage, plus or minus, required to change the forecast such that the event would have been correctly predicted.

3. Either forecast river stage or forecast river crest is a suitable measure of the degree or magnitude of flooding predicted.

4. A verification event occurs whenever the river rises from one flood category into another. If, during the course of a rise, a river passes through one or more levels of flood for which no forecast was issued, only the highest degree of flood will be subject to verification, whether a final crest prediction is made or not.

5. Only the initial (first) forecast of the category of flood anticipated counts for verification. Subsequent river forecasts for stage within this same category are ignored until the river recedes to a lower level, lesser degree (category) of flood.

6. Every forecast that predicts a new (higher) level (category) of flood is verified, even if subsequent forecasts downgrade the earlier prediction, and even if the flood occurs earlier or later than predicted.

7. In the case of a bracket forecast, the highest or lowest stage values, whichever is appropriate, specified by the bracket, will be used to compute flood event error.

8. A non-stage specific forecast, stated categorically or otherwise, will not be verified.

9. Any river rise, regardless of how small, that brings river level into a new category of flood, will be verified.

10. Site-specific flash flood warnings will be verified similarly to flood warnings for the larger rivers.

11. Credit will be given for site-specific flash flood warnings based on area warnings, according to the stipulations

outlined herein, if the Agency so desires.

12. A river already at the warning flood level (category) when the "forecast" is issued, is a "zero lead time warning", and will be verified as an event with no forecast issued.

13. A verified flood event is one in which there occurs during the forecast period, an observed or estimated stage or crest in the same category of flood as was predicted.

14. Categorical Forecast Lead Time for verified events is the difference, in hours, between time of forecast issuance and the time of forecast stage or crest.

15. Observed Lead Time for verified events is the difference, in hours, between time of forecast issuance and the time the event occurred.

And how would, in practice, the verification plan be implemented? I envision a man/machine mix, more machine than man, however: no total automation. Any verification effort brings with it an overhead that cannot be avoided - only minimized - and I am fearful of <u>any</u> piece of software that moves data blindly into <u>any</u> verification algorithm just so the workforce can enjoy hands-off production. It is dangerous business, and the outcome would undoubtedly be badly contaminated statistics, given the kind of river information errors all hydrologists must contend with. Also, from a personal philosophy viewpoint, I am of the opinion that river forecaster need a more or less routine hands-on exposure to verification for the "how are we doing" educational value, as opposed to a once yearly dump of "how we did". I recognize the fact that some hydrologists advocate an "out of mind" approach to verification, lest the forecaster develop some kind of bad forecasting habits in order to verify well. I do not agree with this view. I can't imagine what kinds of habits could develop that lead to good verification but bad flood forecasts. As far as I am concerned, any river forecaster should be armed with an appreciation that "X feet tomorrow" means a flood of "Y magnitude", and a change in that prediction by some incremental value of stage may result in a predicted flood of different magnitude. I think the hydrology profession in the National Weather Service should have little complaint about a reasonable verification workload.

Finally, is the proposed verification plan objective? I think it is both simple (?) and objective. "Yes it happened, no

it didn't," and if yes, "when", is about as simple and objective as life on earth gets.

6. THE STATISTICS AND VERIFICATION SUMMARIES

We have developed procedure that generates raw data in the following form: (1) A flood - no (Cat 1); yes (Cat 2-6 flood, Cat 2-4 flash flood). (2) Was the flood predicted - no, yes. (3) If no, the stage error for the missed event. (4) If yes, the forecast and observed warning times for the predicted event. (5) The number and kind of flood events for which no forecast was issued. This is priceless information to management. Now what, besides tabulation, can be derived from these data that measure service? Let's borrow some ideas from meteorology. The mathematical formulations for a variety of common verification statistics are defined in the National Verification Plan (NWS, 1982), and also appear in other NWS verification publications, so it should be unnecessary to repeat such in this report. The following verification scores could prove highly useful in evaluating the flood forecasting service:

1. Percent Correct (PC) - the fraction of the time a correct flood forecast was made, regardless of category, or by category, expressed in percent.

2. Bias by Category (BIAS) - measures the tendency to overforecast (BIAS>1) or underforecast (BIAS<1) a particular category of flood. A BIAS of one indicates no overforecasting or underforecasting the occurrence of the event.

3. False Alarm Ratio (FAR) - the fraction of the forecasts for flood events that did not verify. FAR is a measure of overwarning, and would be computed for each flood category.

4. Mean Algebraic Error (ME) - in terms of river stage, indicates whether the forecasts for each category of flood, were, overall, too high or too low, and how much.

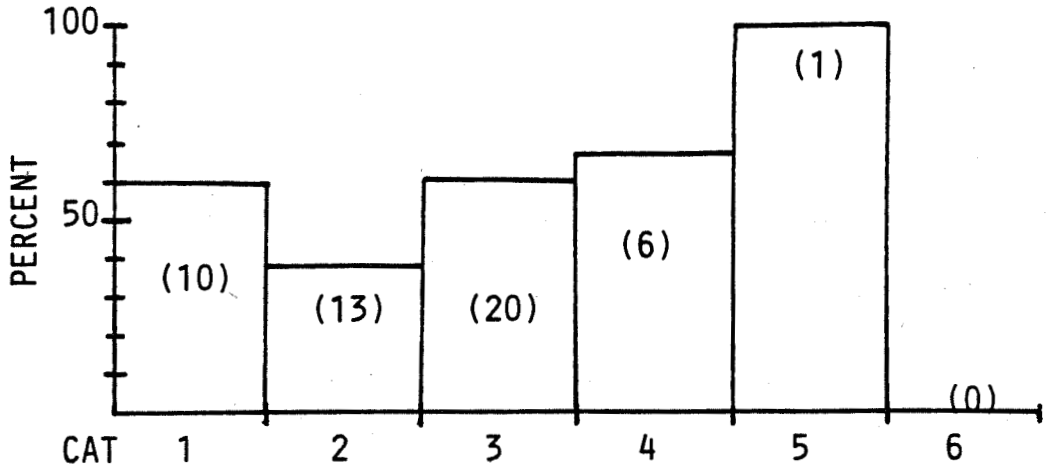
5. Mean Absolute Error (MAE) - in terms of river stage, measures the error in forecasting a category of flood, without regard to sign.

There are other statistics, of course, that could be considered, like CSI (Critical Success Index), but I question the need or utility of such numbers for flood forecast evaluation. However, the final decision regarding which statistics to compute rests, of course, with National Weather

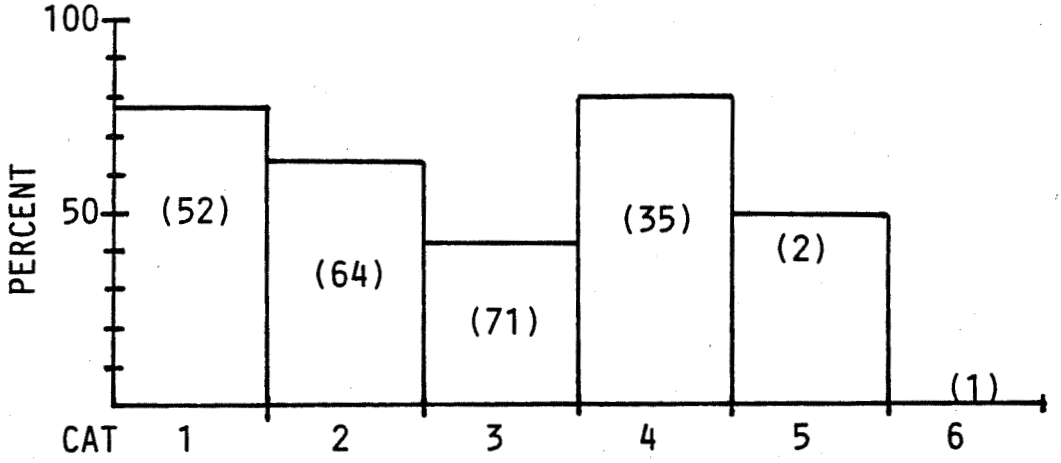
Service Headquarters, according to the information goals set by Hydrology. This author has no strong feelings about such matters, as the primary intent of this report is to develop a verification procedure for the flood prediction service, and illustrate its value in summarizing performance.

Given the kinds of data the proposed verification system generates, and the above suggested statistics, we can think of numerous kinds of verification summaries that would be possible. However, in the interest of "time and space", only a few will be illustrated. I will pattern the examples somewhat after conventional meteorological summaries. Figure 8 is a histogram of rises along the Trinity River in Texas for a hypothetical year, with grouped events according to flood category. The actual (observed) number of events in each category is noted within parentheses. In this summary and all verification summaries in this report, the numbers are not real, but were created only as reasonable values to provide illustration. Detailed discussion of Figure 8, and all subsequent Tables and Figures, is typically not necessary, as the information presented in self-evident. For Dallas, in Figure 8, we see that one near record flood (Cat 5) occurred, and the event was correctly predicted. A similar graph for other individual forecast stations within the Trinity Basin could, of course, be drawn. Histogram (B) is for the entire Trinity, all forecast points. There were 52 no-flood rises, and 78% were predicted. At the other extreme, there were two near record flood events, one (50%) was predicted, and there was one record event (Cat 6), not predicted.

Figure 9 looks at site-specific flash flood events during the year. We see that Grand Prairie had 8 out of 31 rises forecast as no flood within the Trinity Basin; 88% were correctly predicted, a number fairly close to the percentage for the entire Trinity watershed. Within the Trinity, there were three severe flash flood events; only one was predicted, and that forecast was issued at Grand Prairie. Good information here? I do think so! Figure 10 is a summary of flood events for the entire West Gulf RFC area of responsibility, the value of which is obvious. Once this kind of data becomes available, year after year, certain trends should appear, whether one looks at an individual forecast point, a particular river system, or the total area of responsibility for a selected office. For sure, we acquire a "bird's eye view" of our ability to predict the magnitude flood, "here, there, and most everywhere". Figure 11 scopes out the same information for an entire Weather Service Region. Should be of interest to a Regional Director. The same histogram could be generated for all Regions combined, of

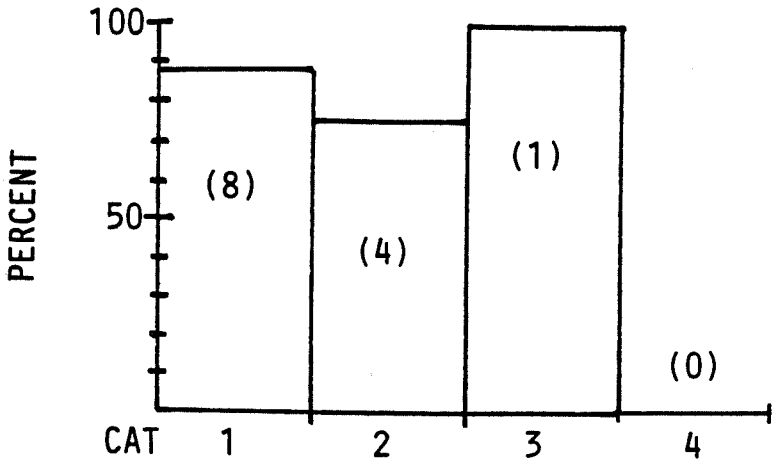


(A) THE TRINITY RIVER AT DALLAS

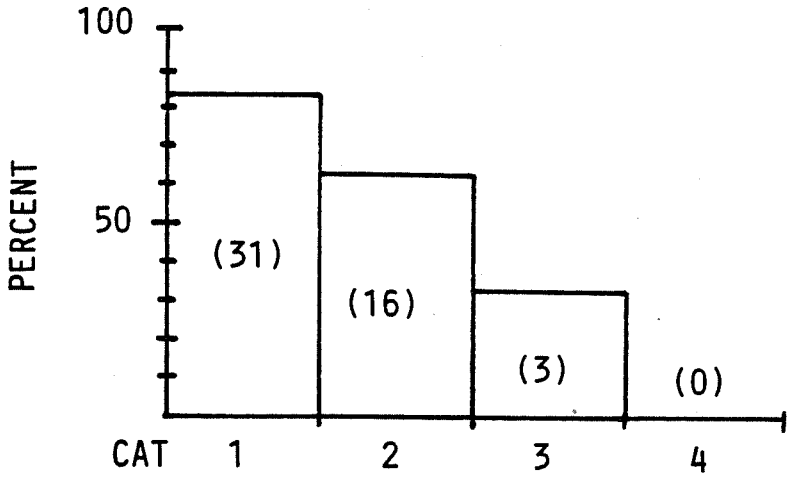


(B) THE TRINITY RIVER SYSTEM

FIG. 8. PERCENTAGE OF FLOODS CORRECTLY PREDICTED IN ONE YEAR. DALLAS (A), AND ALL FORECAST POINTS ALONG THE TRINITY RIVER SYSTEM (B). ACTUAL NUMBER OF EVENTS IN PARENTHESES, FLOOD CATEGORIES 1-6.



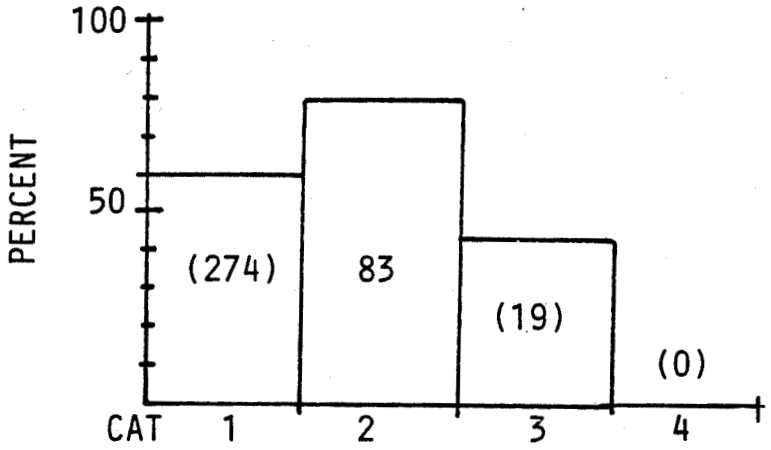
(A) JOHNSON CK NEAR GRAND PRAIRIE



(B) THE TRINITY RIVER SYSTEM

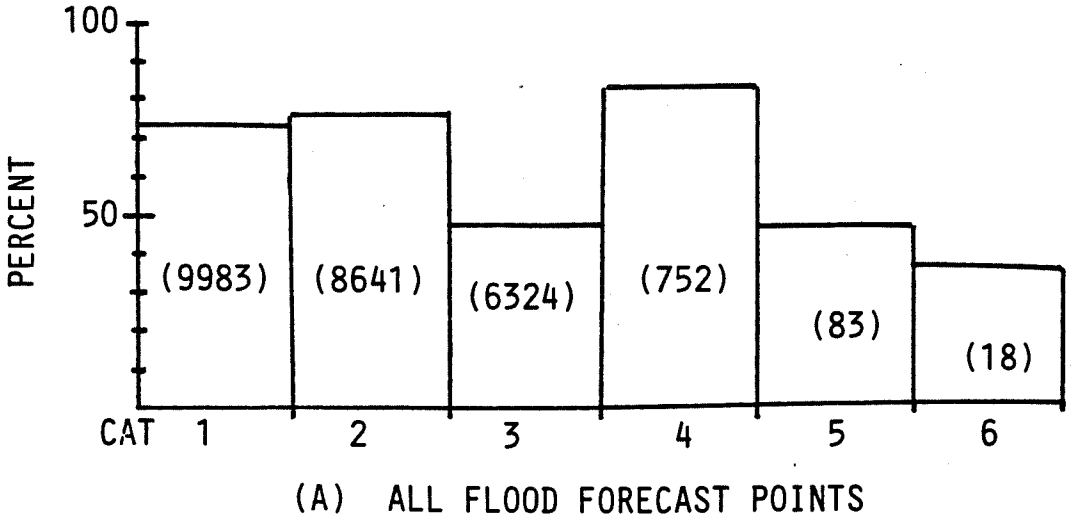
FIG. 9. PERCENTAGE OF FLASH FLOODS CORRECTLY PREDICTED IN ONE YEAR. GRAND PRAIRIE (A), AND ALL FLASH FLOOD FORECAST POINTS ALONG THE TRINITY RIVER SYSTEM (B). ACTUAL NUMBER OF EVENTS IN PARENTHESES, FLASH FLOOD CATEGORIES 1-4

(A) ALL FLOOD FORECAST POINTS

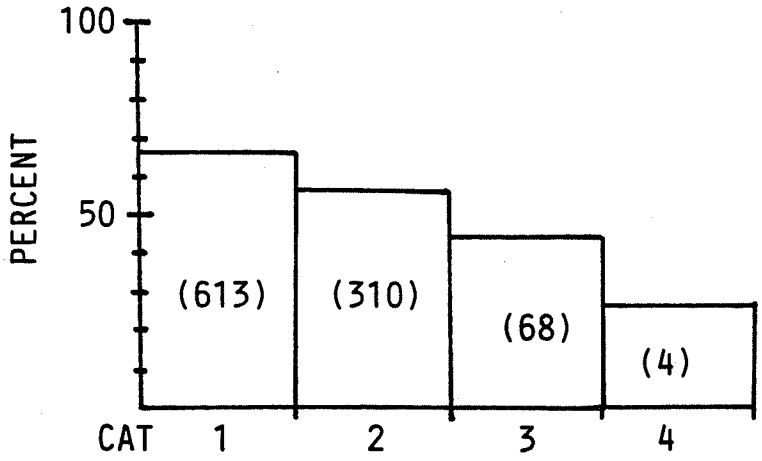


(B) ALL FLASH FLOOD FORECAST POINTS

FIG. 10. PERCENTAGE OF FLOODS (A) AND FLASH FLOODS (B) BY CATEGORY CORRECTLY PREDICTED IN ONE YEAR WITHIN THE WEST GULF RFC TOTAL DRAINAGE AREA OF RESPONSIBILITY. ACTUAL NUMBER OF EVENTS IN PARENTHESES.



(A) ALL FLOOD FORECAST POINTS



(B) ALL FLASH FLOOD FORECAST POINTS

FIG. 11. PERCENTAGE OF FLOODS (A) AND FLASH FLOODS (B) BY CATEGORY CORRECTLY PREDICTED IN ONE YEAR WITHIN THE SOUTHERN REGION. ACTUAL NUMBER OF EVENTS IN PARENTHESES.

course.

Figure 12 graphs forecast lead time for major floods, occurring in some hypothetical year, for the Trinity River at Dallas, and the entire river system. We could also draw, superimposed, the observed lead times for comparison, and undoubtedly would in practice. There is no relation between number of events in Figures 12 plus and earlier graphs - there could be, but there is not. All figures and tables are only illustrative. Saves time. We could also present graphs of lead time for the other levels of flood (Cat 1, 2, 3, 5, 6), but one serves the purpose. The zero lead time warning/no forecast event, of course, falls within the 0-0 time block. Otherwise, lead times are 0-3 hours, 3-6 hours, etc. Track this information over years, and one might see evidence of the improvement or erosion in forecast lead time. For national verification, standard lead time blocks would need to be decided upon. Figure 13 displays forecast lead time for flash flood points, and in this case Grand Prairie, followed by summary for the entire Trinity. These are lead times for site-specific flash flood forecast stations only. Only Cat 2 flash floods are included, and a separate figure would be required to display lead times for the Cat 1 (no flash flood), Cat 3 (severe flash flood), or Cat 4 (extreme flash flood) events. Both Figures 12 and 13 could also be developed, obviously, for RFC-wide drainage or Regional areas, to similarly display lead times.

Figure 14 reveals the forecast error (stage) in predicting floods along the Brazos River in Texas for one year. These event errors for the "busted" forecasts were averaged and the graph then drawn. Construct such a chart for all forecast points and all river systems, and the RFC develops a clear picture of the typical stage error in predicting floods of various magnitude. The change in event forecast error over time might well serve as evidence that improved procedure or technology is resulting in better forecasts to the public. And, I should think, if an office is doing a better job in predicting flood magnitude, it is also likely generating better stage forecasts. Naturally, a Figure 14 could be drawn for flash flood events as well.

Table 5 is powerful information if you run the National Weather Service - or some big piece of it. Here we summarize Cat 4 major flood events for some year. The same information could be and would be generated for other categories of flood (Cat 1, 2, 3, 5, 6), and for flash flood (Cat 1, 2, 3, 4). The event stage error is for the busted forecasts, while event lead time is for the "good" predictions. The zero lead time/no

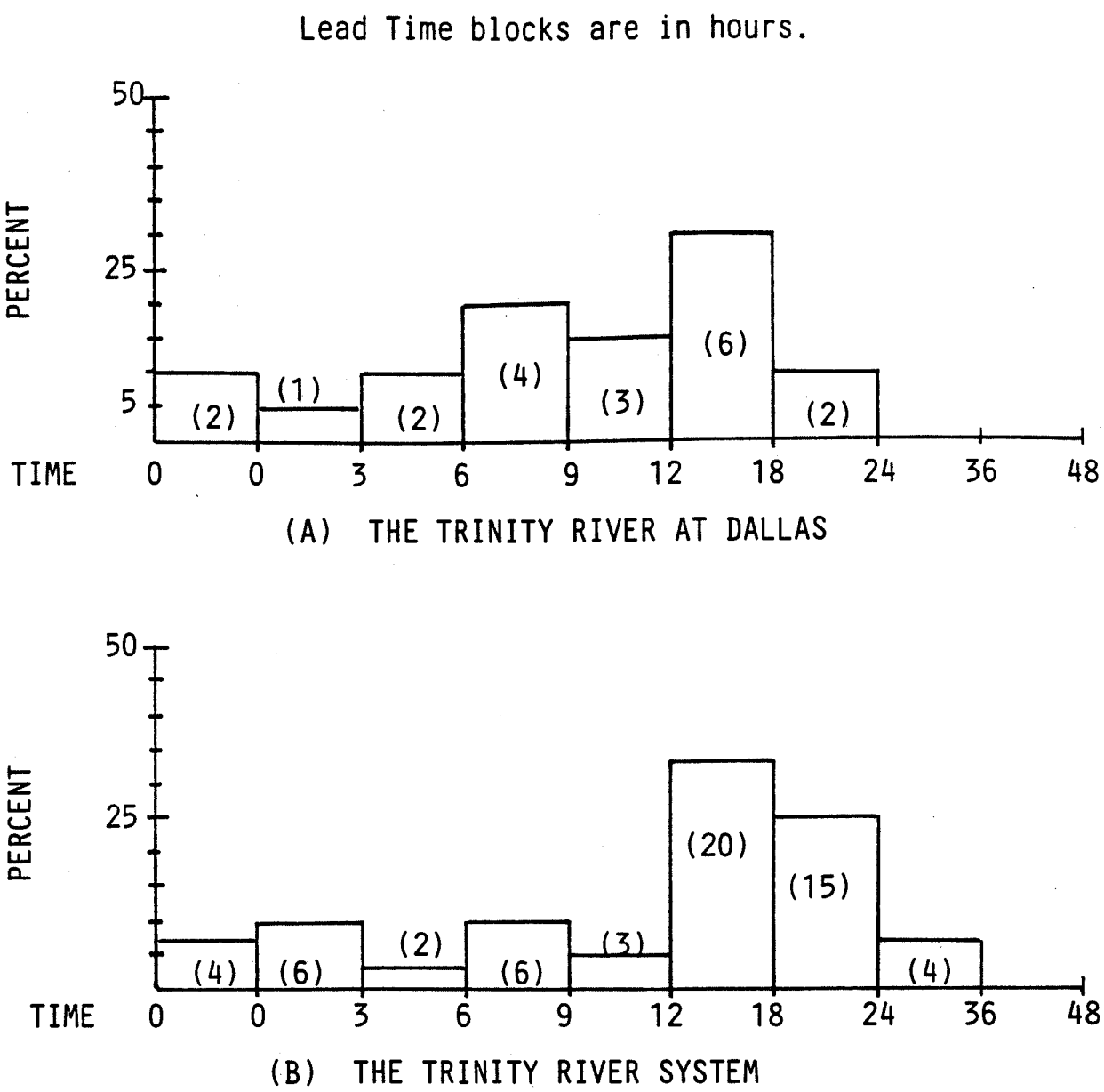
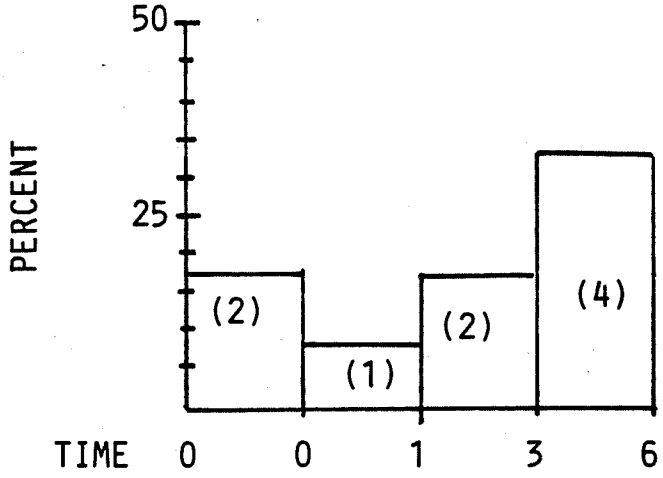
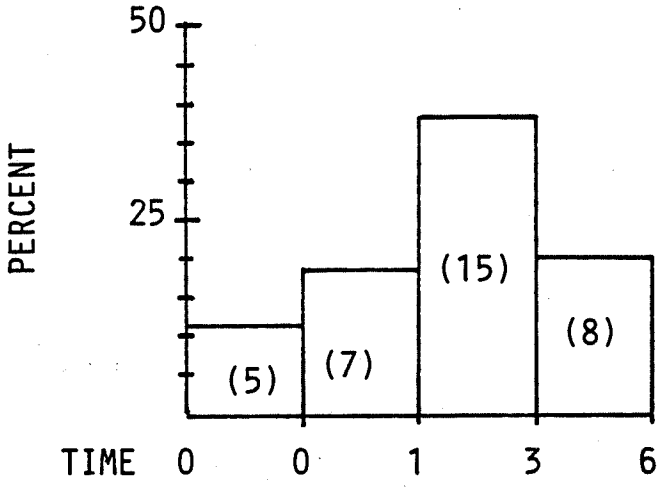


FIG. 12. FORECAST LEAD TIME PERCENTAGES FOR MAJOR FLOOD, ONE YEAR, DALLAS (A), AND ALL FORECAST POINTS ALONG THE TRINITY RIVER SYSTEM (B). ACTUAL NUMBER OF CAT 4 FLOOD EVENTS IN PARENTHESES. OBSERVED LEAD TIME PERCENTAGES NOT SHOWN.

Lead Time blocks are in hours.

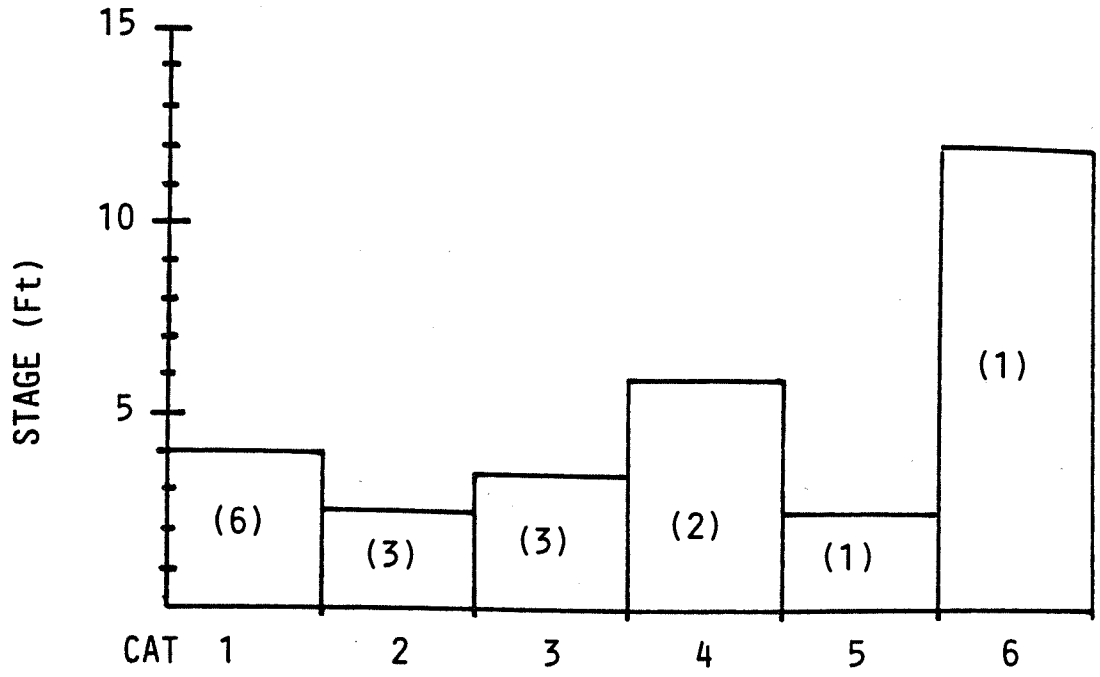


(A) JOHNSON CK NEAR GRAND PRAIRIE

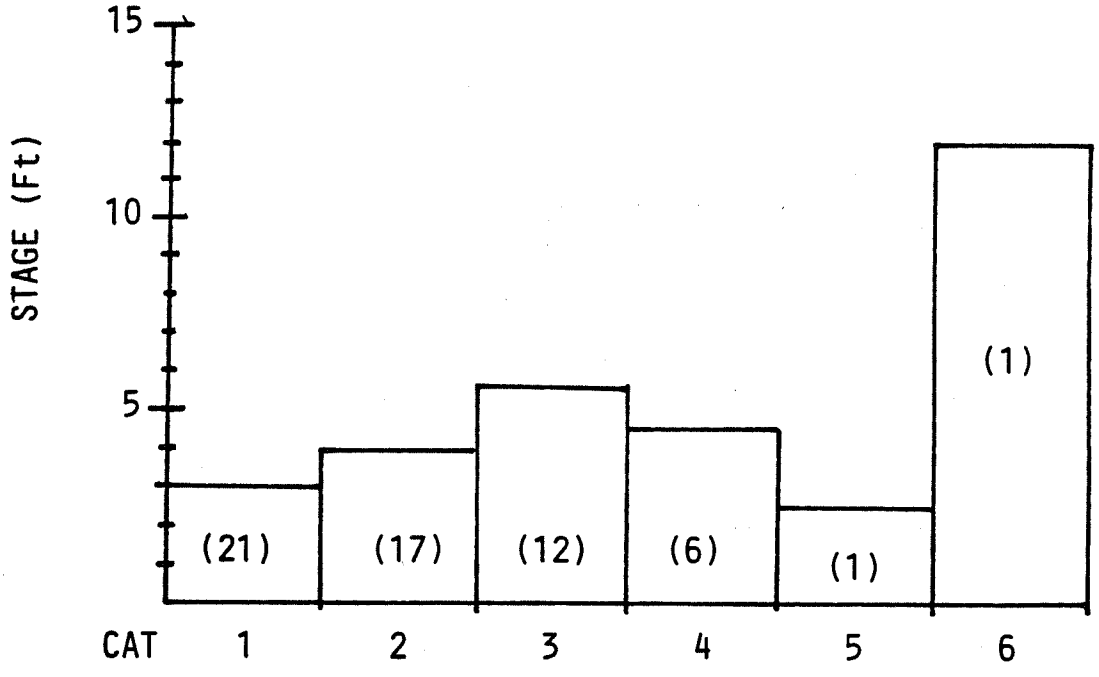


(B) THE TRINITY RIVER SYSTEM

FIG. 13. FORECAST LEAD TIME PERCENTAGES FOR FLASH FLOOD, ONE YEAR, GRAND PRAIRIE (A), AND ALL FLASH FLOOD FORECAST POINTS ALONG THE TRINITY RIVER SYSTEM (B). ACTUAL NUMBER OF CAT 2 FLASH FLOOD EVENTS IN PARENTHESES. OBSERVED LEAD TIME PERCENTAGES NOT SHOWN.



(A) THE BRAZOS RIVER AT SOUTH BEND



(B) THE BRAZOS RIVER SYSTEM

FIG. 14. AVERAGE EVENT ERROR FOR PREDICTING FLOODS IN CATEGORIES 1-6, ONE YEAR. ACTUAL NUMBER OF EVENTS NOT CORRECTLY FORECAST IN PARENTHESES.

forecast events fall in the zero lead time block. All numbers are hypothetical and are only intended to illustrate. No computation was actually made to determine average lead times for any Region. Aside from the first complete accounting of the flood prediction service in a given year, these kinds of tabulations over time should reveal important trends.

What is powerful information at one level of the Agency may be little more than "interesting to look at" at another level. Line hydrologists and research types like hard data and statistical numbers to draw conclusions from. Table 6 is a "first cut" idea at developing such output. NU is the number of observed events, not forecast events. PC, ME, BIAS, and FAR were defined earlier in this report. There is no connection between any numbers listed in Table 6 - all values were created to illustrate only. But it is clear that such tabular data provide keen insight to the flood forecasting business. Take the Cat 5 events, for example, WGRFC FTW. During the year there were 10 near record floods; of those forecast, 60% were correct (predicted the magnitude of flood); the mean stage error for these events not forecast was -3.6 feet, and the mean absolute stage error was 4.5 feet for those same "misses". BIAS for the Cat 5 events not predicted turns out to be 0.4, which indicates that FTW WGRFC had a substantial tendency to underforecast near record floods. Lastly, FAR reflects the fraction of forecast Cat 5 floods the public prepared for, but that did not verify. All in all, a pretty valuable data set. Plot variables like BIAS and FAR, and a history of sorts develops for the flood prediction service.

What about crest error - that thing we buried many pages ago. Does it have any redeemable value in verification? Perhaps, finally yes. If for data summaries like Table 5, where floods are categorized and analyzed, we also list a "MACE" (mean absolute crest error), we pigeon hole crest error numbers in a range of stage (flood category), and thus put those numbers in a fairly clear perspective. For floods we "hit", categorically speaking, we could then look at crest error. For floods we do not hit, the categorical error is more significant than crest error, at least, from an Agency verification viewpoint, and likely also from a public value viewpoint. However, should the profession decide it can never live without crest error stats, the price we pay is the requirement to collect as much crest information as possible in order to minimize estimation routine error. I personally do not feel it is worth the effort, but I bend, as always, to the will of the masses.

Very early in this report I raised the question of stage

Table 5. SUMMARY OF REGIONAL AND NATIONAL STATISTICS FOR MAJOR FLOOD (CAT 4) EVENTS, ONE YEAR. OBSERVED LEAD TIME TOP VALUE: FORECAST LEAD TIME BENEATH.

REGION	NUMBER OF EVENTS	PERCENT FORECAST	PERCENT MISSED	EVENT AVG. STAGE ERROR (FT)	PERCENTAGE DISTRIBUTION LEAD TIME IN HOURS							AVERAGE (HRS) LEAD TIME
					0	0<6	>6≤12	>12≤18	>18≤24	>24≤36	>36≤48	
EAST	285	73	27	3.1	5 5	15 12	21 18	20 23	10 16	15 20	14 6	11.6 17.3
SOUTH	390	61	39	4.9	14 14	24 18	20 23	18 25	13 12	9 6	2 2	9.1 11.5
CENTRAL	362	52	48	2.0	9 9	8 14	20 17	14 20	33 29	9 7	9 4	18.2 14.8
WEST	154	83	17	1.6	15 15	11 21	21 28	40 33	8 2	4 1	1 0	13.5 8.8
ALASKA	96	81	19	2.7	0 0	2 0	1 2	20 18	17 20	29 24	31 36	24.3 29.4
NATIONAL	1287	70	30	2.9	8.6 8.6	12.0 13.0	16.6 17.6	22.4 23.8	16.2 15.8	13.0 11.6	11.2 9.6	15.3 16.4

Table 6. SELECTED RFC VERIFICATION SCORES FOR ONE YEAR, ALL CATEGORIES OF FLOOD

FTW FLOOD	NU	PC	ME	MAE	BIAS	FAR
CAT 1	120	63	+1.8	2.6	0.6	0.31
CAT 2	163	44	+0.9	2.1	1.6	0.50
CAT 3	205	73	+3.1	4.0	0.7	0.22
CAT 4	102	59	-2.8	3.6	1.4	0.39
CAT 5	10	60	-3.6	4.5	0.4	0.35
CAT 6	3	50	-4.0	5.0	0.8	0.50
FLASH FLOOD						
CAT 1	38	66	-0.9	1.5	0.5	0.12
CAT 2	43	81	-1.1	2.8	0.8	0.19
CAT 3	18	39	+3.0	4.1	0.4	0.53
TUL FLOOD						
CAT 1	147	61	+0.8	2.0	0.4	0.35
CAT 2	151	53	+2.2	3.3	0.5	0.40
CAT 3	198	76	+1.7	2.9	1.3	0.20
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

NOTE: There were no Cat 4 flash floods observed or forecast during the year.

error significance in flood forecast verification. What is the significance of a one foot error, or two, or three....somewhere along the Trinity River versus somewhere along the Mississippi River? Well, by virtue of definition, a major flood is a major flood, wherever, a "hit is a hit", wherever, and proper credit is given in the verification plan. Nothing like this can come out of crest or stage verification. But a major flood in, say, Dallas could very well cause more grief than a major flood in, say, Vicksburg, and a two foot miss in category at Dallas might be more serious than at Vicksburg. So while it is true that two feet off at Vicksburg likely represents more percent error in forecast volume than at Dallas, we have no real basis to conclude that therefore a two foot error in forecast flood magnitude at Vicksburg is more significant in verification. No improvement here in our ability to interpret stage error. But, and it is a big but, we at least have acquired the means to put stage error in relative perspective to rise, and we as a Service would understandably be more concerned with indications of diminished forecast performance at higher categories of flood than for floods of lesser magnitude. There will never be a pat answer to "what is the significance of a two degree error in forecast temperature?". The best answer is "where does the error fall - near freezing level, or 100 degrees, or .....". I rest the case for hydrology.

7. CONCLUSIONS

Verification statistics and tabulation need to be looked at carefully; what is presented in this report is, at best, first draft. A statistic that takes into account the number of forecast events when computing stage error data would be helpful. Thought has been expressed by hydrologists that hydrology needs to develop "criteria" that also reflect "how well we could or should do." Verification should not wait for this scientific breakthrough. It is a worthy ideal; I cannot argue otherwise, but I do not have the foggiest idea how to begin, considering the variables, vagaries, and constraints in the flood forecasting business. How well we have done, and how well we are doing is, by itself, 100% more information than the Agency has ever had available for deliberation.

Verification of flood forecasts is a "hard nut to crack." Other hydrologic products, from water supply to inflow predictions, strike me as being relatively simple. However, a tremendous amount of software development and data would be required.

Is there possible problem with the suggested verification system, in terms of service evaluation? Yes. It is possible (although not likely, I think) that the procedures herein could make us look better at flood forecasting than we are. It largely depends upon how flood classification is performed at each forecast point. There is always the individual who sees opportunity to beat the system, and it is possible to construct the flood levels at some stations for "maximum gain" - like 90% of the rises just happen to always reach "major flood"; major flood starting just above bankfull and continuing upward to the snows of Mount Yuk. This is blatant dishonesty. If flood classification is accomplished as prescribed, there should be no performance evaluation problem. But as with most anything radically new, only time and test shows true merit. Compared with most of the verification in meteorology, this verification for hydrology is quite sophisticated. But it should be, by comparison,as the hydrologist has a simple verification problem - he does not have to chase floods in time and space.

Categorical verification dovetails nicely with the hopes for computer worded forecasts, the prospect of which delights many people. Once forecast point flood classification is completed (Fig 1 example), and computer resident, the information becomes germane to public statements, and as such has value beyond verification. There seems to be an on-going

concern in the Weather Service in issuing consistently effective flood warnings - there is always someone ready to advise field personnel on the "best wording". Perhaps computer delivered flood data and text would improve matters a bit. At least time would be saved.

Throughout this report the term "Agency" has been used. By Agency, I have intended to imply all levels of the organization - RFC, up through NWS Headquarters, up through NOAA, and up through Commerce. But verification for hydrology has been viewed by me, from the onset, as a more top down need than down up. It is higher management that has the most critical need for the verification data that sufficiently summarizes our ability to predict flood. I do not think the proposed verification, as detailed by me, is perfect. I do think someone with a keen mind might take the ideas presented herein and improve on them. But until then, I hope "The Plan", as stands, has sufficient merit to warrant careful review and consideration by both the Agency and the Hydrology Profession.

8. ACKNOWLEDGMENTS

Considerable thanks is due Mrs. Anne Smith, now retired Hydrologic Technician, for typing the entire manuscript. She is both personally and professionally the best of the best. Mr. Ernie Cathey, Staff Hydrologist, kindly did most of the drafting of the figures. I am also indebted to Walt Sittner, retired NWS Research Hydrologist, and Eric Anderson, Office of Hydrology/HRL, both of whom made numerous suggestions for improving the report. Lastly, appreciation is expessed to Dave Smith, Regional Hydrologist, SRH, and the WGRFC staff for serving as "sounding board" for many ideas that ended up being both typed and drafted.

9. REFERENCES

Campbell, A. K., 1985: 1982 and 1983 Watch/Warning Verification Flash Flood, Winter Storm and High Wind. NOAA Technical Memorandum NWS FCST-30, Office of Meteorology, Silver Spring, MD, p.6.

Linsley, R. K., Kohler, M. A., and Paulhus, L. H., 1973: Hydrology for Engineers. Second Edition, McGraw-Hill Book Co., P.361-368

Morris, D. G., 1983: January 14 Memo to Regional Hydrologist, SR, Categorical Flood Forecast Verification. 3 pp.

National Weather Service, 1980: Operations Manual. Part E, Chapter 13, p.3.

National Weather Service, 1982: National Verification Plan. U. S. Department of Commerce, NOAA, p.23.

Sittner, W. T., 1973: Determination of Flood Forecast Effectiveness By Use of Mean Forecast Lead Time. NWS, Office of Hydrology, unpublished reference, 19 pp.

Sittner, W. T., 1977: Determination of Flood Forecast Effectiveness By Use of Mean Forecast Lead Time. NOAA Technical Memorandum NWS HYDRO-36, Office of Hydrology, Silver Spring, MD, 22 pp.

Sittner, W. T. and Krouse, K. M., 1979: Improvement of Hydrologic Simulation By Utilizing Observed Discharge As An Indirect Input. NOAA Technical Memorandum NWS HYDRO-38, Office of Hydrology, Silver Spring, MD, 125 pp.

Sittner, W. T., 1987: Private communication.