VERIFICATION PROCEDURE REFERENCE GUIDE
July 16, 2015

Table of Contents: **Page**

1 Public and Fire Weather Forecasts and Warnings 5

1.1 Public Forecasts 5
1.1.1 Verification Sites 5
1.1.2 Web Interface 5
1.1.3 Data 5
1.1.4 Projections 6
1.1.5 Verification Reports 6

1.2 Winter Weather Warnings 10
1.2.1 Matching Warnings and Events 10
1.2.2 Quality Assurance 11
1.2.3 Extensions 11
1.2.4 Lead Time 12
1.2.5 Timing Error 12
1.2.6 Watches and Advisories 12
1.2.7 Backup Mode for Warnings 12

1.3 High Wind Warnings 12
1.3.1 Matching Warnings and Events 13
1.3.2 Quality Assurance 13
1.3.3 Extensions 13
1.3.4 Lead Time 13
1.3.5 Timing Error 14
1.3.6 Watches and Advisories 14
1.3.7 Backup Mode for Warnings 14

1.4 National Fire Danger Rating System (NFDRS) Forecasts 14
1.4.1 Verification Sites 14
1.4.2 Web Interface and Verification Reports 14
1.4.3 Elements 15

1.5 Red Flag Warnings (RFW) 15
1.5.1 Defining Events and Warnings 15
1.5.2 Matching Warnings and Events and Performing Verification 16
1.5.3 Extensions 16
1.5.4 Lead Time 16
1.5.5 Regional Reports 17

2 Convective Severe Weather 17
2.1 Storm-based Warnings 17

2.1.1 Quality Assurance 18
2.1.2 Matching Warnings and Events 18
2.1.3 *POD*, *FAR*, and CSI Calculations 20
2.1.4 Lead Time 20
2.1.5 Backup Mode for Warnings 21

2.2 County-based Warnings 21
2.2.1 Quality Assurance 22
2.2.2 Matching Warnings and Events 22
2.2.3 Lead Time 23
2.2.4 Backup Mode for Warnings 23
2.3 Watches 23

3 Marine Forecasts and Warnings 23

3.1 Wind and Wave Height Forecasts 23
3.1.1 Web Interface 24
3.1.2 Verification Sites 24
3.1.3 Verification Reports 24

3.2 Coastal Flood and Lakeshore Flood Warnings (CFW) 27
3.2.1 Matching Warnings and Events 27
3.2.2 Quality Assurance 28
3.2.3 Extensions 28
3.2.4 Lead Time 28
3.2.5 Timing Error 28
3.2.6 Watches 28
3.2.7 Backup Mode for Warnings 28

3.3 Storm-Based Special Marine Warnings (SMW) 28
3.3.1 Quality Assurance 29
3.3.2 Matching Warnings and Events 29
3.3.3 *POD*, *FAR*, and *CSI* Calculations 30
3.3.4 Lead Time 31
3.3.5 Backup Mode for Warnings 32

3.4 Zone-Based SMWs 32
3.4.1 Quality Assurance 32
3.4.2 Matching Warnings and Events 33
3.4.3 Lead Time 33
3.4.4 Backup Mode for Warnings 33

4 Hydrologic Warnings and Forecasts 33

4.1 Storm-Based FFWs 34
4.1.1 Quality Assurance 34

4.1.2 Matching Warnings and Events 34
4.1.3 *POD*, *FAR*, and *CSI* Calculations 34
4.1.4 Lead Time 35
4.1.5 Backup Mode for Warnings 36

4.2 County-Based FFWs 36
4.2.1 Quality Assurance 36
4.2.2 Matching Warnings and Events 37
4.2.3 Lead Time 37
4.2.4 Backup Mode for Warnings 37

4.3 Point-Based Flood Warnings (FLW) 37
4.3.1 Quality Assurance 38
4.3.2 Matching Warnings and Events 38
4.3.3 Lead Time 38
4.3.4 Absolute Timing Error 38
4.3.5 Backup Mode for Warnings 39

4.4 RFC River Stage Forecasts 39

5 Quantitative Precipitation Forecasts (QPF) 39
5.1 Weather Prediction Center (WPC) QPFs 39

6 Aviation Forecasts 40

6.1 Terminal Aerodrome Forecasts (TAF) 40
6.1.1 Verification Sites 40
6.1.2 Data 40
6.1.3 Web Interface 40
6.1.4 Projections 42
6.1.5 Verification Reports 42
6.1.6 Elements 42
6.1.7 Forecast Types 46

6.2 Aviation Weather Center (AWC) Forecasts 56
6.2.1 Background 56
6.2.2 Domestic Products Verified and Statistics Calculated 56

7 Tropical Cyclones 57

7.1 Tropical Cyclone Forecasts/Advisories 57
7.1.1 Verification Elements 57
7.1.2 Verification Process 57

7.2 Models 58
7.2.1 Verification Elements 58

7.2.2 Verification Process 58

7.3 Verification Reports 58

8 Climate Outlooks. 58
8.1 Medium Range and Seasonal Outlooks 58
8.2 U.S. Hazards Assessment Product 59

9 Models 60

10 References 61

Appendix

A Verification Scores A-1
B Glossary of Contractions and Terms B-1

1 Public and Fire Weather Forecasts and Warnings

The National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) public forecasts issued by all Weather Forecast Offices (WFO) are verified at select points.

1.1 Public Forecasts (PFM Verification)

After May 31, 2012, this program is no longer supported. Historic data up to this date are available, but new data are no longer imported into the system. A new public forecast verification program is being developed, and its implementation is expected in 2016.

1.1.1 Verification Sites

All sites forecast in the point forecast matrices (PFM) that issue routine Meteorological Aviation Reports (METAR) and Special Aviation Weather Reports (SPECI) are verified unless the local WFO determined that a particular site is unrepresentative of its surroundings or inappropriate for verification. An interactive station directory of all active verification sites is maintained on the Public Verification Home Page of the NWS Performance Management Website.

1.1.2 Web Interface

NWS employees access verification statistics through the Public Verification Home Page of the NWS Performance Management Website. *Stats on Demand* accesses an interactive database of monthly data and generates verification statistics customized to the user's request. The user finds PFM verification by selecting the "Verification" and "Public" menus and scrolling down to "PFM Verification." Data may be requested for Max/Min Temperature, Probability of Precipitation (PoP), or Sky Cover for one or more

a. months,

b. scheduled forecast issuance times, i.e., early morning, late afternoon,

c. forecast projections, and

d. verification sites, i.e., single site, multiple sites.

If desired, matching forecasts for all of the above parameters from a single Model Output Statistics (MOS) guidance product may also be selected. MOS guidance beyond the 60-hour projection is not available for the sky cover element.

1.1.3 Data

Public forecast data come from the scheduled PFMs issued by each WFO twice a day at 0400 and 1600 Local Time (LT). The latest 0400 (1600) LT PFM issued between 0000 and 0559 (1200 and 1759) LT, including corrections, are accepted. Amendments are not verified. Guidance forecasts come from available model and MOS products. The verifying observations primarily come from the METAR and SPECI reports issued for each location in the PFMs.

Daytime maximum and nighttime minimum temperatures are inferred from the instantaneous temperatures in each observation and the 6-hour maximum/minimum temperature groups. The satellite cloud product is also used as an observation source for sky cover verification. All METARs and SPECIs are tested for reliability and consistency, and suspicious data are removed. A description of these quality control algorithms is found through a link on the Public Verification Home Page of the NWS Performance Management Website.

1.1.4 Projections

Projections for public elements are defined in terms of the number of 12-hour forecast periods that have elapsed since the forecast issuance time, which is assumed to be 1200 UTC for the early morning PFM issuance and 0000 UTC for the late afternoon issuance. Unless otherwise stated for the individual element, these 12-hour forecast periods are defined as 1200 to 0000 UTC and 0000 to 1200 UTC. Forecasts are made out to Day 7, totaling thirteen forecast periods for the early morning PFM issuance and fourteen for the late afternoon issuance.

1.1.5 Verification Reports

Each report contains verification statistics tailored to the parameters specified through the web interface. Contingency tables, accuracy measures, skill scores, mean errors, and histograms of error categories are included.

Forecast data in the PFMs issued at 0400 LT, when matched to MOS guidance, are matched to the 0000 UTC model cycle from the same date. Forecast data in the PFMs issued at 1600 LT, when matched to MOS guidance, are matched to the 1200 UTC cycle from the same date for the first five 12-hour forecast periods. Beyond period 5, the forecast data in the PFMs issued at 1600 LT are matched to the 0000 UTC cycle from the same date.

1.1.5.1 Max/Min Temperatures

The forecast period for daytime maximum temperatures is defined as 7 am to 7 pm Local Standard Time (LST), except in Alaska Region, where the forecast period is defined 5 am to 8 pm Alaska Standard Time (AKST). The forecast period for nighttime minimum temperatures is defined as 7 pm to 8 am LST, except in Alaska Region, where the forecast period is defined 5 pm to 11 am AKST. Projections (one through 14) are expressed as these forecast periods. All forecast and observed temperatures are expressed in whole degrees Fahrenheit. The following statistics are available in the *Stats on Demand* reports:

a. Number of cases in the sample.

b. Mean absolute error *(MAE)* for the entire sample (defined in Appendix A, section 4.1 b). The percent improvement of the PFM over a single guidance product is also provided for this statistic whenever that guidance product was selected.

c. Mean (algebraic) error *(ME)*, see Appendix A, section 4.1 a.

d. Root mean square error *(RMSE)*, see Appendix A, section 4.1 c.

e. Number of cases whenever the absolute error of the PFM or the selected guidance product was greater than or equal to 6º F.

f. *MAE* whenever the absolute error of the PFM or the selected guidance product was greater than or equal to 6º F.

g. Number of cases when the PFM, the selected guidance product, or the observation changed (in either direction) from the previous 24 hours by at least 10º F. Data are not provided for the first two 12-hour forecast projections.

h. *MAE* when the PFM, the selected guidance product, or the observation changed (in either direction) from the previous 24 hours by at least 10º F. The *MAE* percent improvement of the PFM over the selected guidance product during these circumstances is also provided. Data are not provided for the first two 12-hour projections.

i. Number of cases when the PFM was changed from the selected guidance product by 4º F or greater.

j. *MAE* when the PFM was changed from the selected guidance product by 4º F or greater. The *MAE* percent improvement of the PFM over the selected guidance product during these circumstances is also provided.

k. For minimum temperatures only, when the previous day's minimum temperature was 40º F or greater, the following statistics are provided for forecast temperatures equal to or less than 32º F: probability of detection *(POD*), false alarm ratio *(FAR)*, and critical success index *(CSI)*. See Appendix A, section 3, for the definitions of *POD*, *FAR*, and *CSI*.

l. Histogram of the absolute errors for PFM and the selected guidance product using the following absolute error categories in degrees Fahrenheit: 0-3, 4-5, 6-10, 11-15, greater than 5, greater than 10, and greater than 15. The value of each error category is provided as a percentage of the total sample. The percent improvement of the PFM over guidance is provided for the following absolute error categories: greater than 5º F and greater than 10º F.

1.1.5.2 Probability of Precipitation

Probability of 0.01 inch or greater liquid equivalent precipitation (PoP) is verified within the 12-hour forecast periods defined in section 1.1.4, except in Pacific Region, where the 12-hour PoP periods are defined six hours later than everywhere else, i.e., 0600 to 1800 UTC and 1800 TO 0600 UTC. The following forecast values are allowed in the PFM and are used in verification: {0, 5, 10, 20, 30, ... , 80, 90, 100}. MOS PoPs, forecast to the nearest percent, are rounded to the nearest allowable PFM value. From METARs, 12-hour precipitation amounts to the nearest hundredth of an inch are inferred for the aforementioned periods. All precipitation gage reports

are automatically quality controlled using the following: (a) internal consistency checks with other parts of the METAR report, (b) Stage III quantitative precipitation estimates issued by the River Forecast Centers, and (c) data from the national snow analysis issued by the National Operational Hydrologic Remote Sensing Center. The following statistics are available in the *Stats on Demand* reports:

a. Number of forecast periods.

b. Number of observed precipitation cases.

c. Observed precipitation frequency, i.e., bullet b. divided by bullet a.

d. Mean PoP Forecast: the mean PoP value for all chosen forecasts.

e. Mean PoP Forecast with Precipitation: the mean PoP value for all chosen forecasts whenever 0.01 inch or greater (measurable) precipitation occurred.

f. Mean PoP Forecast without Precipitation: the mean PoP value for all chosen forecasts whenever no measurable precipitation occurred.

g. Brier score (defined in Appendix A, section 4.2 a.). The percent improvement of the PFM over a single guidance product is also provided whenever that guidance product was selected.

h. Brier Score whenever PFM PoP was 30% or greater. The Brier score percent improvement of the PFM over a single guidance product during these circumstances is also provided. The number of these cases is provided in parentheses next to the percent improvement score.

i. Brier score whenever measurable precipitation occurred. The Brier score percent improvement of the PFM over a single guidance product during these circumstances is also provided. The number of these cases is provided in parentheses next to the percent improvement score.

j. Brier score whenever the PFM differed from the selected guidance product by at least 20%. The Brier score percent improvement of the PFM over a single guidance product during these circumstances is also provided. The number of these cases is provided in parentheses next to the percent improvement score.

k. PoPs are interpreted as binary (yes/no) forecasts for measurable precipitation. PoPs greater than or equal to 50% are interpreted as “yes.” PoPs less than 50% are interpreted as “no.”

l. The relative frequency of measurable precipitation is provided for the times when the following PoP thresholds were forecast: 0, 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 percent. Probabilistic forecasts are perfectly reliable when each of the

PoP thresholds equals the relative frequency of measurable precipitation events that occurred when it was forecast. For example, if measurable precipitation occurs 30% of the time that you forecast a 30% PoP, your 30% PoP forecasts were reliable.

1.1.5.3 Sky Cover

a. WFO Forecasts and MOS Guidance: Sky cover is forecasted categorically by the PFM and MOS alphanumeric products as clear, few, scattered, broken, or overcast. This forecast is made instantaneously every 3 hours for the first 60 hours of the PFM. During this time, the first five forecast periods are defined as >0 to 12 hours, >12 to 24 hours, >24 to 36 hours, >36 to 48 hours, and >48 to 60 hours, defining the "zero-hour projection" as 0000 or 1200 UTC, whichever is closer to the scheduled issuance time of the PFM. For projections beyond 60 hours, the PFM issues 6-hour mean forecasts of sky cover so each fits into the 12-hour projection bins defined in section 1.1.3. Due to the lack of availability of sky cover guidance forecasts beyond 60 hours in the format used in the PFM, guidance forecasts at these projections are not available for comparison to the PFM. However, beyond the 60-hours projection, gridded forecasts from the National Digital Forecast Database (NDFD) are compared to gridded guidance products (i.e., model data and gridded MOS data) at the grid point nearest to the verification site, sometimes called the "nearest neighbor grid point."

b. Observations: The verifying observation uses an algorithm that combines the METAR (i.e., 12,000 feet and below) with the alpha-numeric satellite cloud product (SCP) (i.e., above 12,000 feet). In addition to reporting a categorical sky cover value each hour, the SCP also provides an hourly effective cloud amount (ECA), which is a numerical estimate of cloud opacity above 12,000 feet. The algorithm, which is fully described on the Public Verification Home Page of the NWS Performance Management Website, provides an estimate of the total opaque sky cover (clear, few, scattered, broken, or overcast), and this is what is used to verify the forecast categories from the PFM and MOS.

1.1.5.4 Wind Direction, Sustained Wind Speed, and Wind Gusts

a. WFO Forecasts and MOS Guidance: Wind direction (expressed as one of the eight points of the compass) and sustained wind speed are forecasted every 3 hours out to the 60-hour projection in the PFM and the alphanumeric MOS messages; an instantaneous forecast is made every 3 hours. Wind gust forecasts appear on a separate line whenever the forecast gust exceeds the sustained speed by a given threshold. The first five forecast periods are defined as >0 to 12 hours, >12 to 24 hours, >24 to 36 hours, >36 to 48 hours, and >48 to 60 hours, defining the "zero-hour projection" as 0000 or 1200 UTC, whichever is closer to the scheduled issuance time of the selected PFM. Sustained wind speed and wind gusts in the PFM are expressed in statute miles per hour (mph), whereas the sustained speed units in the alphanumeric MOS bulletins are nautical miles per hour (knots). Verification statistics of sustained wind speed and wind gusts are

always expressed as mph. For projections beyond 60 hours, 6-hourly wind direction and sustained wind speed forecasts are taken from the gridded NDFD, and guidance forecasts are taken from the gridded model or MOS products.

b. Verifying observations are taken from the METARs.

1.2 Winter Weather Warnings

a. NWS employees access these verification statistics through the Public Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each data request, the user provides the following definitions and boundaries:

(1) Type of warning (generic or one of the event-specific varieties listed in Table A-1).

(2) Beginning and ending dates.

(3) One or more zones, WFOs, states, or NWS regions.

b. All winter weather warning verification is performed using one of the following methodologies. Advisories are not verified. The user of *Stats on Demand* specifies which method to use when requesting data. The default selection is "All Winter Events (Generic)."

(1) <u>All Winter Events (Generic)</u>. Any type of winter weather event that meets warning criteria verifies any type of winter weather warning, and any winter weather warning covers any winter event that meets warning criteria. See Table A-1. This is the most frequently used method, and the method used in all Government Performance Results Act (GPRA) reports. It is also the default setting on the *Stats on Demand* winter weather warning request interface.

(2) <u>Event Specific</u>. Each warning is verified with the exact, event specific *Storm Data* entry, e.g., an ice storm warning is verified with an ice storm entry in *Storm Data* and vice versa. See Table A-1.

1.2.1 Matching Warnings and Events

All warning data are automatically taken from the warning products issued to the public. Each public forecast zone is treated as a separate verification area. Therefore, a warning covering three zones is counted as three warned areas or three warnings. All events that meet regional/local warning criteria (see Table 1, two middle columns) are automatically taken from the certified *Storm Data* reports.

Table 1. *Storm Data* entries (events), warning types, and verification methods.

Warning Type	**Event Specific Verification** *Storm Data* entries required for event specific verification of the warnings in left column and vice versa	**Generic Verification** *Storm Data* entries required for generic verification of any warning in left column and vice versa	**Warning Criteria Not Met** *Storm Data* entries that do <u>not</u> verify the warnings in left column
Winter Storm	Not applicable	Winter Storm, Heavy Snow, Sleet, Ice Storm, Lake Effect Snow, or Blizzard	Winter weather
Ice Storm	Ice Storm only		
Lake Effect Snow	Lake Effect Snow only		
Blizzard	Blizzard only		

The following event times, defined in NWSI 10-1605, Storm Data Preparation, are provided for each event listed in *Storm Data* and are used for verification:

a. Beginning time.

b. Criteria time.

c. Ending time.

Warnings and events that meet warning criteria are recorded in separate verification databases. Whenever the time period between the criteria time and the ending time of an event coincides with any part of the valid period of a warning, one warned event and one verified warning are counted. Unwarned events and unverified warnings are also counted. From these statistics, the *POD*, *FAR*, and *CSI* are computed (see sections 3.1 to 3.3 of Appendix A) and listed in the verification reports. Numerous examples of specific verification scenarios are provided through the Winter Storm Warning Verification training module link on the Public Verification Home Page of the NWS Performance Management Website.

1.2.2 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the Valid Time and Event Code (VTEC) line are checked for consistency with the World Meteorological Organization (WMO) warning header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

1.2.3 Extensions

Warnings may be extended in area and/or time. Extensions of warnings to new areas (zones) are counted as new warnings, i.e., one warning per zone. Each time extension of a zone already

warned is counted as a new warning only if the earlier warning did not verify during its valid period. Examples of the verification of warning extensions are provided through the Winter Storm Warning Verification training module link on the Public Verification Home Page of the NWS Performance Management Website.

1.2.4 Lead Time

A lead time (in hours) is computed for each zone that experiences an event meeting warning criteria. If the event criteria time does not occur during the valid period of a warning, the lead time for that event is zero. If the event criteria time occurs during the valid period of a warning, the lead time for that event is computed by subtracting the warning issuance time from the event criteria time. The warning issuance time comes from the WMO header of the warning. Negative lead times are set to zero. The average lead time is computed from all lead times listed in the event database, including zeroes.

1.2.5 Timing Error

The timing error (in hours) for each warned event is defined as the event beginning time minus the warning beginning time. For each data request, the mean absolute error and mean algebraic error (bias) are provided.

1.2.6 Watches and Advisories

While watches and advisories are not verified in the same manner as warnings, the following statistics are provided:

a. The percentage of unwarned events that occurred with an advisory in effect.

b. The percentage of unwarned events that occurred with a watch in effect.

1.2.7 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO (World Meteorological Organization) header of the warning.

1.3 High Wind Warnings

NWS employees access these verification statistics through the Public Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each data request, the user provides the following boundaries:

a. Beginning and ending dates.

b. One or more zones, WFOs, states, or NWS regions.

1.3.1 Matching Warnings and Events

All warning data are automatically taken from the warning products issued to the public. Each public forecast zone is treated as a separate verification area. Therefore, a warning covering three zones is counted as three warned areas or three warnings.

All events that meet warning criteria are automatically taken from certified *Storm Data* reports. The following event times, defined in NWSI 10-1605, Storm Data Preparation, are provided for each event listed in *Storm Data* and are used for verification:

a. Beginning time.

b. Ending time.

Warnings and events that meet warning criteria are recorded in separate verification databases. Whenever an event that meets warning criteria (defined temporally as the period between its beginning and ending times) coincides with any part of the valid period of a warning, one warned event and one verified warning are counted. Unwarned events and unverified warnings are also counted. From these tallied statistics, the *POD*, *FAR*, and *CSI* are computed (see sections 3.1 to 3.3 of Appendix A) and listed in the verification reports.

1.3.2 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

1.3.3 Extensions

Warnings may be extended in area and/or time. Extensions of warnings to new areas (zones) are counted as new warnings, i.e., one warning per zone. Each time extension of a zone already warned is counted as a new warning only if the earlier warning did not verify during its valid period.

1.3.4 Lead Time

A lead time (in hours) is computed for each zone that experiences an event meeting warning criteria. If the event beginning time does not occur during the valid period of a warning, the lead time for that event is zero. If the event beginning time occurs during the valid period of a warning, the lead time for that event is computed by subtracting the warning issuance time from the event beginning time. The warning issuance time comes from the WMO header of the warning. Negative lead times are set to zero. The average lead time is computed from all lead times listed in the event database, including zeroes.

1.3.5 Timing Error

The timing error (in hours) for each warned event is defined as the event beginning time minus the warning beginning time. For each data request, the mean absolute error, the mean algebraic error (bias), and a distribution of errors are provided.

1.3.6 Watches and Advisories

While watches and advisories are not verified in the same manner as warnings, the following statistics are provided:

a. The percentage of unwarned events that occurred with an advisory in effect.

b. The percentage of unwarned events that occurred with a watch in effect.

1.3.7 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

1.4 National Fire Danger Rating System (NFDRS) Forecasts

After May 31, 2012, this program is no longer supported. Historic data up to this date are available, but new data are no longer imported into the system.

Forecasts and observations in this automated program come from the fire weather product with the AWIPS product identifier (PIL) NMCFWOrr, NMCFWOss or NMCFWOxxx, where rr refers to one of the four CONUS NWS regions, ss refers to a state, and xxx refers to a specific WFO.

Both NFDRS forecasts and observations are valid at 1300 LST but are issued as separate bulletins with the same product name. The forecasts are issued approximately 22 hours prior to the forecast valid time, and the verifying observations are disseminated shortly after 1300 LST the next day. For example, a forecast valid at 1300 LT will be issued at approximately 1500 LT the previous day. The forecasts verifying observations are subsequently matched and verified.

1.4.1 Verification Sites

A database of the NFDRS observation sites used in verification is posted to the Fire Weather Verification Home Page of the NWS Performance Management Website.

1.4.2 Web Interface and Verification Reports

NWS employees access verification statistics from the Fire Weather Verification Home Page of the NWS Performance Management Website. Data are only available for the CONUS. *Stats on Demand* accesses an interactive database and generates verification statistics customized to the user's request. With each data request, the user provides the following definitions and boundaries:

a. Element. See section 1.3.3 of this appendix.

b. Beginning and ending dates. Specific months within a longer specified valid period may also be selected, e.g., select all June, July, and August data from the valid period January 1, 2004, thru December 31, 2008.

c. Spatial domain, to include (1) one or more individual verification sites, (2) one or more fire weather forecast zones, (3) one or more WFO forecast areas, or (4) the entire Nation, excluding Alaska.

d. Threshold error value. For temperature, relative humidity and wind speed, the user includes a threshold absolute error value. This value is entered by the user after selecting the desired forecast element (i.e., temperature, relative humidity, or wind speed) and is used to calculate the percentage of time the absolute error was greater than or equal to the user-specified value. Examples: 5ºF (temperature), 10% (relative humidity), 10 mph (wind speed).

e. Threshold window for *POD, FAR,* and *CSI*. For temperature, relative humidity, and wind speed, the user specifies the window of values, i.e., lowest and highest, from which the *POD, FAR,* and *CSI* will be calculated. These values are entered by the user after selecting the desired forecast element (i.e., temperature, relative humidity, or wind speed). Examples: between 90 ºF and 120 ºF (temperature), between 30 and 100 mph (wind speed).

1.4.3 Elements

a. State of Weather. Each state of weather is designated by a weather code (single integer) value from zero to 9. Each weather code is assigned to one of following three groups: group i (weather codes zero and 1), group ii (weather codes 2 and 3), and group iii (weather codes 4 thru 9). A forecast is counted as a hit if it falls within the same group as the observation.

b. Temperature.

c. Relative Humidity.

d. Wind Speed.

1.5 Red Flag Warnings (RFW). Perform RFW verification manually at each WFO with an RFW program.

1.5.1 Defining Events and Warnings. For verification purposes, an event is defined (a) when observations are queried in a given zone to determine if weather conditions meet or exceed the locally established warning criteria, <u>and</u> (b) when local land management personnel determine prior to warning issuance that the fuels meet or exceed the critical burning threshold. Each WFO and its local users determine the specific, unique weather criteria for issuance of a RFW in its

area of responsibility. When observations are not available in a zone, the determination of an event should be based on the objective opinion of an experienced forecaster. Events are not determined by the number of fire starts or by querying users to determine if they feel you "hit" or "missed" warnings.

In summary, warnings are issued based on two factors: weather and fuel conditions. The former is determined by the forecaster, and the latter is determined by the user. The latter is determined in advance of the warning issuance and doesn't change when the verification is done.

1.5.2 Matching Warnings and Events and Performing Verification. Treat each fire weather zone as a separate verification area. Therefore, count a warning covering three zones as three warned areas or three warnings. Record warnings and events in separate databases. All listings in the event database must meet weather warning criteria. Warnings are verified based on whether the zone experienced locally-established weather warning criteria.

Count one verified warning and one warned event whenever an event meeting weather warning criteria occurs in a warned zone. Count one missed event if an event meeting weather warning criteria occurs in a zone with no warning. However, if weather warning criteria were met, but a warning was not issued because the users determined that the fuels were insufficient to warrant a warning, then a missed event is not recorded.

Count one unverified warning (or false alarm) for each warned zone that does not meet weather warning criteria.

The majority of RFWs include wind and humidity criteria or some index based on these parameters. However, in some areas, some warnings are issued due to the occurrence of dry lightning. These two types of events can exhibit big differences in lead time, Probability of Detection (POD), False Alarm Ratio (FAR), and Critical Success Index (CSI). Therefore, all offices doing RFW verification will verify RFWs three ways: first, by tracking and verifying all events; second, by tracking and verifying just wind/humidity events or their equivalent; and third, by tracking and verifying just dry lightning events. All three sets of verification must be sent to the regional fire weather program manager and/or verification program manager at the end of the calendar year. If an office does not have any criteria for dry lightning events, that office will report "n/a" for their dry lightning verification.

1.5.3 Extensions. Warnings may be extended in area and/or time. Count extensions of warnings to new areas (zones) as new warnings, i.e., one warning per zone.

1.5.4 Lead Time. Compute a lead time for each zone that experiences an event. Subtract the time of warning issuance from the time when the event first met warning criteria in the zone. If warning criteria at a particular WFO are subject to a temporal limit (e.g., the criteria must be met for a minimum of three consecutive hours), then the lead time is computed from the first observed occurrence of that temporal criteria. For example, a warning was issued at 0600 LST and the weather criteria were first met at 1200 LST. However, based upon the established temporal limit, the third hour of the weather criteria was not observed until 1500 LST. Assuming warning criteria as stated in the local Annual Operating Plan (AOP) have been met,

then the calculated lead time would be 6 hours, i.e., 1200 LST (first occurrence) minus 0600 LST (warning issuance time).

Set negative values to zero. If a zone experiences an event meeting warning criteria when a warning is not in effect, assign that event a lead time of zero. Compute average lead time from all the lead times listed in the event database, including zeroes.

1.5.5 Regional Reports. The NWS regional headquarters report the annual verification statistics to the National Fire Weather Operations Coordinator (NFWOC). The report should contain the following elements by office: Number of RFWs issued, average lead time in hours, number of correct warnings, number of warnings that did not verify and number of unwarned events. These elements need to be reported for all events, just wind/humidity events or their equivalent, and just dry lighting events. Also include the number of spot forecasts issued by each office. The NFWOC will send the regional fire weather program managers a spreadsheet to fill in these numbers the first week of January. The NWS regions will report these numbers to the NFWOC by the last day in January. The NFWOC will compute the POD, FAR and CSI for each office, each region and nationally, as well as the average lead time for each region and nationally. POD, FAR and CSI are computed as follows:

a. Number of correct warnings (A)

b. Number of unwarned events (B)

c. Number of warnings that did not verify (C)

d. POD = A/(A+B)

e. FAR = C/(A+C)

f. CSI = A/(A+B+C)

2 Convective Severe Weather

This section describes the verification of all severe thunderstorm and tornado watches and warnings.

2.1 Storm-based Warnings

Storm-based warning issuance replaced county-based warning issuance October 1, 2007, so storm-based warning verification should be used for warnings issued on or after this date. For warnings issued before October 1, 2007, see section 2.2 for a description of county-based warning verification. NWS employees access verification statistics through the Severe Weather Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each request, the user provides the following definitions and boundaries:

a. Type of warning (method).

b. Beginning and ending dates.

c. One or more WFOs or NWS regions.

d. Severity of event, based on total cost of damage, number of fatalities, and/or tornado EF-scale (optional).

2.1.1 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

2.1.2 Matching Warnings and Events

All warning data are automatically extracted from the warning products issued to the public. The basic area for a tornado or severe thunderstorm warning is the polygon boundary outlined by the latitude-longitude coordinates located at the bottom of the product. Therefore, for verification purposes, the area within the latitude-longitude boundaries is counted as the warning.

Verification statistics are computed for tornado and severe thunderstorm warnings and events using one of three methods. The user of *Stats on Demand* selects the method. The first method combines severe thunderstorms and tornadoes together and treats them as a single event type. The latter two methods are event specific—they treat non-tornadic severe thunderstorms and tornadoes as separate types of events. See Table A-2 for illustration.

Table A-2. *Storm Data* entries (events) used to verify local severe storm warnings.

Warning Type	**Event Specific Verification** Each warning type in the left column is only verified by the corresponding event type from the same line of this column. Each event type in this column is be covered by the corresponding warning type from the same line in the left column.	**All Severe Thunderstorm and Tornado (Generic) Verification** Each warning type in the left column is verified by *any* of the event types in this column. Any event type in this column is be covered by one of the warning types in the left column.
Severe thunderstorm (SVR product)	Non-tornadic severe thunderstorm, e.g., hail or thunderstorm wind meeting NWS warning criteria	Non-tornadic severe thunderstorm or tornado
Tornado (TOR product)	Tornado	

All event data are automatically taken from certified *Storm Data* reports. A Tornado (TOR) or Severe Thunderstorm Warning (SVR) is verified by a confirmed event of the type specified in

Table A-2 and occurring within the temporal and areal boundaries of the warning. Unlike the county-based severe weather verification method, multiple severe thunderstorm wind and hail events in the same county, separated by less than 10 miles and 15 minutes, are ***not*** considered duplicates.

Each warning is checked to see if a verifying event occurred within its temporal and areal boundaries and is categorized as verified or unverified.

Events are logged in *Storm Data* using one of two methods. The first method is an isolated event at a single location (referred to as an instantaneous event). An example would be an isolated hail event reported at a single time. The second method is used for an event that starts at one location and moves to another location over a period of time (referred to as a track event). An example would be a tornado that moved from one location to another. Both methods are evaluated differently.

a. Evaluation of Instantaneous Events. A check is performed on each instantaneous event to see if a warning was in effect at the time and location of the event. If so, the event was warned. If not, the event was unwarned.

b. Evaluation of Track Events. Before track events are evaluated, two assumptions are made:

(1) The event travels in a straight path between the event beginning and ending locations logged in *Storm Data.*

(2) The event travels at a constant speed between the event beginning and ending locations logged in *Storm Data.*

Once these assumptions are made, the location of the event is estimated every minute for the duration of the event. The event is then evaluated at each of those locations and times. For example, a tornado event lasting from 0100 to 0105 and traveling three miles would be evaluated at six locations and times, i.e., six segments of the event. A check is then performed at each point along the track of the event to see if a warning was in effect. If so, the event was warned at that point. If not, the event was unwarned at that point.

Next, the ***percentage of the event warned*** *(PEW)* is calculated for each event. For an instantaneous event, the *PEW* is zero for an unwarned event and 100 for a warned event. For a track event, the *PEW* is calculated linearly, dividing the total number of warned one-minute segments by the total time length of the event. In the case of the example in the previous paragraph (a tornado lasting from 0100 to 0105), if the tornado was inside the warning polygon at 0100 and 0101 (the warned segments of the event) and moved outside the polygon starting at 0102 (the unwarned segments of the event lasted from 0102 to 0105), the *PEW* for the entire event would be two divided by six or 33.3 percent.

2.1.3 *POD, FAR,* and *CSI* Calculations

Once a *PEW* is calculated for each event, the *POD* may be calculated:

$$POD_{SB} = \frac{0.01}{N}\sum_{i=1}^{N} PEW_i$$

where:

POD is the probability of detection,

SB is for storm-based warnings,

PEW is the total percentage of each event, *i*, warned, expressed as a value from zero to 100, and

N is the total number of events.

The best possible *POD* is one; the worst is zero. Additional information, with examples for severe thunderstorms and tornadoes, is found through the verification training module link on the Severe Weather Verification Home Page of the NWS Performance Management Website.

The *FAR* for storm-based warnings is the same calculation that was used for the old county-based system:

$$FAR = \frac{C}{A + C}$$

where:

FAR is the false alarm ratio.

A is the number of verified warnings.

C is the number of unverified warnings (also known as false alarms).

The CSI_{SB} for storm-based is computed directly from the POD_{SB} and *FAR* calculations for storm-based warnings.

$$CSI_{SB} = [(POD_{SB})^{-1} + (1 - FAR)^{-1} - 1]^{-1}$$

2.1.4 Lead Time

The methodologies for computing the lead time for tornado, severe thunderstorm, and generic severe thunderstorm/tornado events are identical.

a. Detailed Report: The detailed report lists each event and each warning. Lead times (in minutes) are found in the event list and are defined as the event time minus the warning issuance time, with more specifics to follow concerning the

event time. The time of warning issuance is taken from the WMO header of the warning, and the event times are taken from *Storm Data*.

(1) Lead Time of Instantaneous Events. A check is performed on instantaneous events to identify if a warning was valid at the time and location the event occurred. If a warning was valid at that time and location, the event is assigned a lead time based on the time the warning was issued. If no warning was valid at that time and location, the event is assigned a lead time of zero.

(2) Lead Time of Track Events. The methodology of how track events are evaluated in one-minute points is given in section 2.1.2. A check is performed at each one-minute point, for the duration of the event, to identify if a warning was valid at the time and location along the path that the event occurred. If a warning was valid at the time and location, the point is assigned a lead time based on the time the warning was issued. If no warning was valid at that time and location, the point is assigned a lead time of zero. This process is repeated at every point for the duration of a given event, and the lead time listed for that event is calculated by averaging the lead times stored for all one-minute segments of the track.

Based on the calculations described in (1) and (2), a *lead time* is listed for each event. The values in the *initial lead time* column are similar to those in the *lead time* column, except the initial one-minute track point lead time is used instead of the average lead time for a given track event. The *initial lead time* is generated so the NWS has a comparable lead time to the one calculated for county-based warning verification.

b. Summary Statistics: A *mean lead time* is calculated for the summary statistics by averaging the *lead times* listed in the detailed report. This includes all instantaneous and track events. The *initial lead time* in the summary statistics is calculated by averaging the *initial lead times* listed in the detailed report. This includes all instantaneous and track events.

2.1.5 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

2.2 County-based Warnings

County-based warning issuance ceased October 1, 2007, so county-based warning verification should be used for warnings issued before this date. Storm-based warning issuance commenced on October 1, 2007; see section 2.1 for a description of storm-based warning verification. NWS employees access verification statistics through the Severe Weather Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each data request, the user

provides the following definitions and boundaries:

a. Type of warning.

b. Beginning and ending dates.

c. One or more counties, WFOs, states, or NWS regions.

d. Severity of event, based on total cost of damage, number of fatalities, and/or tornado EF-scale (optional).

Verification statistics are computed for tornado and severe thunderstorm warnings and events using one of three methods. The user of *Stats on Demand* selects the method. The first method combines severe thunderstorms and tornadoes together and treats them as a single event type. The latter two methods are event specific—they treat non-tornadic severe thunderstorms and tornadoes as separate types of events. See Table A-2 (located in section 2.1.2) for illustration.

2.2.1 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the warning header (top two lines of the warning). Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

2.2.2 Matching Warnings and Events

All warning data are automatically extracted from the warning products issued to the public. Each county included in a warning statement is counted as a separate warning. The warning issuance and expiration times are taken from the VTEC line of the warning text.

All events are automatically taken from certified *Storm Data* reports. Each warning (SVR or TOR) is verified by a confirmed event of the type specified in Table A-2 (section 2.1.2). For verification purposes, multiple severe thunderstorm wind and hail events in the same county separated by less than 10 miles and 15 minutes are considered duplicates; only the first entry is used for verification. This rule has the following exceptions:

a. Any event that causes death or injury is included in the event database.

b. Any event that causes crop or property damage in excess of $500,000 is included in the event database.

c. Any report of winds 65 knots or greater is included in the event database.

d. Any hail size report of 2 inches or greater is included in the event database.

e. An event is not considered a duplicate if it is the only event verifying a warning.

Any event not recorded in the verification database due to the aforementioned duplicate rule may still appear in the publication *Storm Data*. An event moving into a second county creates an additional event for the database.

Warnings and events qualified for use in verification are recorded in separate databases. Whenever an event occurs in a warned county during any part of the valid period of the warning, one verified warning and one warned event are counted. Unwarned events and unverified warnings are also counted. From these databases, the *POD*, *FAR*, and *CSI* are computed (see sections 3.1 to 3.3 of Appendix A) and listed in the verification reports.

2.2.3 Lead Time

The methodologies for computing the lead time (in minutes) in each county for tornado, severe thunderstorm, and generic severe thunderstorm/tornado events are identical. For verification purposes, the definition of the term "event" is given in section 2.2.2. The lead time for each event is computed by subtracting the warning issuance time from the time when the event was first reported in the county. The warning issuance time is taken from the WMO header of the warning, and the start time of the event is taken from *Storm Data*. Negative lead times are set to zero. If one or more events occur in a county not covered by a warning, each unwarned event is assigned a lead time of zero. An event moving into a second county creates an additional event for the database. The lead time for the second event is based on the time the event first entered the second county. Average lead time is computed from all lead times listed in the event database, including zeroes. The percentage of events with a lead time greater than zero is also computed.

2.2.4 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

2.3 Watches

The Storm Prediction Center (SPC) is responsible for verifying the tornado and severe thunderstorm watches it issues. The area enclosed by a watch is verified without regard to the number of counties affected. Weiss et al. (1980) describes how SPC accounts for variations in the size of convective watch areas. All events are taken from *Storm Data*. Statistics are stratified for tornado and severe thunderstorm watches combined and for tornado watches only.

3 Marine Forecasts and Warnings

3.1 Wind and Wave Height Forecasts

Marine wind and wave forecasts are verified at fixed point locations out to Day 5 (waves) and Day 7 (winds). Forecasts are taken directly from the NDFD (grib2) files and matched in time and space to observation points, known as verification sites. The NDFD is issued hourly, but only the 0000 and 1200 UTC cycles are verified. Data in this program began October 1, 2013, but a legacy marine program contains data from October 1994 to July 2013. The legacy marine program is formatted in a similar but not identical manner to the current program.

3.1.1 Web Interface

NWS employees access verification statistics through the Marine Verification Home Page of the NWS Performance Management Website. The *Stats on Demand* interface under wind and wave forecast verification is selected. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. The user provides the following parameters:

a. Start and end dates. The user decides whether to query the database by (1) observation/forecast valid dates and times or (2) forecast issuance dates.

b. Forecast cycles of NDFD (0000 UTC and 1200 UTC). Matching guidance products usually come from cycles 12 hours older than the NDFD cycle. Forecast cycles only appear in the request Graphic User Interface (GUI) when the database is queried by forecast issuance dates.

c. Projections. Each forecast for a point in space and time is matched to the appropriate observation and, when requested, guidance product. The user requests projections by one or more hours, 12-hour period(s), or 24-hour "Day(s)," e.g., Day 1, Day 2. Waves are verified out to Day 5. Winds are verified out to Day 7.

d. Verification sites (single site, multiple sites, WFO/national center area, regional using NWS-defined regions, and NWS-wide).

e. Matching guidance product. At the request of the user, one guidance product at a time may be matched to temporally- and spatially-matched forecast/observation pairs. When selected, guidance products usually come from a cycle 12 hours earlier than the NDFD cycle.

3.1.2 Verification Sites

An interactive directory of legacy marine verification stations appears on the Marine Verification Home Page of the NWS Performance Management Website.

3.1.3 Verification Reports

Verification reports are prepared for wind direction, frequent wind gusts, sustained wind speed, or significant wave height. Each report contains verification statistics tailored to the parameters specified in the web interface (see section 3.1.1). Contingency tables, accuracy measures, skill scores, mean errors, and histograms of error categories are listed. The percentage of time with sustained wind speed or wind gust errors less than 5 knots, and the percentage of time with significant wave height errors less than 2 feet are part of the histograms and are shaded yellow. However, a graduated scale is used to count larger errors as "correct" whenever the observed wind speed or wind gust equals or exceeds 20 knots or the observed significant wave height equals or exceeds 10 feet; these "correct" percentages are shaded green. The *percent correct*

statistic is calculated from this histogram, and it appears near the end of each sustained wind speed or significant wave height data report; these values are used in Government Performance and Results Act of 1993 (GPRA) reports. The data report for each of the following elements includes a multi-category contingency table of forecasts versus observations; the category boundaries are listed in Table 5. See Table 6 for the rules concerning how observations are selected to verify the forecast at each valid time.

Table 5. Categories used in the marine forecast contingency table analyses.

Wind Speed	Wind Direction	Significant Wave Heights	Wind Gusts
• Under 7 knots • 7 to 12 knots • 13 to 17 knots • 18 to 22 knots • 23 to 27 knots • 28 to 33 knots • Over 33 knots	• North (338° to 22°) • Northeast (23° to 67°) • East (68° to 112°) • Southeast (113° to 157°) • South (158° to 202°) • Southwest (203° to 247°) • West (248° to 292°) • Northwest (293° to 337°)	• Under 2 feet • 2 to 3 feet • 4 to 5 feet • 6 to 8 feet • 9 to 12 feet • 13 to 18 feet • Over 18 feet	• Under 20 knots • 20 to 24 knots • 25 to 33 knots • 34 to 47 knots • Over 47 knots

a. Sustained Wind Speed. Each 3-hourly forecast (6-hourly beyond the 60-hour projection) and, when it is requested, the matching guidance product used to prepare that forecast, are matched in time and space to the appropriate verifying observation(s). Appropriate observations are defined in Table 6 (section 1). Anemometers vary with height so each observed wind speed is corrected to 10 meters above sea level, except for Great Lakes wind speeds, which are corrected to 10 meters above lake level. Liu et al. 1979 is used for these altitude corrections; when the air and sea surface temperatures are available, this algorithm uses them to incorporate the low-level stability.

b. Wind Direction. Each 3-hourly forecast (6-hourly beyond the 60-hour projection) and, when it is requested, the matching guidance product used to prepare that forecast, are matched in time and space to the appropriate verifying observation(s). Appropriate observations are defined in Table 6 (section 1); however, each forecast matched with an observed sustained wind speed under 7 knots is not verified.

c. Significant Wave Heights. Each 6-hourly forecast and, when it is requested, the matching guidance product used to prepare that forecast, is matched in time and space to the appropriate verifying observation(s). Appropriate observations are defined in Table 6.

d. Wind Gusts. Each 3-hourly forecast and, when it is requested, the matching guidance product, are matched in time and space to the appropriate verifying observation(s). Appropriate observations are defined in Table 6. The same altitude correction used for sustained wind speed is also applied to wind gusts.

Table 6. Algorithm used to select wind, wave height, and wind gust observations.

1. The National Data Buoy Center (NDBC)* Standard Meteorological Data (SMD) are used to verify wind direction, sustained wind speed, and significant wave height with the observation closest in time to the forecast valid time; however, the observation time must be within ± 2 hours of the forecast valid time. Whenever two different observation times are competing to be matched to a forecast, and those times are equidistant from the forecast valid time (e.g., 1750 vs. 1810 for the 1800 UTC forecast valid time), the earlier observation (e.g., in this case, the 1750 UTC observation) is used. If the observations for a verification site are not archived on the NDBC web site (e.g. *WeatherFlow*), a similar algorithm is followed; however, data frequency varies from one data source to another.

Example: At NDBC-owned data sites*, a forecast valid at 1800 UTC is usually verified with the 1750 UTC observation; these observations are archived once per hour. One or more missing reports result in the following ordinal priorities. If the 1750 UTC observation is missing, the 1850 UTC observation, if available, is used. Then, if the 1850 UTC observation is missing, the 1650 UTC observation, if available, is used. Then, if the 1650 UTC observation is missing, the 1950 UTC observation, if available, is used. Then, if the 1950 UTC observation is missing, the forecast valid at 1800 UTC is not verified.

2. For non-NDBC owned sites*, the SMD are also used to verify <u>frequent</u> wind gusts. The frequent wind gusts at a given forecast valid time, e.g., 1800 UTC, are approximated by using the median value from all wind gust reports within ± 1 hour of the forecast valid time. If one or more normally reported wind gust observations are missing from that time window, then the frequent wind gusts are approximated by using the median value from all wind gust reports within ± 2 hours of the forecast valid time. At least two wind gust reports must exist within that time window, or the observation is recorded as missing, and the corresponding forecast is not verified. If the observations for a verification site are not archived on the NDBC web site (e.g. *WeatherFlow*), a similar algorithm is followed; however, data frequency varies from one data source to another.

Example (1): To estimate the frequent wind gusts at 1800 UTC from a National Ocean Service (NOS) station that reports data in SMD format every six minutes, the median value from the gusts reported at the following times is used: 1700, 1706, 1712, … , 1848, 1854, and 1900 UTC. If one or more of those normally reported observations are missing, then the median value from the gusts reported at the following times is used: 1600, 1606, 1612, … , 1948, 1954, and 2000 UTC.

Example (2): To estimate the frequent wind gusts at 1800 UTC from a point that reports data in SMD format every half hour at 20 and 50 minutes past each hour, the median gust value from the following observations is used: 1720, 1750, 1820, and 1850 UTC. If one or more of those normally reported wind gust observations are missing, then the median gust value from the following set of observations is used: 1620, 1650, 1720, … , 1850, 1920, and 1950 UTC.

3. For NDBC-owned sites, the Continuous Wind Data (CWD) are used to verify frequent wind gusts; the SMD are used if the CWD are not available. Wind observations are archived every 10 minutes, and once per hour, usually at 50 minutes past the hour, the peak gust from the past hour is reported in the CWD gust column. At these sites, the peak hourly wind gust reported at the time closest to the forecast valid time is compared to the peak hourly wind gust reported one hour later. The frequent wind gusts at the forecast valid time are approximated by using the lesser of the two peak hourly values. If one of those values is missing, then the lesser of two alternative consecutive hourly reports within ± 2 hours of the forecast valid time is used, following the ordinal pattern in this example:

Example: To approximate the frequent wind gust reported at 1800 UTC, take the lower of the peak hourly wind gust values reported in the CWD data at 1750 and 1850 UTC. Then, if either of those peak hourly values is missing, the lower of the following peak gust reports, if both are available, are used: 1650 and 1750 UTC. Then, if either of those peak hourly values is missing, the lower of these alternative peak gust reports, if both are available, are used: 1850 and 1950 UTC. Then, if either of those peak hourly values is missing, the frequent wind gust estimate at 1800 UTC is recorded as missing, and the forecast valid at 1800 UTC is not verified.

* NDBC quality controls and archives meteorological and significant wave height data from instruments owned by NDBC. NDBC also archives data from other organizations; however, NDBC does not maintain or quality control data they do not own. National Ocean Service coastline stations and SCRIPPS wave-rider buoys are examples of such stations. The Northeast Regional Association of Coastal Ocean Observing Systems (NERACOOS) buoys are not quality controlled nor archived by NDBC. Hence, NERACOOS data must be collected in the near real time.

3.2 Coastal Flood and Lakeshore Flood Warnings (CFW)

NWS employees access these verification statistics through the Marine Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each request, the user provides the following boundaries:

a. Beginning and ending dates.

b. One or more zones, WFOs, states, or NWS regions.

3.2.1 Matching Warnings and Events

All warning data are automatically taken from the warning products issued to the public. Each public forecast zone is treated as a separate verification area. Therefore, a warning covering three zones is counted as three warned areas or three warnings.

a. All events are automatically taken from certified *Storm Data* reports. Only the following reportable events are used to verify a CFW:

   (1) Coastal Flood.

   (2) Lakeshore Flood.

   (3) Seiche.

b. See NWSI 10-1605, Storm Data Preparation, for descriptions of each of the above events. Minor coastal or lakeshore flooding, such as nuisance flooding, is treated as a non-event for verification purposes. The following event times, defined in NWSI 10-1605, Storm Data Preparation, are provided for each event listed in *Storm Data* and are used in verification:

   (1) Beginning time.

   (2) Ending time.

c. Warnings and reportable events are recorded in separate verification databases. Whenever a reportable event (defined as the period between its beginning and ending times) coincides with any part of the valid period of a warning, one warned event and one verified warning are counted. Unwarned events and unverified warnings are also counted. From these databases, the *POD*, *FAR*, and *CSI* are computed (see sections 3.1 to 3.3 of Appendix A) and listed in the verification reports.

3.2.2 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

3.2.3 Extensions

Warnings may be extended in area and/or time. Extensions of warnings to new areas (zones) are counted as new warnings, i.e., one warning per zone. Each time extension of a zone already warned is counted as a new warning only if the earlier warning did not verify during its valid period.

3.2.4 Lead Time

A lead time (in hours) is computed for each zone that experiences a reportable event. If the event beginning time does not occur during the valid period of a warning, the lead time for that event is zero. If the event beginning time occurs during the valid period of a warning, the lead time for that event is computed by subtracting the warning issuance time from the event beginning time. The warning issuance time comes from the WMO header of the CFW. Negative lead times are set to zero. The average lead time is computed from all lead times listed in the event database, including zeroes.

3.2.5 Timing Error

The timing error (in hours) for each warned event is defined as the event beginning time minus the warning beginning time. For each data request, the mean absolute error, the mean algebraic error (bias), and a distribution of errors are provided.

3.2.6 Watches

While watches are not verified in the same manner as warnings, the percentage of unwarned events that occurred with a watch in effect is provided.

3.2.7 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

3.3 Storm-Based Special Marine Warnings (SMW)

Any SMW issued for a coastal or Great Lake marine zone, Lake Okeechobee, or Lake Pontchartrain is verified. Storm-based SMW issuance replaced marine zone-based SMW issuance October 1, 2007, so storm-based SMW verification should be used for warnings issued on or after this date. For warnings issued before October 1, 2007, see section 3.4 for a description of marine zone-based SMW verification.

NWS employees access storm-based SMW verification statistics through the Marine Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an

interactive database to provide verification statistics customized to the user's request. With each request, the user provides the following definitions and boundaries:

a. Beginning and ending dates.

b. One or more WFOs or NWS regions. National statistics are also available.

c. Severity of event, based on total cost of damage and/or number of fatalities (optional).

3.3.1 Quality Assurance

All data imported into the warning database are taken directly from the warning text. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

3.3.2 Matching Warnings and Events

All warning data are automatically extracted from the warning products issued to the public. The basic area for a SMW is the polygon boundary outlined by the latitude-longitude coordinates located at the bottom of the product. Therefore, for verification purposes, the area within the latitude-longitude boundaries is counted as the warning.

Only the following reportable events in certified *Storm Data* reports occurring within the temporal and areal boundaries of an SMW verify that warning:

- Marine hail, 3/4 inch or greater.
- Marine thunderstorm wind, 34 knots or greater.
- Waterspouts.
- Marine strong wind.
- Marine high wind.

Each warning is checked to see if a verifying event occurred within its temporal and areal boundaries and is categorized as verified or unverified.

Events are logged in *Storm Data* using one of two methods. The first method is an isolated event at a single location (referred to as an instantaneous event). An example would be an isolated marine hail event reported at a single time. The second method is used for an event that starts at one location and moves to another location over a period of time (referred to as a track event). An example would be a waterspout that moved from one location to another. Both methods are evaluated differently.

a. Evaluation of Instantaneous Events. A check is performed on each instantaneous event to see if a warning was in effect at the time and location of the event. If so, the event was warned. If not, the event was unwarned.

b. Evaluation of Track Events. Before track events can be evaluated, two assumptions are made:

(1) The event travels in a straight path between the event beginning and ending locations logged in *Storm Data.*

(2) The event travels at a constant speed between the event beginning and ending locations logged in *Storm Data.*

Once these assumptions are made, the location of the event is estimated every minute for the duration of the event. The event is then evaluated at each of those locations and times. For example, a tornado event lasting from 0100 to 0105 and traveling three miles would be evaluated at six locations and times, i.e., six segments of the event. A check is then performed at each point along the track of the event to see if a warning was in effect. If so, the event was warned at that point. If not, the event was unwarned at that point.

Next, the ***percentage of the event warned*** *(PEW)* is calculated for each event. For an instantaneous event, the *PEW* is zero for an unwarned event and 100 for a warned event. For a track event, the *PEW* is calculated linearly, dividing the total number of warned one-minute segments by the total time length of the event. In the case of the example in the previous paragraph (a tornado lasting from 0100 to 0105), if the tornado was inside the warning polygon at 0100 and 0101 (the warned segments of the event) and moved outside the polygon starting at 0102 (the unwarned segments of the event lasted from 0102 to 0105), the *PEW* for the entire event would be two divided by six or 33.3 percent.

3.3.3 *POD, FAR,* and *CSI* Calculations

Once a *PEW* is calculated for each event, the *POD* may be calculated:

$$POD_{SB} = \frac{0.01}{N} \sum_{i=1}^{N} PEW_i$$

where:

POD is the probability of detection,

SB is for storm-based warnings,

PEW is the total percentage of each event, *i*, warned, expressed as a value from zero to 100, and

N is the total number of events.

The best possible *POD* is one; the worst is zero.

The *FAR* for storm-based warnings is the same calculation that was used for the old marine zone-based system:

$$FAR = \frac{C}{A + C}$$

where:

FAR is the false alarm ration.

A is the number of verified warnings.

C is the number of unverified warnings (also known as false alarms).

The CSI_{SB} for storm-based is computed directly from the POD_{SB} and *FAR* calculations for storm-based warnings.

$$CSI_{SB} = [(POD_{SB})^{-1} + (1 - FAR)^{-1} - 1]^{-1}$$

3.3.4 Lead Time

a. Detailed Report: The detailed report lists each event and each warning. Lead times (in minutes) are found in the event list and are defined as the event time minus the warning issuance time, with more specifics to follow concerning the event time. The time of warning issuance is taken from the WMO header of the warning, and the event times are taken from *Storm Data*.

(1) Lead Time of Instantaneous Events. A check is performed on instantaneous events to identify if a warning was valid at the time and location the event occurred. If a warning was valid at that time and location, the event is assigned a lead time based on the time the warning was issued. If no warning was valid at that time and location, the event is assigned a lead time of zero.

(2) Lead Time of Track Events. The methodology of how track events are evaluated in one-minute points is given in section 3.3.2 of this manual. A check is performed at each one-minute point, for the duration of the event, to identify if a warning was valid at the time and location along the path that the event occurred. If a warning was valid at the time and location, the point is assigned a lead time based on the time the warning was issued. If no warning was valid at that time and location, the point is assigned a lead time of zero. This process is repeated at every point for the duration of a given event, and the lead time listed for that event is calculated by

averaging the lead times stored for all one-minute segments of the track.

Based on the calculations described in (1) and (2), a *lead time* is listed for each event. The values in the *initial lead time* column are similar to those in the *lead time* column, except the initial one-minute track point lead time is used instead of the average lead time for a track event. The *initial lead time* is generated so the NWS has a comparable lead time to the one calculated for county-based warning verification.

b. Summary Statistics: A *mean lead time* is calculated for the summary statistics by averaging the *lead times* listed in the detailed report. This includes all instantaneous and track events. The *initial lead time* in the summary statistics is calculated by averaging the *initial lead times* listed in the detailed report. This includes all instantaneous and track events.

3.3.5 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

3.4 Zone-Based SMWs

For verification of SMWs issued on or after October 1, 2007, see section 3.3, Storm-based SMW Verification.

Zone-based SMWs were issued before October 1, 2007. Therefore, all SMWs issued for a coastal or Great Lake marine zone, Lake Okeechobee, or Lake Pontchartrain is verified with these rules.

NWS employees access marine zone-based SMW verification statistics through the Marine Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each data request, the user provides the following definitions and boundaries:

a. Beginning and ending dates.

b. One or more marine zones, WFOs, bodies of water, or NWS regions. National statistics are also available.

c. Severity of event, based on total cost of damage and/or number of fatalities (optional).

3.4.1 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

3.4.2 Matching Warnings and Events

All warning data are automatically taken from the warning products issued to the public. Each marine forecast zone represents a separate verification area. Therefore, a warning issued for two zones counts as two separate warnings. Only the following reportable events in the certified *Storm Data* reports verify the SMW:

a. Marine hail, 3/4 inch or greater.

b. Marine thunderstorm wind, 34 knots or greater.

c. Waterspouts.

d. Marine strong wind.

e. Marine high wind.

Warnings and reportable events are recorded in separate databases. Whenever a reportable event occurs in a warned marine zone during any part of the valid period of the warning, one verified warning and one warned event are counted. Unwarned events and unverified warnings are also counted. From these databases, the *POD*, *FAR*, and *CSI* are computed (see Appendix A, sections 3.1 to 3.3) and listed in the verification reports.

3.4.3 Lead Time

The lead time (in minutes) for each reportable event is computed separately for each marine zone by subtracting the time of warning issuance from the time when the reportable event was first reported in the marine zone. The time of warning issuance is taken from its WMO header, and the time when the reportable event was first reported in the marine zone is taken from *Storm Data*. Negative lead times are set to zero. If one or more events occur in a zone with no warning in effect, each unwarned event is assigned a lead time of zero. Average lead time is computed from all lead times listed in the event database, including the zeroes. The percentage of events with a lead time greater than zero is also computed and listed in the verification reports.

3.4.4 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

4 Hydrologic Warnings and Forecasts

Hydrologic verification consists of the verification of county-based and storm-based flash flood warnings (FFW) and point-based flood warnings (FLW).

4.1 Storm-Based FFWs

See section 4.2 for a description of county-based FFW verification.

NWS employees access these verification statistics through the Hydrology Verification of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each request, the user provides the following definitions and boundaries:

a. Beginning and ending dates.

b. One or more WFOs or NWS regions.

c. Severity of event, based on total cost of damage and/or number of fatalities (optional).

4.1.1 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

4.1.2 Matching Warnings and Events

All warning data are automatically extracted from the warning products issued to the public. The basic area for a FFW is the polygon boundary outlined by the latitude-longitude coordinates located at the bottom of the product. Therefore, for verification purposes, the area within the latitude-longitude boundaries is counted as the warning.

All event data are automatically taken from certified *Storm Data* reports of the event type "flash flood." Each warning comes from an FFW product, and each FFW is checked to see if a confirmed event occurred within its temporal and areal boundaries. Each warning is thereby categorized as verified or unverified.

Events are logged in *Storm Data* as areal events, which means that each event is entered for an area of land. The area of the event reported and the forecast area of the warning are overlaid to compute the ***percentage of the event warned*** *(PEW).*

4.1.3 *POD, FAR,* and *CSI* Calculations

Once a *PEW* is calculated for each event, the *POD* is calculated:

$$POD_{SB} = \frac{0.01}{N} \sum_{i=1}^{N} PEW_i \quad ,$$

where:

POD is the probability of detection,

SB is for storm-based warnings,

PEW is the total percentage of each event, i, warned, expressed as a value from zero to 100, and

N is the total number of events.

The best possible *POD* is one; the worst is zero.

The *FAR* for storm-based warnings is the same calculation that was used for the old county-based system:

$$FAR = \frac{C}{A + C} \ ,$$

where:

FAR is the false alarm ratio.

A is the number of verified warnings.

C is the number of unverified warnings (also known as false alarms).

The CSI_{SB} for storm-based is computed directly from the POD_{SB} and *FAR* calculations for storm-based warnings.

$$CSI_{SB} = [(POD_{SB})^{-1} + (1 - FAR)^{-1} - 1]^{-1} \quad .$$

4.1.4 Lead Time

The lead time for flash floods (in minutes) is called the ***maximum event lead time.*** The verification matching process generates this maximum event lead time for all flash flood events. It is calculated by analyzing every flash flood event to see if a warning was valid over any area of the event when the event first began. Any amount of areal overlap of the warning and event polygons is acceptable. The key is to use only those warnings valid for the time and area <u>when the event first begins</u>.

There are three possible scenarios:

a. <u>No warning is valid for the area in which an event begin</u>s. The Maximum Event Lead Time equals zero minutes. This also applies to situations where warnings are issued after the event has already begun.

b. <u>A single warning is valid for the area in which an event begins</u>. The Maximum Event Lead Time equals the Event Beginning time minus the Warning Issuance

Time.

c. Multiple warnings are valid for the area in which an event begins. In this case, the lead time calculation is based only on the warning that was issued earliest (first). The Maximum Event Lead Time equals the Event Beginning Time minus the Warning Issuance Time (of first issued warning).

The time of warning issuance is taken from the WMO header of the FFW, and the event beginning time is taken from *Storm Data*. Negative lead times are set to zero.

If 100% of the event was not warned (see the last paragraph in section 4.1.2), the maximum event lead time may not be representative of the entire area of the event. In these situations, an ***area weighted lead time*** is calculated to correct for the portion of the flooded area that was not warned. For example, the percentage of an event warned was 83, and the lead time across the warned portion of the flooded area was 60 minutes. To properly account for the 17% of the event not warned, the 60-minute lead time is multiplied by 83%, resulting in an area weighted lead time equal to 49.8 minutes.

4.1.5 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

4.2 County-Based FFWs

See section 4.1 for a description of storm-based FFW verification.

NWS employees access these verification statistics through the Hydrology Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each request, the user provides the following definitions and boundaries:

a. Beginning and ending dates.

b. One or more counties, WFOs, states, NWS regions, or the contiguous United States.

c. Severity of event, based on total cost of damage and/or number of fatalities (optional).

County-based FFW verification is performed to generate statistics that can be compared to those produced since 1986, which can be used in the analysis of long-term trends in warning quality.

4.2.1 Quality Assurance

All data imported into the warning database are taken directly from the warning. The issuing WFO and warning type in the VTEC line are checked for consistency with the WMO header.

Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

4.2.2 Matching Warnings and Events

All warning data are automatically extracted from the warning products issued to the public. Since each county specified in a warning represents a separate verification area, a warning covering three counties is counted as three warnings. Events are automatically taken from certified *Storm Data* reports. *Storm Data* reports entered as the event type "flash flood" verify an FFW.

Warnings and events are recorded in separate databases. Whenever an event occurs in a warned county during any part of the valid period of the warning, one verified warning and one warned event are counted. Unwarned events and unverified warnings are also recorded and tallied. From these databases, the *POD*, *FAR*, and *CSI* are computed (see Appendix A, sections 3.1 to 3.3) and listed in the verification reports.

4.2.3 Lead Time

For verification purposes, the definition of the term "event" is given in section 4.2.2. The lead time (in minutes) for each flash flood event is computed separately for each county by subtracting the time of warning issuance from the time when the event first occurred in the county. The time of warning issuance comes from the WMO header of the FFW, and the event beginning time for the given county is taken from *Storm Data*. Negative lead times are set to zero. If one or more events occur in a county with no warning in effect, each unwarned event is assigned a lead time of zero. Average lead time is computed from all lead times listed in the event database, including zeroes. The percentage of events with lead time greater than zero is also computed.

4.2.4 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, listed in the WMO header of the warning.

4.3 Point-Based Flood Warnings (FLW)

This section is only for point flood warnings and does not apply to areal FLWs, for which verification does not yet exist.

NWS employees access these verification statistics through the Hydrology Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to provide verification statistics customized to the user's request. With each request, the user inputs one or more of the following to define the search:

a. Beginning and ending dates.

b. Search area: National, NWS region, RFC, WFO, one or more locations in a WFO area, one or more locations in a RFC area, or one or more locations in a state.

c. Category of river response time to significant input (e.g., heavy rainfall): slow (greater than 60 hours), medium (24 to 60 hours), or fast (less than 24 hours).

4.3.1 Quality Assurance

All data imported into the warning and event databases are taken directly from the warning. The issuing WFO and warning type in the Primary VTEC (P-VTEC) line are checked for consistency with the WMO header. Inconsistent warnings are not counted for verification, and products issued with the improper coding may not be correctly imported into the database.

4.3.2 Matching Warnings and Events

All warning data are automatically extracted from the warning products issued to the public. Since each river forecast point specified in a warning represents a separate verification point, a warning covering three river forecast points is counted as three verifiable warnings. Flood events are automatically defined using the Flood Beginning Date/Time (Z_B) and Flood Ending Date/Time (Z_E) indicators of the Hydrologic VTEC (H-VTEC) string in the final flood statement (FLS) product issued after the original FLW product.

During any part of the valid period of the warning, whenever a flood event begins, which is defined to occur when flood stage is reached at a river/stream forecast point, one "verified warning" and one "event <u>during</u> a valid warning" are counted. Whenever a flood event begins at a warned point before any part of the valid period of a warning, one "verified warning" and one "event <u>before</u> a valid warning" are counted. Unverified warnings are also recorded and tallied. Since the verification system can only define flood events through the use of H-VTEC strings in FLS products following FLW products, there is no mechanism for storing information in the database on flood events for which a warning was never issued. Therefore, it is not possible to compute the *POD* and *CSI* statistics. From the warning database, the *FAR* and the Frequency of Hits are computed and listed in the verification reports (see Appendix A, sections 3.2 and 3.4, respectively).

4.3.3 Lead Time

For verification purposes, the definition of the term "event" is given in section 4.3.2. The lead time (in hours) for each flood event is computed separately for each river point by subtracting the time of first warning issuance from the time when the flood event first occurred at the river point. The time of warning issuance comes from the WMO header of the FLW, and the event beginning time for the given river point is taken from the Flood Beginning Date/Time Indicator (Z_B) of the H-VTEC code in the final FLS product issued after the original FLW product. Negative lead times are set to zero. If an event begins at a location prior to a warning being valid, the event is assigned a lead time of zero. Average lead time is computed from all lead times listed in the event database, including zeroes.

4.3.4 Absolute Timing Error

The absolute timing error (in hours) for each warned event is defined as the measure of the absolute difference between the time when flooding was first observed and the time when

flooding was first forecast to have occurred in the initial warning for that event. An absolute timing error close to zero is desirable. For each data request, the average absolute timing error is provided.

4.3.5 Backup Mode for Warnings

All warnings issued by the backup office are attributed to the primary WFO, as listed in the WMO header of the warning.

4.4 RFC River Stage Forecasts

The RFCs operate the river stage forecast verification software, and the Water Resources Support Branch (within the Analyze, Forecast and Support Office) maintains policy. For a selected set of locations, both stream water level observations (stage) and stage forecasts issued by RFCs are posted to a verification database at each RFC. Forecast values are matched with concurrent observations. From these pairs, verification statistics measuring the performance of the forecast system are calculated. The initial phase of river forecast verification is based on calculations of mean, mean absolute, and root mean square differences between observed and forecast values for each verification site on the river. RFCs automatically transmit their monthly river stage forecast verification statistics to the Performance and Evaluation Branch by the 20th of the subsequent month.

NWS employees access verification statistics on RFC river stage forecasts through the Hydrology Verification Home Page of the NWS Performance Management Website. *Stats on Demand* uses an interactive database to generate verification statistics customized to the user's request. The system allows verification statistics for locations to be grouped together by forecast lead time as well as hydrologic characteristics, i.e., (1) locations responding rapidly to rainfall, (2) locations with intermediate responses, and (3) locations with slow responses.

5 Quantitative Precipitation Forecasts (QPF)

5.1 Weather Prediction Center (WPC) QPFs

The WPC computes verification statistics for its QPFs and corresponding model QPFs. These data have been calculated since 1971 and are posted on the WPC Website.

6 Aviation Forecasts

6.1 Terminal Aerodrome Forecasts (TAF)

This *Stats on Demand* program is the official NWS TAF verification tool. TAFs are evaluated every five minutes, twelve times per hour or 288 times for an entire 24-hour TAF. The 5-minute interval times end in either a "0" or "5." Forecast conditions at the end of each 5-minute interval are matched with the most recently reported METAR/SPECI, and each element (e.g., ceiling) is verified separately. Routine hourly METARs not received just before the hour are assumed to be missing, and all 5-minute verification intervals following that scheduled METAR are discarded as missing until a new METAR or SPECI is reported.

6.1.1 Verification Sites

All terminals for which the NWS issues TAFs may be verified. An interactive station directory appears on the Aviation Verification Home Page of the NWS Performance Management Website.

6.1.2 Data

The Performance and Evaluation Branch automatically collects all data from operational products. Forecast data come from the TAFs and observation data come from the METAR/SPECIs. All METARs and SPECIs are tested for reliability and consistency, and suspicious data are removed. A description of these quality control algorithms is found through a link on the Aviation Verification Home Page of the NWS Performance Management Website.

Each WFO may occasionally need to request one or more deletions to the observation database after becoming aware that one or more corrupted Meteorological Aviation Reports (METAR) or Special Aviation Weather Reports (SPECI) were issued for a point in its county forecast and warning area. To ensure that the verification database agrees with the records at the National Climatic Data Center (NCDC), the WFO should email a WS Form B-14 (Notice of Corrections to Weather records) to Surface.QC@noaa.gov and NWS.Verification@noaa.gov. If a Datzilla ticket is required instead of a B-14, the WFO should forward a copy of the Datzilla receipt to NWS.Verification@noaa.gov. These reports should be specific and state the exact problem(s) with each observation reported.

Guidance data come from available alphanumeric MOS and Local AWIPS MOS Program (LAMP) products. The latest version of guidance available at TAF issuance time is used. The persistence forecast, defined as the observed conditions at TAF issuance time, is also available. Forecaster identification, when appropriate, is read from a separate AWIPS product transmitted by the WFO with the WMO header: NTXX98 Kccc, where ccc is the WFO forecast office identifier.

6.1.3 Web Interface

NWS employees access TAF verification statistics through the Aviation Verification Home Page of the NWS Performance Management Website. The *Stats on Demand* interface under TAF

Verification is selected. The user is able to request data for any TAF element (see section 6.1.6), a single forecast type (e.g., prevailing, TEMPO) and, if desired, corresponding data from a single guidance product, as defined in section 6.1.2, for one or more:

a. Dates, defined by the beginning and ending dates (format *mm/dd/yyyy).* Data more than 18 months old are only available in terms of full months.

b. Scheduled TAF beginning times, i.e., 0000, 0600, 1200, 1800 UTC. Amended TAFs are grouped with the scheduled TAF issued before them. For example, TAF amendments issued between 0000 and 0600 UTC are grouped as with the 0000 UTC scheduled TAF.

c. Projection period groups (see section 6.1.4).

d. Verification sites (single site, multiple sites, WFO forecast area, states, NWS regions).

The user of *Stats on Demand* also specifies one of the following options concerning scheduled and amended TAFs: (a) verify scheduled TAFs only, (b) verify amended TAFs only, or (c) verify scheduled and amended TAFs. Amended TAFs are only verified for the first six hours of the valid period.

For ceiling, visibility, and flight category, the categories in the contingency tables are fixed; however, an optional two-category contingency table is displayed near the end of the verification report if requested by the user. The user can choose to

a. not receive the two-category contingency table (select none for "ceilings below" and/or "visibilities below"),

b. receive the two-category contingency table(s) by selecting a threshold value for ceiling (next to "ceilings below") and another for visibility (next to "visibilities below") to define the breakpoints between the two categories, or

c. receive two-category contingency table(s) by selecting "published local airfield minima." These minima (one for ceiling and one for visibility) define the upper and lower categories significant to operations.

For example with option b., if the element is ceilings, the user may define the two categories as less than 700 feet (lower category) and 700 feet or greater (upper category). If the element is visibility, the user may define the two categories as less than one statute mile (lower category) and one statute mile or greater (upper category). If the element is flight category, the user would need to pick the breakpoints between the two categories for ceilings <u>and</u> visibilities, e.g., (1) ceilings below 1200 feet <u>or</u> visibilities below 2 statute miles (lower category), (2) ceilings at least 1200 feet <u>and</u> visibilities at least 2 statute miles (upper category).

6.1.4 Projections

Scheduled TAFs are issued and verified for projections of 24 or 30 hours beyond the initial valid time of the most recent scheduled TAF. For verification purposes, projections are defined from the initial valid time of the TAF, which is 0000, 0600, 1200, or 1800 UTC for scheduled TAFs and the issuance time for amendments.

a. When the user requests verification statistics for scheduled TAFs only, he/she may select one or more of the following projection period groupings:

- Greater than zero to 3 hours.
- Greater than 3 to 6 hours.
- Greater than 6 to 9 hours.
- Greater than 9 to 12 hours.
- Greater than 12 to 18 hours.
- Greater than 18 to 24 hours.
- Greater than 24 to 30 hours.

b. When the user requests verification statistics for amended TAFs only or scheduled and amended TAFs combined, he/she selects one or both of the following projection periods:

- Greater than zero to 3 hours.
- Greater than 3 to 6 hours.

6.1.5 Verification Reports

Verification reports are prepared for each TAF element, and each report contains verification statistics tailored to the parameters specified in the web interface. Accuracy measures and skill scores are listed in these reports. All elements, except for the significant weather types, are verified in multi-category contingency tables, with the respective element divided into fixed categories. For ceiling, visibility and flight category, the user may request an optional two-category contingency table near the end of the report, where the two categories are not fixed but defined by the person requesting the verification statistics. More details are provided in sections 6.1.6.1 through 6.1.6.3.

Since forecasts are evaluated every 5 minutes, the contingency tables usually contain twelve entries per hour per verification site. Through a set of switches in the report header, the user can set the contingency tables to display data in terms of frequency (number of 5-minute intervals), number of hours, number of minutes, or percentage of the contingency table total.

6.1.6 Elements

The user of *Stats on Demand* specifies a single element with each request. To receive results for multiple elements, the user runs *Stats on Demand* separately for each element desired.

6.1.6.1 Ceiling

Ceiling is recorded in hundreds of feet but verified in fixed categories (see Table 7).

Table 7. Ceiling categories used in contingency tables.

• Less than 200 feet
• 200 to 400 feet
• 500 to 900 feet
• 1000 to 1900 feet
• 2000 to 3000 feet
• Greater than 3000 feet (including situations with no ceiling)

At the request of the user of *Stats on Demand,* an optional 2-category verification is also available in the same verification report, where the two categories are defined by the user:

- High category: ceilings at or above *x* hundred feet,
- Low category: ceilings below *x* hundred feet.

where, *x* is selected by the user.

6.1.6.2 Visibility

Visibility is recorded in the database in statute miles and fractions thereof. For observations taken on or after October 19, 2007, the surface visibility is always used. Whenever the surface visibility is higher than the tower visibility, the surface visibility is taken from the remarks section of the observation. For observations taken before October 19, 2007, the visibility reported in the main body of the observation is used, regardless of whether it was a surface or tower visibility. Six fixed categories for the contingency tables are defined in Table 8.

Table 8. Visibility categories used in contingency tables.

• Less than ½ statute mile
• ½ to less than 1 statute mile
• 1 to less than 2 statute miles
• 2 to less than 3 statute miles
• 3 to 5 statute miles
• Greater than 5 statute miles

At the request of the user of *Stats on Demand,* an optional 2-category verification is also available in the same verification report, where the two categories are defined by the user:

- High category: visibilities at or above *y* statute miles,
- Low category: visibilities below *y* statute miles.

where, *y* is selected by the user.

6.1.6.3 Flight Category

Flight Category is determined from Table 9. The categories for ceiling (e.g., 1500 feet, MVFR) and visibility (e.g., ¾ statute mile, LIFR) are first determined separately. The overall flight category is defined as the lower category of the two. Using the examples of ceiling and visibility cited in this paragraph, an overall LIFR flight category results.

Table 9. Use of ceiling and visibility to determine flight category.

CATEGORY	CEILING (feet)	VISIBILITY (statute miles)
Very Low Instrument Flight Rules (VLIFR)	less than 200	less than ½
Low Instrument Flight Rules (LIFR)	200 to 400	½ to less than 1
Instrument Flight Rules (IFR)	500 to 900	1 to less than 3
Marginal Visual Flight Rules (MVFR)	1000 to 3000	3 to 5
Visual Flight Rules (VFR)	no ceiling or greater than 3000	greater than 5

At the request of the user of *Stats on Demand*, an optional 2-category verification is also available in the same verification report, where the user selects a critical threshold value for ceiling (*x* hundred feet) and a critical threshold value for visibility (*y* statute miles) to define the two categories:

- High category: ceilings greater than or equal to *x* hundred feet <u>and</u> visibilities greater than or equal to *y* statute miles.
- Low category: ceilings less than *x* hundred feet <u>or</u> visibilities less than *y* statute miles.

6.1.6.4 Wind Direction

Wind direction is recorded in the database in degrees of the compass, adjusted to true-north. The categories in the multi-category contingency table are defined by the eight points of the compass and are found in Table 10.

Table 10. Wind direction categories used in contingency tables.

- North (340 to 20 degrees)
- Northeast 30 to 60 degrees
- East (70 to 110 degrees).
- Southeast (120 to 150 degrees)
- South (160 to 200 degrees)
- Southwest (210 to 240 degrees)
- West (250 to 290 degrees)
- Northwest (300 to 330 degrees)

6.1.6.5 Sustained Wind Speed

Sustained wind speed is recorded in the database in knots. The categories in the multi-category contingency tables are fixed with arbitrary limits defined in Table 11.

Table 11. Sustained wind speed categories used in contingency tables.

• Less than 8 knots
• 8 to 12 knots
• 13 to 17 knots
• 18 to 22 knots
• 23 to 27 knots
• 28 to 32 knots
• Greater than 32 knots

6.1.6.6 Wind Gusts

Wind gusts are recorded in the database in knots. The categories in the multi-category contingency tables are fixed with arbitrary limits defined in Table 12.

Table 12. Wind Gust categories used in contingency tables.

• No gusts or less than 18 knots
• 18 to 22 knots
• 23 to 27 knots
• 28 to 32 knots
• 33 to 37 knots
• 38 to 42 knots
• 43 to 47 knots
• Greater than 47 knots

6.1.6.7 Significant Weather Types

Each significant weather type is verified separately with a 2-category contingency table of forecasts versus observations. A second 2-category contingency table (guidance versus observations) is provided for each element if guidance products were included in the request. The two categories comprising each of these contingency tables are occurrence and non-occurrence of the weather type. Precipitation intensity is not verified. Note: To get the most complete set of scores, this element should be verified *without* guidance since all guidance products issue these forecasts for a very limited number of weather types.

- Liquid precipitation—rain (RA), rain showers (SHRA), drizzle (DZ).
- Snow types—snow (SN), snow showers (SHSN), snow grains (SG).
- Ice types, i.e., ice crystals (IC), ice pellets (PL), showers of ice pellets (SHPL), small (less than 1/4 inch diameter) hail/snow pellets (GS), showers of GS (SHGS).
- Freezing precipitation—freezing rain (FZRA), freezing drizzle (FZDZ).
- Fog/Mist—Fog (FG), mist (BR), and freezing fog (FZFG).

- Haze (HZ) and smoke (FU).

- Thunderstorms (TS), including funnel clouds (FC) and tornadoes/waterspout (+FC). Some observation stations do not report thunderstorms. These METARs use the TSNO remark. Thunderstorms are not verified whenever the TSNO remark appears in the observation. VCTS are not considered in verification and are ignored whenever they appear in forecasts or observations. Note: Vicinity Thunderstorms (VCTS) means thunderstorms are forecast or were observed within a 5- to 10-mile radius from the center of a terminal.

- Hail (1/4 inch or greater diameter) (GR) and showers of GR (SHGR).

- Squalls (SQ).

- Blowing snow (BLSN), drifting snow (DRSN).

- Blowing spray (BLPY).

- Volcanic ash (VA).

- All dust and sand events, i.e., widespread dust (DU), blowing dust (BLDU), drifting dust (DRDU), dust storm (DS), sand/dust whirls (PO), blowing sand (BLSA), drifting sand (DRSA), and sandstorm (SS).

6.1.7 Forecast Types

TAFs primarily predict prevailing conditions and use the "from" (FM) change indicator to introduce changes to the forecast prevailing conditions. Prevailing forecast verification is described in section 6.1.7.1. Another "type" of forecast, called the operational impact forecast, is defined in section 6.1.7.2. Sometimes a TEMPO or PROB change indicator is used to respectively designate a temporarily fluctuating or probabilistic forecast condition. When a TEMPO or PROB change indicator is used, two forecasts are valid for the same time. TEMPO and PROB forecast evaluation are explained, respectively, in sections 6.1.7.3 and 6.1.7.4. The following terms are repeated several times and are defined:

a. A *change* from one observation to a subsequent observation:

(1) For ceiling and visibility, change is defined as at least a one-category change. Respectively, the six fixed categories for ceiling and visibility are located in sections 6.1.6.1 and 6.1.6.2.

(2) Each of the thirteen significant weather types is a binary variable, and change is defined as the starting or stopping of that weather type. Precipitation intensities are ignored.

b. A *hit* for a TEMPO or PROB forecast is defined:

(1) For ceiling and visibility, a forecast hit is defined as the forecast category equaling the observation category. Respectively, the fixed categories are defined in sections 6.1.6.1 and 6.1.6.2.

(2) For each of the thirteen significant weather types, a hit occurs when the forecast and observation agree on the occurrence of that weather type.

c. Less [More] in Error. When comparing two forecast types (i.e., prevailing and TEMPO, prevailing and PROB):

(1) For ceiling or visibility, less [more] in error means the TEMPO or PROB forecast was not a hit and had a smaller [larger] absolute categorical error than the prevailing forecast. Respectively, the fixed categories used to define categorical error for ceiling and visibility are located in sections 6.1.6.1 and 6.1.6.2.

(2) All thirteen significant weather types are binary variables, so the term "less [more] in error" is not used when referring to any of them.

d. More [Less] Favorable Flight Conditions. When comparing two forecast types (i.e., prevailing and TEMPO, prevailing and PROB):

(1) For ceiling or visibility, the more [less] favorable flight conditions are defined as the higher [lower] category forecast, using the fixed categories for ceiling and visibility, respectively located in sections 6.1.6.1 and 6.1.6.2.

(2) For each of the thirteen significant weather type forecasts, the more [less] favorable flight condition is defined as the negative [positive] forecast of that event.

6.1.7.1 Prevailing Forecast

The prevailing forecast is defined as (1) the forecast conditions that are in the initial time period of the TAF and (2) any forecast conditions that immediately follow a FM change indicator. For the element specified by the user of *Stats on Demand* (e.g., ceiling), the prevailing forecast is evaluated at the end of every 5-minute interval of the TAF by comparing it to the most recent METAR/SPECI available. Most verification is categorical, using the fixed categories defined in the verification reports, and results are recorded twelve times per hour in contingency tables of forecasts versus observations. Prevailing forecasts are evaluated by themselves, or they are matched with one guidance product at a time, producing an additional contingency table of guidance forecasts versus observations. Conventional verification statistics are computed from the contingency tables, and comparisons may be drawn between prevailing forecast and guidance performance.

6.1.7.2 Operational Impact Forecast (OIF)

TAFs are sometimes formatted in a manner whereby two forecasts are valid for a single terminal at the same time. One of the following circumstances applies to all NWS TAFs at all times: (1) Just the prevailing forecast is in effect, (2) The prevailing forecast is in effect simultaneously with a forecast for temporary conditions (TEMPO), or (3) The prevailing forecast is in effect simultaneously with a 30% probabilistic forecast (PROB). For verification, the OIF is defined as the forecast in effect that is most likely to have the largest impact on operations. The following rules are used to determine the OIF:

a. The OIF is not provided for any of the wind elements.

b. If no TEMPO or PROB forecast is in effect for the user-specified element, then the OIF for that element is defined as the prevailing forecast.

c. If a PROB forecast is in effect for the user-specified element, then the OIF for that element is defined as the forecast (prevailing or PROB) of the less favorable flight conditions, i.e., lower ceiling category, lower visibility category, or the occurrence of the weather type.

d. If a TEMPO forecast is in effect for the user-specified element, then the OIF for that element is defined through a two step process.

(1) *First step—the variability test.* The legitimacy of the TEMPO forecast is first evaluated by a variability test that is performed at the end of every 5-minute interval of the TAF TEMPO forecast. If the observation database changes twice or more during the variability period, then that 5-minute interval of the TEMPO forecast passes the variability test. The term "change" for each forecast element in the TAF is defined in section 6.1.7. The beginning time of the variability period for each 5-minute window in the TEMPO forecast is defined as 90 minutes prior to the ending time of the 5-minute interval being tested. The end time of the variability period for each 5-minute window in the TEMPO forecast is defined as 90 minutes after the ending time of the 5-minute interval being tested. Note: This test only measures condition variability—it does not measure forecast correctness or accuracy.

(2) *Second step.* If the TEMPO forecast fails the variability test for a given 5-minute interval, then the OIF for that interval is defined as the forecast with the less favorable flight conditions, i.e., lower ceiling category, lower visibility category, or the occurrence of the significant weather type.

If the TEMPO forecast passes the variability test for a given 5-minute interval, then the OIF for that interval is (1) defined as the forecast with the smallest categorical error for ceiling and visibility (in a tie, the OIF is

set equal to the prevailing forecast category); or (2) set equal to the observation (no error) for each significant weather type.

The OIF for flight category is determined by first calculating the OIF separately for ceiling and visibility. Then, Table A-8 (located in section 6.1.6.3) is used to find the OIF for flight category. The lower category of the two is the flight category OIF.

Just like the prevailing forecast, the OIF is evaluated only for the element specified by the user of *Stats on Demand* at the end of every 5-minute interval that the TAF is valid. At each of these times, the OIF is compared to the most recent METAR/SPECI available. Most verification is categorical, using the fixed categories defined in the verification reports, and results are recorded twelve times per hour in contingency tables of forecasts versus observations. OIFs are evaluated by themselves, or they are matched with one guidance product at a time, producing an additional contingency table of guidance forecasts versus observations. Conventional verification statistics are computed from the contingency tables, and comparisons may be drawn between OIF performance and guidance performance.

6.1.7.3 TEMPO Forecast

TEMPO forecast statistics provide feedback concerning the effectiveness of TEMPO usage. TEMPO forecasts are evaluated at the end of every 5-minute interval for which a TEMPO forecast was valid, even though TEMPO forecasts are prepared with hourly temporal resolution. These statistics are not matched to guidance. They are provided for ceiling, visibility, or the thirteen significant weather types. The following statistics are tallied:

a. Number of hours each significant weather type was observed. This is the number of hours each significant weather type was observed regardless of whether or not it was included in a prevailing, TEMPO or PROB30 forecast. By definition, this statistic is only provided for significant weather types and not the other elements.

b. Number of hours TEMPO forecast. This is the total number of hours (the number of 5-minute intervals divided by 12) TEMPO forecasts were issued. This value is provided to the nearest hour.

c. Justified TEMPO. This is the total number of hours when the TEMPO forecast passed the variability test. A TEMPO forecast that passes the variability test is *justified.* The variability test is performed at the end of every 5-minute interval of the valid period of the TEMPO forecast. For a given 5-minute interval, if the observation changed twice or more during the variability period ("variability period" is defined in the next paragraph), then the TEMPO forecast passed the variability test for that 5-minute interval. The term *change* is defined separately for each individual element in section 6.1.7 (paragraph a).

   The variability period for each 5-minute interval in a TEMPO forecast is defined as follows: it begins 90 minutes prior to the end of the 5-minute interval being tested and ends 90 minutes after the conclusion of that same 5-minute interval.

This makes the variability period a sliding three-hour window that precedes and follows the TEMPO valid period. This means that whenever a TEMPO forecast is valid at the start time of a given TAF, the variability period for its first 5-minute interval precedes the start time of the TAF by 85 minutes. This statistic only measures condition variability—it does not measure forecast correctness or accuracy.

Example: A TEMPO group is in effect from 0800 until 1200 UTC. The end of every 5-minute interval is checked for justification. Start with the end time of 0800-0805 UTC and see if two or more changes occur between 0635 and 0935 UTC (0805 UTC ±90 minutes). If a 1500-foot ceiling at 0635 UTC rises to 2500 feet at 0720 UTC, and then drops to 1200 feet at 0840 UTC, then two changes occurred between 0635 and 0935 UTC, making the TEMPO group justified for the 0800-0805 UTC interval. Repeat this process for every five minute interval until you finish the TEMPO group at noon (last 5-minute interval is 1155-1200 UTC). Assuming no more ceiling category changes occurred after 0840 UTC, the number of 5-minute intervals with a justified TEMPO forecast was 10, which converts to 0.8 (10/12) hour. Ideally, this number should equal the total number of TEMPO forecast hours, which in this example was 4.0.

d. Justified TEMPO–Hit (%). Considering only the 5-minute intervals when the TEMPO forecast was justified for the user-specified element (see paragraph c. of this section for the definition of *justified*), this is the percentage of time that the TEMPO forecast was a hit. The term *hit* is defined separately for each individual element in section 6.1.7 (paragraph b). Ideally, this statistic ranges between 10 and 49. *Example: Between 0600 and 0820 UTC, the observations indicated that ceilings varied sufficiently to justify a TEMPO group. The TAF prevailing group forecast ceilings at 800 feet, the TEMPO group forecast ceilings at 300 feet, and ceilings 200 to 400 feet, inclusive, were observed at the end of 40% of the 5-minute intervals between 0600 and 0820UTC. Justified TEMPO - Hit (%): 40.*

e. Justified TEMPO–Improved the TAF (%). Considering only the 5-minute intervals when the TEMPO forecast was justified (see paragraph c. of this section for the definition of *justified*), this is the percentage of time the TEMPO forecast was less in error than the prevailing forecast, and the TEMPO forecast was not a hit. Since each of the thirteen significant weather types is a binary variable and can, therefore, only be a hit or miss, this statistic is not provided for the significant weather types. *Example: Between 0600 and 0820 UTC, the observations indicated that ceilings varied enough to justify a TEMPO group. The TAF prevailing group forecast ceilings at 1200 feet, the TEMPO group forecast ceilings at 700 feet, and ceilings between 200 and 400 feet were observed at the end of 40% of the 5-minute intervals between 0600 and 0820 UTC. Justified TEMPO–Improved TAF (%): 40.*

f. Unjustified TEMPO (hours). This is the total number of hours that the TEMPO forecast was unjustified, and therefore, failed the variability test (see paragraph c.

of this section for a description of the variability test). This statistic is determined by subtracting the number of hours of justified TEMPO forecasts from the total number of TEMPO forecast hours. Ideally, this statistic is zero.

g. Unjustified TEMPO - Should Be FM (%). Considering only the 5-minute intervals when the TEMPO forecast was not justified (see paragraph c. of this section for the definition of justified), this statistic is the percentage of time when the TEMPO forecast was a hit, resulting in an incorrect prevailing forecast. Ideally, this statistic is zero. *Example: During the period that the observations indicated that ceilings did not vary enough to justify a TEMPO group, the TAF prevailing group forecast ceilings at 1200 feet, the TEMPO group forecast ceilings at 800 feet, and ceilings were observed between 500 and 900 feet all the time. TEMPO S/B FM (%): 100.*

h. Unjustified TEMPO - Benign (%). Considering only the 5-minute intervals when the TEMPO forecast was not justified (see paragraph c. of this section for the definition of justified), this statistic is the percentage of time whenever (1) the TEMPO forecast was more in error than the prevailing forecast, and (2) the TEMPO forecast predicted *more* favorable flight conditions than the prevailing forecast. In these cases, poor TEMPO forecasts are benign to flight operations because the pilot has already planned for the less favorable flight conditions in the prevailing forecast. Ideally, this statistic is zero. *Example: The TAF prevailing group forecast ceilings at 700 feet, the TEMPO group forecast ceilings at 1200 feet, and ceilings were observed between 500 and 900 feet at the end of 90% of the 5-minute intervals that failed the variability test. Tempo Benign (%): 90.*

i. Unjustified TEMPO - Hurt (%). Considering only the 5-minute intervals when the TEMPO forecast was not justified (see paragraph c. of this section for the definition of "justified"), this statistic is the percentage of time whenever (1) the TEMPO forecast was more in error than the prevailing forecast, and (2) the TEMPO forecast predicted *less* favorable flight conditions than the prevailing forecast. In these cases, poor TEMPO forecasts hurt flight operations because the pilot is forced to plan for the less favorable flight conditions that did not occur. Ideally, this statistic is zero. *Example: The TAF prevailing group forecast ceilings at 1400 feet, the TEMPO group forecast ceilings at 600 feet, and ceilings were observed between 1000 and 1900 feet at the end of 90% of the 5-minute intervals that failed the variability test. TEMPO Hurt (%): 90.*

The aforementioned statistics are summarized in Table 13.

Table 13. Summary of TAF TEMPO Group Verification Statistics.

Statistic	Definition	Ideal value or range
Number of hours	Number of hours each significant weather type was observed regardless of whether or not it was included in the prevailing, TEMPO or PROB forecast.	N/A
Number of hours TEMPO forecast	Total number of hours (number of 5 minute intervals divided by 12) that TEMPO forecasts were issued (specified to the nearest hour).	N/A
Justified TEMPO (hours)	Total number of hours when the TEMPO forecast passed the variability test.	Total number of TEMPO forecast hours
Justified TEMPO – Hit (%)	Considering only 5 minute intervals when the TEMPO forecast was justified, this is the percentage of time the TEMPO forecast was a hit.	10% to 49%
Justified TEMPO – Improved the TAF (%)	Considering only 5 minute intervals when the TEMPO forecast was justified , this is the percentage of time the TEMPO forecast was less in error than the prevailing forecast, without being a hit.	N/A
Unjustified TEMPO (hours)	Total number of hours that the TEMPO forecast failed the variability test.	Zero
Unjustified TEMPO – Should be FM (%)	Considering only 5 minute intervals when the TEMPO forecast was ***not*** justified, this is the percentage of time the TEMPO forecast was a hit (this implies the prevailing forecast was a miss).	Zero
Unjustified TEMPO – Benign (%)	Considering only 5 minute intervals when the TEMPO forecast was ***not*** justified, this is the percentage of time the TEMPO forecast was more in error than the prevailing forecast, and the TEMPO forecast predicted ***more*** favorable conditions than the prevailing forecast.	Zero
Unjustified TEMPO – Hurt (%)	Considering only 5 minute intervals when the TEMPO forecast was ***not*** justified, this is the percentage of time the TEMPO forecast was more in error than the prevailing forecast, and the TEMPO forecast predicted ***less*** favorable conditions than the prevailing forecast.	Zero

Most of these statistics are used for every element, including the elements collectively called "significant weather type." However, the following exceptions exist:

Each significant weather type (e.g., rain, fog, snow) is a binary variable, i.e., it occurs or it doesn't occur (yes or no). Therefore, statistic (e), "improved the TAF" has no meaning for any of the significant weather types because a binary forecast situation that is a "miss" cannot simultaneously improve the forecast, i.e., the hits have already been counted by statistic (d).

Statistic (a) is only used for the significant weather types because it lists the total number of hours that each significant weather type occurred and has no meaning for elements such as ceiling, visibility, flight category, wind speed and wind gusts. All of the significant weather types collectively fit into the format of Table 14 in the *Stats on Demand* reports.

Table 14. Example of TEMPO table for the significant weather types.

	Condition Observed (Hours)	TEMPO Forecast (Hours)	Justified TEMPO		Unjustified TEMPO			
			Number of Hours	Hits % of (d)	Number of Hours	TEMPO S/B FM %of (f)	TEMPO Benign % of (f)	TEMPO Hurt % of (f)
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
LIQUID	1296	1305	533.7	42%	771.3	37%	1%	55%
etc.								

6.1.7.4 PROB Forecast

The PROB forecast is evaluated at the end of every 5-minute interval for which a PROB forecast was valid. PROB forecast evaluation statistics are not matched with guidance. Statistics are provided for any element selected by the user; however, wind direction and flight category are not available. PROB forecasts during the first nine hours of the TAF are not allowed in the NWS so these statistics should be viewed beyond the 9-hour projection. The following statistics are tallied:

a. Number of Hours: This is the total number of hours (the number of 5-minute intervals divided by twelve) that PROB groups were valid for the user-specified element. This value is provided to the nearest hour.

b. PROB Hit (Element + precip/TS) (%): This is the percentage of all 5-minute intervals when (1) the PROB forecast was a hit and (2) precipitation or a thunderstorm (i.e., TS, FC, or +FC) occurred simultaneously. Credit is not granted if the element is a hit, but precipitation or a thunderstorm did not occur. If the user selects *significant weather type* as the element, the rows representing the various precipitation types and thunderstorms (i.e., TS, FC, or +FC) are "blacked out." Ideally, this statistic is between 30 and 40. *Example: The prevailing forecast is 4000 feet, the PROB forecast is 1500 feet, and light rain is forecast with the lower ceilings. Ceilings between 1000 and 1900 feet with light snow were observed at the end of 30% of the 5-minute intervals. Prob Hit w/*

precip/TS: 30. Note: The 30% hit rate occurred even though rain was forecast with the lower ceilings and snow was observed. For this statistic, any type of precipitation or a thunderstorm verifies the ceiling. The significant weather type (incorrect rain forecast) is verified separately in the significant WX type rows. If no precipitation had occurred with the lower ceilings, the Prob Hit w/ precip/TS would have been zero.

c. PROB Hit w/out precip/TS (%): This is the percentage of all 5-minute intervals when the PROB forecast was a hit; however, precipitation or a thunderstorm (i.e., TS, FC, or +FC) did not occur simultaneously. If the user selects *significant weather type* as the element, the rows representing the various precipitation types and thunderstorms (i.e., TS, FC, or +FC) are "blacked out." *Example: The prevailing ceiling forecast is 4000 feet, the PROB forecast is 1500 feet, and light rain is forecast with the lower ceilings. Ceilings between 1000 and 1900 feet were observed at the end of 30% of the 5-minute intervals, but no precipitation or thunderstorm events occurred at the end of these 5-minute intervals. Prob Hit w/out precip/TS: 30.*

d. PROB Hit (Precip/TS only) (%): This is the percentage of all 5-minute intervals when a PROB forecast for precipitation or thunderstorms (i.e., TS, FC, or +FC) was a hit. This statistic is only provided when the element requested is *significant weather type*. The only rows with data are the various precipitation types and thunderstorms (i.e., TS, FC, or +FC); all others are "blacked out." Ideally, this statistic is between 30 and 40.

e. PROB Improved the TAF (%): This is the percentage of all 5-minute intervals when the PROB forecast was not a hit, but the PROB forecast was less in error than the prevailing forecast. Unlike the *PROB Hit* statistic (bullet b.), credit is granted whenever precipitation or a thunderstorm did not occur. This statistic is provided for any element selected by the user, except for the thirteen significant weather types. Ideally, this statistic is zero. *Example: The TAF prevailing group forecast ceilings at 1200 feet, the PROB group forecast ceilings at 700 feet, ceilings below 200 to 400 feet were observed 40% of the time, and ceilings 1000 feet or higher were observed 60% of the time. Prob Imp (%): 40.*

f. PROB Benign (%). This is the percentage of all 5-minute intervals when (1) the PROB forecast was more in error than the prevailing forecast, and (2) the PROB forecast predicted *more* favorable flight conditions than the prevailing forecast. In these situations, poor PROB forecasts are benign to flight operations because the pilot has already planned for the less favorable flight conditions in the prevailing forecast. Ideally, this statistic is zero. *Example: The TAF prevailing group forecast ceilings at 700 feet, the PROB group forecast ceilings at 1200 feet, and ceilings were observed between 500 and 900 feet at the end of 90% of the 5-minute intervals. Prob Benign (%): 90.*

g. PROB Hurt (%). This is the percentage of all 5-minute intervals when (1) the PROB forecast was more in error than the prevailing forecast, and (2) the PROB forecast predicted *less* favorable flight conditions than the prevailing forecast. In these cases, the poor PROB forecasts hurt flight operations because the pilot was forced to plan for the less favorable flight conditions that did not occur. Ideally, this statistic is zero. *Example: The TAF prevailing group forecast ceilings at 1400 feet, the PROB group forecast ceilings at 600 feet, and ceilings were observed between 1000 and 1900 feet 90% of the time. Prob Hurt (%): 90.*

The aforementioned statistics are summarized in Table 15.

Table 15. Summary of TAF PROB Group Verification Statistics.

Statistic	Definition	Ideal value or range
Number of Hours	Total number of hours (number of 5 minute intervals divided by 12) that PROB groups were valid for the user-specified element (to nearest hour).	N/A
PROB Hit (Element +Pcpn/Tstm) (%)	Percentage of all 5 minute intervals within PROB groups when the PROB forecast was a hit, and the Pcpn/Tstm occurred.	30% to 40%
PROB Hit without Pcpn/Tstm (%)	Percentage of all 5 minute intervals within PROB groups when the forecast was a hit, and the Pcpn/Tstm did not occur.	Zero
PROB Hit (Pcpn/Tstm only) (%)	Percentage of all 5 minute intervals within PROB groups when the precipitation or thunderstorm forecast was a hit.	30% to 40%
PROB Improved the TAF (%)	Percentage of all 5 minute intervals within PROB groups when the PROB forecast was not a hit, but it was less in error than the prevailing forecast.	Zero
PROB Benign (%)	Percentage of all 5 minute intervals within PROB groups when the PROB forecast was more in error than the prevailing forecast, and the PROB group forecasted ***more*** favorable conditions than the prevailing forecast.	Zero
PROB Hurt (%)	Percentage of all 5 minute intervals within PROB groups for user specified element, where the PROB forecast was more in error than the prevailing forecast, and the PROB group forecasted ***less*** favorable flight conditions than the prevailing forecast.	Zero

6.2 Aviation Weather Center (AWC) Forecasts

6.2.1 Background

The AWC uses the automated Real-Time Verification System (RTVS), created specifically for verifying AWC's manually produced forecasts and various associated automated forecast algorithms. RTVS is continuously under review and revision as more and better sources of aviation verification observations are implemented. Verification techniques are under constant scrutiny in an effort to improve upon the subjectivity of pilot reports and other observations/observation products used in many aviation forecast verification procedures. Additionally, the RTVS' convective verification procedures are often revised and refined in an effort to provide the AWC with the best possible statistics for describing the accuracy of its convective forecasts. The National Convective Weather Diagnostic algorithm is currently used to verify AWC's convective products. While RTVS provides a baseline and a starting point for verification trend monitoring, the statistics are subject to change as RTVS evolves into a more mature system meeting the AWC's needs. Statistics are also prone to substantial monthly and seasonal variability based on the subjectivity and unreliable frequency of pilot reports. No standardized observing network exists for verifying aviation forecast variables, such as icing and turbulence. Despite these problems, statistics are presented as 12-month running averages.

6.2.2 Domestic Products Verified and Statistics Calculated

a. Airman's Meteorological Information (AIRMET)

(1) Icing (AIRMET Zulu) and Turbulence (AIRMET Tango). The following verification statistics, defined in section 4.4 of Appendix A, are calculated separately for AIRMET Zulu and AIRMET Tango: *POD, POD* of no observations (*POD[N]*), the percent area of AIRMET coverage across the domestic airspace (% Area), and the percent volume of AIRMET coverage across the domestic airspace.

(2) Instrument Flight Rules (IFR) Conditions (AIRMET Sierra). The following verification statistics are calculated: *POD, FAR,* and % Area.

b. Convective Forecasts

(1) Convective Significant Meteorological Information (SIGMET). The following verification statistics are calculated: *POD, FAR,* % Area.

(2) Collaborative Convective Forecast Product: The following verification statistics are calculated: *POD, FAR,* and % Area.

7 Tropical Cyclones

The National Hurricane Center (NHC) and the Central Pacific Hurricane Center (CPHC) verify tropical cyclone track and intensity forecasts.

7.1 Tropical Cyclone Forecasts/Advisories

NHC and CPHC issue Tropical Cyclone Forecast/Advisory products. The Tropical Cyclone Forecast/Advisory product will be referred to as the TCM product in this instruction. The first TCM product associated with each tropical system is normally issued when meteorological data indicate the formation of a tropical or subtropical cyclone. Subsequent advisories are issued at 0300, 0900, 1500, and 2100 UTC. Special forecasts/advisories are issued if significant changes to the forecast occur. Each advisory product contains 12-, 24-, 36-, 48-, 72-, 96-, and 120-hour forecast positions and maximum sustained wind speed. Forecast positions are rounded to the nearest tenth of a degree of latitude and longitude, and forecast intensities are rounded to the nearest 5 knots.

7.1.1 Verification Elements

The following TCM elements are verified at 12, 24, 36, 48, 72, 96, and 120 hours:

a. Maximum Sustained Surface Wind. A tropical cyclone's intensity is verified by the maximum sustained surface wind, defined as the highest 1-minute average wind (at an elevation of 10 m with an unobstructed exposure) associated with the cyclone at a particular point in time. Units for this element are "knots."

b. Location. The position of the tropical cyclone center is determined from several parameters at multiple layers, including vorticity maxima and the cyclone's the cyclone's minimum wind or minimum surface pressure. The units for this element are degrees latitude and longitude.

7.1.2 Verification Process

Each TCM product contains an operational estimate of the tropical cyclone's current location and maximum sustained surface wind speed. These estimates are determined from a variety of sources, including surface observations from land or marine platforms, aircraft reconnaissance data, radars, and satellites. During a tropical cyclone event as new observations become available, an ongoing evaluation of the operational location and intensity estimates results in the creation of a "working best track", whose points will often differ from the operational values contained in the TCM. A preliminary verification of the TCM forecast parameters can be accomplished by comparison with the working best track.

After each tropical cyclone event has concluded, hurricane specialists review all available data and refine the working best track. The refined set of locations and intensities is known as the "final best track." A cyclone's final verification is performed by comparing the TCM location and intensity forecasts with the final best track. In order to be included in the verification sample, the system has been a tropical (or subtropical) cyclone at both the initial time and the forecast time.

Preparation of a cyclone's final best track is a time-consuming process that may not be completed until several weeks after the conclusion of the event. As a result, final verifications for the season are generally not available at the conclusion of the hurricane season.

7.2 Models

Various models are run operationally and provide forecasted tropical cyclone tracks. Several models provide forecasted tropical cyclone intensities. The models range in complexity from simple statistical models to three-dimensional primitive equation models.

7.2.1 Verification Elements

The following model elements may be verified at 12, 24, 36, 48, 72, 96, and 120 hours:

a. Maximum Sustained Surface Wind. A tropical cyclone's intensity is verified by the maximum sustained surface wind, defined as the highest 1-minute average wind (at an elevation of 10 m with an unobstructed exposure) associated with the cyclone at a particular point in time. Units for this element are "knots."

b. Location. The position of the tropical cyclone center is determined from several parameters at multiple layers, including vorticity maxima and the cyclone's the cyclone's minimum wind or minimum surface pressure. The units for this element are degrees latitude and longitude.

7.2.2 Verification Process

A preliminary verification of model location and intensity forecasts may be made against the working best track. The final verification will be made using the final best track.

7.3 Verification Reports

The NHC and the CPHC maintain verification statistics and post them on their respective websites:

http://www.nhc.noaa.gov/verification
http://www.prh.noaa.gov/cphc

8 Climate Outlooks

8.1 Medium Range and Seasonal Outlooks

The Climate Prediction Center (CPC) verifies its medium range and seasonal outlooks.

a. The following mean temperature and total precipitation forecasts are verified on a grid that covers the contiguous United States:

(1) 6 to 10 day forecast.

(2) Week 2 (8-14 day) forecast.

(3) Monthly, issued monthly for the following month with a 0.5-month lead.

(4) Seasonal, issued monthly for twelve consecutive 3-month seasons. Each of the twelve seasonal forecasts is issued with a 0.5-month through 12.5-month lead time.

b. The data specifications follow:

(1) Data Source: River Forecast Centers – Approximately 5000 stations per day are used, including approximately 1500 stations per day from the Hydrologic Automated Data System (HADS) and several hundred stations per day from the Climate Anomaly Data Base.

(2) Resolution: The station data are fit to a 0.5x0.5-degree grid, and the verification is done on a 2x2-degree grid.

(3) Domain: 20 to 60 degrees North; 60 to 140 degrees West.

(4) Format: The format is sequential 32-bit IEEE floating point created on a big endian platform (e.g. cray, sun, sgi and hp). The undefined (missing) value is 9999.

(5) Window: The Day 1 analysis is valid for the window from 1200 UTC on Day 0 (the day issued) to 1200 UTC on Day 1. Because of report receipt timing, daily minima are available 1 day earlier than the daily maxima and the daily means.

(6) Analysis Scheme: Modified Cressman (1959) scheme (Glahn et al. 1985; Charba et al. 1992). The minimum number of stations required for analysis is 350. Whenever the number of stations is fewer than 350, the analysis is not performed for that particular day.

(7) Quality Control: A climatological standard deviation check is used. If a reported value is more than 4 standard deviations removed from the historical distribution, the value is omitted from the analysis.

A version of the Heidke Skill score (described in section 2.7 of Appendix A) is computed for verification.

8.2 U.S. Hazards Assessment Product

CPC verifies heavy precipitation forecasts in its 3- to 14-day U.S. Hazards Assessment Product. Hazard forecasts of daily (1200 to 1200 UTC) precipitation expected to exceed the hazard threshold at specific grid points on specific dates are made each weekday for the 3- to 14-day

forecast period, e.g., a forecast made on Tuesday is valid 1200 UTC Friday (Day 3) until 1200 UTC on the Tuesday two weeks after the forecast is issued (Day 14). All issuances of the Hazard Assessment are verified. The forecast domain consists of a one-degree-latitude by one-degree-longitude grid (881 points) over the contiguous United States. The daily hazard threshold for each grid point is defined as the greater of one inch of precipitation for a given day or the 95th percentile of the climatology for a given day. For verification, daily (1200 to 1200 UTC) precipitation amounts are analyzed to each of the 881 grid points. One "event" is defined as any grid point where observed precipitation equals or exceeds the daily threshold. A similar procedure is used for verifying severe weather hazards (tornadoes, damaging winds, and large hail) included in the hazard assessment product. Observation data are taken from the SPC preliminary severe weather reports. The following 2x2 contingency table is used to classify all events and non-events with respect to how they were forecast:

Table 16. Special 2x2 contingency table.

		Forecasts	
		Yes	No
Events	Yes	A	B
	No	C	*X*

Any event that occurs on one or more days within the hazardous forecast area during the hazard period is counted as one "hit" (*A* in Table 16). For example, a heavy precipitation hazard was forecast for a particular grid point from November 17 thru 19. That grid point received enough precipitation to exceed its daily threshold on two separate dates: November 17 and 19. Consequently, one "hit" is counted. One "hit" is also counted whenever *no hazard* is forecast, and the observed precipitation does *not* equal or exceed the hazard threshold during any of the eleven forecast days *(X)*. A "miss" is counted whenever an event occurs with none forecast (*B*), or a hazard is forecast with no event reported (*C*; also known as a false alarm). From these counts, the following scores are computed (see Appendix A, sections 3.1 through 3.3): *POD*, *FAR*, and threat score (also called the *CSI)*. The bias of the hazardous events (*A+B*) is computed by dividing all hazards forecasted *(A+C)* by *(A+B)*.

9 Models

The Environmental Modeling Center verifies its numerical models. As part of its World Meteorological Organization responsibilities, the National Centers for Environmental Prediction Central Operations (NCO) sends monthly numerical model verification statistics to all World Forecast Centers. NCO also provides model verification statistics to the annual Numerical Weather Prediction report.

10 References

Charba J. P., A. W. Harrell III, and A. C. Lackner III, 1992: A monthly precipitation amount climatology derived from published atlas maps: Development of a digital database. *NOAA TDL Office Note 92-7*, 20 pp.

Cressman G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87**, 367–374.

Glahn H. R., T. L. Chambers, W. S. Richardson, and H. P. Perrotti, 1985: Objective map analysis for the local AFOS MOS Program. *NOAA Tech. Memo. NWS TDL* ***75,*** 34 pp.

Kluepfel, C.K., A.J. Schreiner, and D.A. Unger, 1994: The satellite-derived cloud cover product (sounder). *NWS Technical Procedures Bulletin No. 410*, NOAA, U.S. Department of Commerce, 15 pp.

Liu, W.T., K.B. Katsaros, and J.A. Businger, 1979: Bulk parameterization of air-sea exchanges of heat and water vapor including the molecular constraints at the interface. *J. Atmos. Sci.*, **36,** 1722-1735.

Weiss, S.J., D.L. Kelly, and J.T. Schaefer, 1980: New objective verification techniques at the National Severe Storms Forecast Center. *Preprints, 8th Conference on Weather Forecasting and Analysis*, Denver, Colorado, American Meteorological Society, 412-419.

APPENDIX A – VERIFICATION SCORES

1 Introduction A-2

2 Generalized Contingency Table A-2
2.1 Percent Hits A-2
2.2 Bias by Category A-3
2.3 Probability of Detection *(POD)* A-3
2.4 False Alarm Ratio *(FAR)* A-4
2.5 Critical Success Index *(CSI)* A-4
2.6 Generalized Skill Score *(SS)* A-5
2.7 Heidke Skill Score *(HSS)* A-5
2.8 Pierce Skill Score *(PSS)* A-6
2.9 Equitable Skill Scores A-7
2.9.1 Subjective Explanation A-7
2.9.2 Mathematical Background A-8
2.9.3 Experimental Score for Wind Direction A-12

3 Specialized Contingency Table A-14
3.1 Probability of Detection A-14
3.2 False Alarm Ratio A-16
3.3 Critical Success Index A-16
3.4 Frequency of Hits A-17

4 Scores Computed for Specific Forecast Elements A-18
4.1 Temperature, Wind Speed and Direction, and Wave Height A-18
4.2 Probability of Precipitation A-19
4.3. QPF A-21
4.4 Ceiling Height and Visibility A-22
4.5 Aviation Weather Center (AWC) Verification Statistics A-23

5 Ruth-Glahn Convergence Score A-23

6 References A-24

1 Introduction

Verification scores are applied at the local, regional, and national levels. Different scores may be applied to the same data. The type of score selected for use depends upon the objective. Frequently used scores are given in this manual and presented within the context of specific elements and events subject to verification. Two excellent references for verification scores are Jolliffe and Stephenson (2003) and Wilks (2011).

2 Generalized Contingency Table

A generalized forecast/observation contingency table (Table A-1) is often used to summarize the forecast performance of a given element by category (the term "category" is sometimes called class). The table is divided into *k* mutually exclusive and exhaustive categories. Each cell of the table, A_{ij} , gives the number of occurrences with the observation in the *ith* category (e.g., 13 to 17 knots for sustained wind speed) and the forecast in the *jth* category (e.g., 18 to 22 knots for sustained wind speed). Categorically correct forecasts (A_{ii} for all i), where all $i = j$, are represented along the upper left to lower right diagonal of the contingency table. The row and column totals, R_i and C_i, respectively, are often called the marginal totals of the contingency table, and they are used in computing forecast bias and skill.

Table A-1. Generalized Contingency Table**.**

	Forecast Category				
Observed Category	1	2	…	k	Total
1	A_{11}	A_{12}	…	A_{1k}	R_1
2	A_{21}	A_{22}	…	A_{2k}	R_2
…	…	…	…	…	…
k	A_{k1}	A_{k2}	…	A_{kk}	R_k
Total	C_1	C_2	…	C_k	N

The following scores may be computed from the data in this contingency table:

2.1 Percent Hits

$$PH = \left(\frac{100}{N}\right) \sum_{i=1}^{k} A_{ii} \ ,$$

where (referring to Table A-1):

PH= A measure of accuracy that calculates the percent hits from data in a multi-category contingency table with *k* categories.

A_{ii} = any situation when the forecast was categorically correct, i.e., the forecast category *i* equaled the observed category *i*,

$i = 1, \ldots, k$ for a contingency table with *k* categories.

N = The total number of forecast/observation pairs.

2.2 Bias by Category

$$BIAS_i = \frac{C_i}{R_i} ,$$

where (referring to Table A-1):

$BIAS_i$ = a measure of the tendency to over-forecast (bias greater than one) or under-forecast (bias less than one) a particular category, *i*, of a multi-category contingency table, where *k* values of bias exist,

$i = 1, \ldots, k$ for a contingency table with *k* categories,

C_i = the column (forecast) total for each category *i*, and

R_i = the row (observation) total for each category *i*.

2.3 Probability of Detection *(POD)*

$$POD_i = \frac{A_{ii}}{R_i} ,$$

where (referring to Table A-1):

POD_i = the probability of detection for each individual category *i* of a multi-category contingency table. It is an accuracy measure that gives the forecaster's success in covering each event of category *i* with a correct forecast (A_{ii}). The POD_i does not penalize the forecaster for incorrect forecasts of category *i*.

$i = 1, \ldots, k$ for a contingency table with *k* categories.

A_{ii} = any situation when the forecast was categorically correct, i.e., the forecast category *i* equaled the observed category *i*.

R_i = the row (observation) total for each category *i*.

Sometimes it is useful to combine two or more categories from a contingency table into a single category and compute a *POD* for the new category. For a description of this type of specialized contingency table and the *POD* formula, see sections 3 and 3.1 of this appendix.

2.4 False Alarm Ratio *(FAR)*

$$FAR_i = \frac{C_i - A_{ii}}{C_i} \ ,$$

where (referring to Table A-1):

FAR_i = the false alarm ratio for each individual category, *i*, of a multi-category contingency table (e.g., Table A-1). It is an accuracy measure that gives the fraction of forecasts of category *i* that was incorrect. It gets its name "false alarm" from the times when category *i* is a rare or extreme event that may require a warning, watch or advisory.

$i = 1, \ . \ . \ . \ , k$ for a contingency table with *k* categories.

A_{ii} = any situation when the forecast was categorically correct, i.e., the forecast category *i* equaled the observed category *i*.

C_i = the column (forecast) total for each category *i*.

Sometimes it is useful to combine two or more categories from a contingency table into a single category and compute an *FAR* for the new category. For a description of this type of specialized contingency table and the *FAR* formula, see sections 3 and 3.2 of this appendix.

2.5 Critical Success Index *(CSI)*

$$CSI_i = \frac{A_{ii}}{R_i + C_i - A_{ii}} \ ,$$

where (referring to Table A-1):

CSI_i = the critical success index for each individual category *i* of a multi-category contingency table (e.g., Table A-1). It is an accuracy measure that gives the forecaster's success in covering each event of category *i* with a correct forecast (A_{ii}) while also penalizing for incorrect forecasts of category *i*. The *POD* doesn't penalize for incorrect forecasts.

$i = 1, \ . \ . \ . \ , k$ for a contingency table with *k* categories.

A_{ii} = any situation when the forecast was categorically correct, i.e., the forecast category *i* equaled the observed category *i*.

C_i = the column (forecast) total for each category *i*.

R_i = the row (observation) total for each category *i*.

Sometimes it is useful to combine two or more categories from a contingency table into a single category and compute a *CSI* for the new category. For a description of this type of specialized contingency table and the *CSI* formula, see sections 3 and 3.3 of this appendix.

2.6 Generalized Skill Score *(SS)*

This generalized skill score measures the fraction of possible improvement of the forecasts over some standard or test set of forecasts.

$$SS = \frac{NC - E}{N - E} \quad ,$$

where:

NC = number of correct forecasts,
E represents some standard or test set of forecasts,
N = the total number of observation/forecast pairs.

2.7 Heidke Skill Score *(HSS)*

Sometimes the standard or test forecasts *(E)* from the generalized skill score (see section 2.6 of this appendix) are the values expected by chance and are computed from the marginal totals of a contingency table . One such score is the *HSS.*

$$HSS = \frac{NC - E}{N - E} \quad ,$$

where (referring to Table A-1):

$$NC \text{ (number correct)} = \sum_{i=1}^{k} A_{ii} \ .$$

$$E = \sum_{i=1}^{k} \frac{C_i R_i}{N} \quad .$$

N = the total number of forecast/observation pairs in the contingency table.

C_i = the column (forecast) total for each category *i*.

R_i = the row (observation) total for each category *i*.

The Heidke skill score is similar to the Peirce skill score (section 2.8). Both scores are equitable, which means a perfect set of forecasts receives a score equal to one, and zero indicates no forecast skill. Negative scores are rare, but mathematically possible.

With three or more categories in the contingency table, Heidke and Peirce only provide credit to categorical forecast hits along the upper left to lower right diagonal of the contingency table. Therefore, they do not penalize large categorical errors more than small categorical errors. This property rules out the possibility for granting “partial credit” to small forecast errors or “near hits.”

Heidke and Peirce differ only in how they estimate the number of correct forecasts that would be expected by chance in their respective denominators; the numerators of the two scores are identical. Heidke penalizes the forecaster more than Peirce for over-forecasting the rare event

categories, and Heidke rewards the forecaster more than Peirce for under-forecasting the rare event categories. However, both scores are similar in how they reward *correct* forecasts of rare events. Correct forecasts of low frequency events are treated the same as correct forecasts of common events so the forecaster is not encouraged to forecast climatologically improbable events. The Gerrity skill score (sections 2.9.1 and 2.9.2 of this appendix) further addresses the matter of rare event forecasting.

The CPC uses a version of the Heidke skill score for its main verification statistic. This is calculated by the formula:

$$HSS = \left(\frac{NC - CH}{NT - CH}\right) 100 \ ,$$

where:

NC is the total number of locations for which the forecast was correct,
NT is the total number of locations for which a forecast was made,
CH is the number of locations which would be forecast correctly, on average, by chance.

In a three class system (which is how all the CPC forecasts are characterized), one third of the locations are expected to be correct by chance. Thus if 99 locations are forecast, 33 are expected to be correctly forecast. This statistic results in scores of 100 if all locations are forecast correctly, zero if 33 are forecast correctly, and -50 if all locations are forecast incorrectly.

2.8 Peirce Skill Score *(PSS)*

The Pierce skill score (Peirce 1884), also known as the Hanssen–Kuipers discriminant (Hanssen and Kuipers, 1965) and the true skill statistic (Flueck 1987), is calculated from a contingency table and is similar to the Heidke skill score. Both scores are equitable, which means a perfect set of forecasts receive a score equal to one, and zero is indicative of no forecast skill. Negative scores are rare, but mathematically possible.

$$PSS = \frac{NC - E}{N - E^*} \ ,$$

where (referring to Table A-1):

$$NC \text{ (number correct)} = \sum_{i=1}^{k} A_{ii} \ .$$

$$E = \sum_{i=1}^{k} \frac{C_i R_i}{N} \ .$$

$$E^* = \sum_{i=1}^{k} \frac{R_i R_i}{N} \ .$$

N = the total number of forecast/observation pairs.

C_i = the column (forecast) total for each category i.

R_i = the row (observation) total for each category i.

With three or more categories in the contingency table, Peirce and Heidke only provide credit to categorical forecast hits along the upper left to lower right diagonal of the contingency table. Therefore, they do not penalize large categorical errors more than small categorical errors. This property rules out the possibility for granting "partial credit" to small forecast errors or "near hits."

Peirce and Heidke differ only in how they estimate the number of correct forecasts that would be expected by chance in their respective denominators; the numerators of the two scores are identical. Heidke penalizes the forecaster more than Peirce for over-forecasting the rare event categories, and Heidke rewards the forecaster more than Peirce for under-forecasting the rare event categories. However, both scores are similar in how they reward *correct* forecasts of rare events. Correct forecasts of low frequency events are treated the same as correct forecasts of common events so the forecaster is not encouraged to forecast climatologically improbable events. The Gerrity skill score (sections 2.9.1 and 2.9.2 of this appendix) addresses the matter of rare event forecasting.

2.9 Equitable Skill Scores (ESS)

2.9.1 Subjective Explanation

Skill scores are often used to evaluate multi-category forecasts with a single score. Equitability is a desirable property for a skill score because equitability has the following characteristics:

a. A set of perfect forecasts (all categorical hits) produces a score equal to one.

b. A set of randomly generated forecasts or a set of forecasts that always predicts the same forecast category results in a "no skill" score equal to zero.

While equitable skill scores, such as Heidke (section 2.7 of this appendix) and Peirce (section 2.8 of this appendix), are convenient (they can be computed with a hand calculator), they only grant credit for categorical forecast hits. Therefore, with three or more categories in the contingency table, Peirce and Heidke do not penalize large categorical errors more than small ones, and this rules out the possibility of receiving partial credit for "near hits." Also, correct forecasts of low frequency events are treated the same as correct forecasts of very common events so the forecaster is not encouraged to forecast climatologically improbable (rare) events.

Gandin and Murphy (1992) developed a mathematical framework for computing equitable scores that allow for a system of graduated, partial credit that considers the size of each miss and the observed frequency of each category. While Gandin and Murphy allowed for forecast systems with a higher number of forecast categories, examples of systems with greater than three categories were beyond the scope of their work. Gerrity (1992) built upon Gandin and Murphy

and derived a general set of formulas that place no upper limit on the number of categories allowed in the system. The Gerrity Skill Score *(GSS)* is applied to scoring forecasts of ordinal variables (order matters) with maximum and minimum values, e.g., temperature, wind speed, ceiling, and visibility. This score is not easily calculated by hand, but it is relatively simple to program. The *GSS* has been implemented operationally in the NWS and has the following reward/penalty characteristics:

a. A relatively small reward is given for correctly forecasting common events.

b. A large reward is given for correctly forecasting rare events.

c. A graduated reward/penalty system is used, whereby a large forecast error for a given observation category is penalized more than a small forecast error in that same observation category.

The otherwise favorable property of giving large rewards to correct forecasts of rare events may make the score volatile, especially with very few occurrences of rare events. For example, if a particular event occurs on a very rare basis, the *GSS* may increase substantially due to just one additional correct forecast of that rare event. Therefore, the *GSS* is not the ideal score for data requests that include relatively small geographic areas and/or short periods of time with little variability in the element. It is also important to exercise care in defining categories in the first place to keep very rare events and volatile scores from becoming a foregone conclusion.

Depending upon the element being verified, the rarest categories tend to be either the lowest or highest categories of the contingency table. With wind speed and significant wave height, the rarest events tend to be the highest categories. With ceiling and visibility, the rarest events tend to be the lowest categories. The *GSS* Low/High Category Delta is defined as the increase that occurs in the *GSS* due to one additional forecast hit in the lowest/highest category whose event count is at least one. Whenever this score is listed in *Stats on Demand* reports, an accompanying delta value is also provided. Any delta value that exceeds 0.05 is indicative of a potentially volatile score. See section 2.9.2 for the mathematical definitions of the Gerrity score, the accompanying delta values, and more specifics about how the *Stats on Demand* programs automatically respond to scores with high delta values.

2.9.2 Mathematical Background

The probability matrix, **P**, comes from the **A** matrix (Table A-1), where all

$$p_{ij} = \frac{A_{ij}}{N} \; ; \quad (i = 1, \dots, k \;\text{ and }\; j = 1, \dots, k)$$

The row totals of the **P** matrix comprise **p**, the climatological probability vector, (p_1 , p_2 , ... , p_k). The column totals of the **P** matrix comprise **q**, the forecast probability vector, (q_1 , q_2 , ... , q_k).

Gandin and Murphy (1992) describe what is meant by an “equitable skill score” for the evaluation of categorical forecasts. The general formula is

$$ESS = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} s_{ij} \ ,$$

where:

p_{ij} are the individual cells of the aforementioned **P** matrix, and

s_{ij} are the individual cells of the reward-penalty matrix, also called the scoring (**S**) matrix.

When an appropriate climatology is used to populate the **S** matrix, a random set of forecasts yields an *ESS* equal to zero, and a perfect set of forecasts (i.e., only the diagonal of the **P** matrix is populated) yields an *ESS* equal to one.

Gerrity (1992) derived the following formulas for populating the **S** matrix in a *k*-category system. These formulas are only appropriate for ordinal variables (i.e., the order of the categories matters) that are not periodic. Wind speed and ceiling height are examples of ordinal, non-periodic variables, with maximum and minimum values.

Wind direction is not appropriate for Gerrity because it is periodic variable with no maximum and minimum values. The order of the wind direction categories still matters, but a wind direction from 360 degrees does not the mean wind direction has reached its maximum value. Clockwise from 360 degrees, the count resets. In an eight-category system, variables appropriate for Gerrity may have an error up to seven categories, whereas an eight-category wind direction forecast has a maximum error equal to four. See section 2.9.3 for an experimental alternative to Gerrity that is appropriate for scoring wind direction forecasts.

Gerrity defines p_i as the relative frequency with which category *i* of an event is observed in a large sample of forecasts and then defines D_n and R_n:

$$D_n \equiv \frac{1-\sum_{i=0}^{n} p_i}{\sum_{i=1}^{n} p_i} \qquad\qquad R_n = \frac{1}{D_n} \ ,$$

Gerrity notes that D_n is the ratio of the probability that an observation falls into a category with an index greater than *n* to the probability that it falls into a category with an index less than or equal to *n*, and R_n is the reciprocal of this ratio of probabilities.

In terms of D_i and R_i, Gerrity expresses the elements of a *k*-category equitable **S** matrix in the following manner:

$$s_{nn} = \frac{1}{k-1}\left[\sum_{i=1}^{n-1} R_i + \sum_{i=n}^{k-1} D_i\right] \ ; \quad n = (1, \dots, k)$$

$$s_{mn} = \frac{1}{k-1}\left[\sum_{i=1}^{m-1} R_i + \sum_{i=m}^{n-1}(-1) + \sum_{i=n}^{k-1} D_i\right]; \quad 1 \le m < k, \qquad m < n \le k$$

$$s_{nm} = s_{mn}\,; \quad 2 \le n \le k\,; \quad 1 \le m < n$$

Burroughs (1993) applies these general equations for populating the **S** matrix to specific *k*-category marine elements.

The **S** matrix is computed directly from the sample of the *Stats on Demand* data request. This practice has one major shortcoming; requests for verification data from relatively small, restrictive samples will tend to produce volatile scores that fluctuate due to random changes in the data set. Ironically, this problem is aggravated in these situations by the otherwise favorable property of giving more weight to rare events.

Depending upon the element being verified, the least frequently observed category in each contingency table is usually the highest or lowest category in the contingency table. For elements such as wind speed, wind gusts, and wave height (ceiling, visibility, and flight category), this is usually the highest (lowest) category. Whenever the highest (lowest) category in a contingency table has zero observed cases, the *GSS* cannot be calculated due to division by zero errors, and the unpopulated category must be combined with its adjacent category. This results in a contingency table that produces **P** and **S** matrices with one less category. Assuming the highest and lowest categories of this new system are populated, the *GSS* is calculated. In the event that one of these categories is still unpopulated, additional category combining iterations may be necessary.

Similar to the situation of an unpopulated high- or low-end category is the situation where one of these categories has a very low population. The calculation of a *GSS* in such situations is mathematically possible, but such *GSS* values are usually quite volatile and difficult to analyze, depending upon the forecast performance of the low frequency event. To help identify potential score volatility, one of the following "deltas" is calculated and listed in the data report whenever the Gerrity skill score appears:

$$\delta_{low} = \frac{s_{aa}}{N}\,,$$

$$\delta_{high} = \frac{s_{bb}}{N}\,,$$

where:

> δ_{low} is defined as the increase that occurs in the *GSS* due to one additional forecast hit in *a*, the lowest category in the contingency table whose total event count is at least one. The δ_{low} value is only listed in data reports of the following elements: ceiling, visibility, and flight category.

δ_{high} is defined as the increase that occurs in the *GSS* due to one additional forecast hit in *b*, the highest category in the contingency table whose total event count is at least one. The δ_{high} value is only listed in data reports of the following elements: wind speed, wind gusts, wave height, and quantitative precipitation categories.

The algorithm in the following four bullets addresses *GSS* volatility by rewarding the forecast/guidance for accurate predictions of rare, possibly high-impact events, that cause the score to be volatile, while not penalizing the forecast/guidance for the negative side of volatility.

- Whenever the *GSS* δ_{high} value exceeds 0.05 for one of the following elements: wind speed, wind gusts, significant wave height, quantitative precipitation, the highest category in the contingency table for forecasts and observations is combined with the second highest category, making these two adjacent categories a single category. In the resulting **P** and **S** matrices, the bottom two rows of each matrix become a single row, and the two right-most columns of each matrix become a single column.

- Whenever the *GSS* δ_{low} value exceeds 0.05 for one of the following elements: ceiling, visibility, or fight category, the lowest category in the contingency table for forecasts and observations is combined with the second lowest category, making these two adjacent categories a single category. In the resulting **P** and **S** matrices, the top two rows of each matrix become a single row, and the two left-most columns of each matrix become a single column.

- The *GSS* is then recalculated, using the **P** and **S** matrices with the reduced number of rows and columns. If the value of the recalculated *GSS* exceeds the value of the *GSS* from the previous calculation, then the recalculated *GSS* and the associated δ_{high} [δ_{low}] value are *stored for possible listing* in the data report. If the value of the recalculated *GSS* is less than or equal to the value of the *GSS* from the previous calculation, then the *GSS* from the previous calculation and the associated δ_{high} [δ_{low}] value are listed in the data report, and the following bullet is skipped, thus ending this loop.

- If the recalculated δ_{high} [δ_{low}] value is 0.05 or less, or the recalculated *GSS* came from **P** and **S** matrices with only three rows and columns, then the recalculated *GSS* and its associated δ_{high} [δ_{low}] value are listed in the data report, and the next sentence is skipped, thus ending this loop. If the recalculated δ_{high} [δ_{low}] value is greater than 0.05, then all bullets in this algorithm are repeated to see if further category reduction is necessary.

If the user doesn't like the 0.05 threshold for the delta value used by the *Stats on Demand* program for combining categories, an *Excel* calculator is linked to the Performance Management web site that allows the user to re-compute the score "manually."

The user of *Stats on Demand* can also calculate the delta for any intermediate category, *i*, in the contingency table by dividing the weight given in the **S** matrix for a correct forecast in the *i*th category (s_{ii}) by the total sample size (N). High deltas occur less frequently among the

intermediate categories, and unpopulated intermediate categories do not cause division by zero errors so no automatic category combining is performed for the intermediate categories.

2.9.3 Experimental Score for Wind Direction

The continuous, periodic nature of wind direction creates a dilemma. While the order of the wind direction categories matters, as it does with most weather elements, wind direction has no maximum and minimum values. This makes Gerrity an inappropriate skill score for verifying wind direction forecasts. This section suggests an alternative score for an eight-category wind direction forecast system, where each category represents one of the eight points of the compass. In such a *k*-category system of equally sized observation categories and nominal variables, the probability of a random forecast hit in any single category is the reciprocal of the number of categories, 1/*k* or 1/8. Therefore, the following system can be used to populate the **S** matrix whenever each category has an equal number of members.

Define n as the absolute categorical error, $| i - j |$, s_{ij} as each individual cell of the **S** matrix, and r_n as the reward/penalty multiplier for a scoring matrix to an eight-category system:

$$s_{ij} = r_n ,$$

for (n = 0, 1, 2, 3, 4) and (i , j = 1, 2, ... ,7, 8). In forming each cell of **S**, Burroughs (1993) showed that whenever each of the eight observation categories has an equal number of members and all $s_{ii} = 1$, the following relationship of r_n values must exist for equitability:

$$2(r_1 + r_2 + r_3) + r_4 = -1 . \qquad (1)$$

While equation (1) can be solved in a number of ways, the following formula distributes incremental scoring points in a linear manner:

$$s_{ij} = r_n = 1 - \frac{n}{2} . \qquad (2)$$

for ($n = | i - j |$ = 0, 1, 2, 3, 4) and (i , j = 1, 2, ... , 7, 8). For category i forecast hits, n = 0 and s_{ii} = 1, and for category i maximum misses, n = 4 and s_{ij} = -1. Two-category errors are scored as zero, while one- and three-category errors are scored, respectively, as +0.5 and -0.5, creating the following **S** matrix:

$$\mathbf{S} = \begin{pmatrix} 1 & 0.5 & 0 & -0.5 & -1 & -0.5 & 0 & 0.5 \\ 0.5 & 1 & 0.5 & 0 & -0.5 & -1 & -0.5 & 0 \\ 0 & 0.5 & 1 & 0.5 & 0 & -0.5 & -1 & -0.5 \\ -0.5 & 0 & 0.5 & 1 & 0.5 & 0 & -0.5 & -1 \\ -1 & -0.5 & 0 & 0.5 & 1 & 0.5 & 0 & -0.5 \\ -0.5 & -1 & -0.5 & 0 & 0.5 & 1 & 0.5 & 0 \\ 0 & -0.5 & 1 & -0.5 & 0 & 0.5 & 1 & 0.5 \\ 0.5 & 0 & -0.5 & -1 & -0.5 & 0 & 0.5 & 1 \end{pmatrix}.$$

Since a frequency distribution with each of the eight categories having exactly the same number occurrences almost never occurs in reality, a more generalized version of equation (2) is required, where the probability of a hit in any given category, i, is $8p_i$ for all i = 1, 2, ... , 7, 8, and p_i is the proportion of observations populating category i. Note that the denominator of Equation (3) equals one whenever p_i equals to 1/8:

$$s_{ij} = \frac{1 - \frac{n}{2}}{8p_i} \quad . \tag{3}$$

Unlike other **S** matrices from the Gandin and Murphy (1992) family of scores, including the Gerrity score, the **S** matrix computed from equation (3) is *not* symmetric, i.e., all $s_{ij} \neq s_{ji}$. Note that equation (3) yields division by zero errors whenever any p_i equals zero; hence, this score is not calculated or listed in the data report under such circumstances.

<u>Score Volatility</u>: Similar to the situation when $p_i = 0$ (with division by zero errors) is the situation when p_i is very small. While the score and all elements of the **S** matrix is computed under these circumstances, some resulting s_{ij} values may have large magnitudes, adding volatility to the score. Score volatility from each of the observed wind directions in the data report can be estimated from the **S** matrix by taking each of the values along the diagonal of forecast hits (s_{ii}) and dividing each by N, the total sample size of the contingency table (data matrix):

$$\delta_i = \frac{s_{ii}}{N} \ ,$$

for each category i.

3 Specialized Contingency Table

The following contingency table (Table A-2) may be used when only two outcomes (yes or no) exist for a given event or forecast, e.g., tornadoes.

Table A-2. Specialized Contingency Table.

		Forecasts	
		Yes	No
Events	Yes	*A*	*B*
	No	*C*	*X*

where:

A is the number of correct yes forecasts for a specific event. In warning verification, it is the number of warned events or verified warnings.

B is the number of specific events observed but not forecast. In warning verification, it is the number of unwarned events.

C is the number of yes forecasts of a specific event which did not verify. In warning verification, it is the number of unverified warnings (also known as false alarms).

X is the number of times the specific event was neither forecast nor observed.

Table 1-2 may be obtained from Table 1-1 by combining multiple categories. For example with marine forecasts, sustained wind speeds are divided into seven categories. Define sustained wind speeds equaling or exceeding 28 knots (categories 6 and 7) as the "yes" outcome for a strong wind forecast or event. In this case, the "no" outcome is all sustained wind speeds less than 28 knots (categories 1 through 5 combined). The resulting contingency table is left with two categories, yes and no.

The scores most frequently computed from this table are:

3.1 Probability of Detection

$$POD = \frac{A}{A + B} ,$$

where:

POD is the probability of detection.

A is the number of correctly forecast actual events (see Table 1-2). In warning

verification, it is the number of warned events.

B is the number of incorrectly forecast actual events. In warning verification, it is the number of un-warned events.

Storm-based Warning Verification: This paragraph pertains to all severe thunderstorm, tornado, and special marine warnings issued on or after October 1, 2007. The *POD* is calculated differently from the county-based system because each event does not fit into the categories of warned and un-warned. Many events are partially warned. Therefore, the percentage of the event warned *(PEW)* is first calculated for each event. This is done differently, depending upon the type of event and is explained in the following sections of NWSI 10-1601: section 2.1.2 for severe thunderstorms and tornadoes, section 3.4.2 for special marine warnings, and section 4.1.2 for flash flood warnings. Once a *PEW* is calculated for each event, the *POD* may be calculated:

$$POD_{SB} = \frac{0.01}{N}\sum_{i=1}^{N} PEW \ ,$$

where:

POD is the probability of detection,
SB = storm-based warnings,
PEW = the total percentage of each event, *i*, warned,
N = the total number of events.

The best possible *POD* is one; the worst is zero. Additional information, with examples for severe thunderstorms and tornadoes, is found on the Severe Weather Home Page of the NWS Performance Management Website.

Old County-Based and Marine Zone-Based Warning Verification: This paragraph pertains to all severe thunderstorm, tornado, and special marine warnings issued before October 1, 2007. The *POD* is computed from the event database and is the number of warned events divided by the total number of events. The more often an event is correctly forecast, the better (higher) the score. The best possible *POD* is one while the worst possible score is zero.

Null events: If *(A+B)* is the total number of events, e.g. turbulence or icing, sometimes it is useful to compute the *POD[N]* of null events, i.e., no turbulence or no icing. This is also called the probability of null events (PON):

$$POD[N] = PON = \frac{X}{X + C} \ ,$$

where:

X = the number of correctly forecast null events,
C = the number of incorrectly forecast null events.

3.2 False Alarm Ratio

$$FAR = \frac{C}{A + C} \ ,$$

where:

FAR is the false alarm ratio.

A is the number of correct yes forecasts (see Table 1-2). In warning verification, it is the number of verified warnings.

C is the number incorrect yes forecasts. In warning verification, it is the number of unverified warnings (also known as false alarms).

For warning verification, the *FAR* is computed from the warning database and is the number of false alarms (unverified warnings) divided by the total number of warnings. This is true for storm-based warning verification and the old county-based or marine zone-based warning verification system. For storm-based verification, any event occurring within the boundaries of the polygon during its valid period is counted as a verified warning. The more often an event is forecast and does not occur, the worse the score. The best possible *FAR* is zero, the worst possible score is one.

The *POD* and *FAR* are most often used in the verification of watches and warnings. However, it is possible to apply the *POD* and *FAR* to many events and forecasts related to public and aviation elements. Two examples are the *POD* for ceilings below 1000 feet and the *FAR* for forecasts of freezing rain.

Over-forecasting an event will achieve a high *POD* but at the expense of a high *FAR*. Overall success may be expressed by the critical success index *(CSI).*

3.3 Critical Success Index

$$CSI = \frac{A}{A + B + C} \ ,$$

where:

CSI is the critical success index.

A is the number of correct yes forecasts for a specific event. In warning verification, it is the number of warned events or verified warnings.

B is the number of specific events observed but not forecast. In warning verification, it is the number of unwarned events.

C is the number of yes forecasts of a specific event which did not verify. In warning verification, it is the number of unverified warnings (also known as false alarms).

The best possible *CSI* is one; the worst is zero. The relationship among *POD*, *FAR*, and *CSI* can also be expressed as follows. This is also the formula used to compute the *CSI* for the current storm-based warning system:

$$CSI = [(POD)^{-1} + (1 - FAR)^{-1} - 1]^{-1} \ .$$

<u>Storm-based Warning Verification</u>: The following variation of the above formula is used to compute the CSI for the current storm-based warning system:

$$CSI_{SB} = [(POD_{SB})^{-1} + (1 - FAR)^{-1} - 1]^{-1} \ .$$

<u>Old County-Based and Marine Zone-Based Warning Verification</u>: The value of *A* varies depending upon whether it is taken from the warning or the event database. This was true because multiple events within a single county were sometimes counted as separate events in the event database, whereas only one warning could be in effect for a particular county at the same time. For this reason, the number of warned events in the event database, denoted below as A_e, may exceed the number of verified warnings in the warning database, denoted below as A_w. Using these conventions, the definitions of *POD* and *FAR* are

$$POD = \frac{A_e}{A_e + B} \ ,$$

$$FAR = \frac{C}{A_w + C} \ .$$

Given these expressions for *POD* and *FAR*, the *CSI* formula is:

$$CSI = \frac{A_w A_e}{A_W A_e + A_w B + A_e C} \ .$$

3.4 Frequency of Hits *(FOH)*

$$FOH = \frac{A}{A + C} \ ,$$

where:

FOH = the frequency of hits.

A = the number of correct yes forecasts (see Table 1-2). In warning verification, it is the number of verified warnings.

C = the number incorrect yes forecasts. In warning verification, it is the number of unverified warnings (also known as false alarms).

In the case of warnings, the *FOH* is computed from the warning database and is the number of

verified warnings divided by the total number of warnings. The more often warnings verify, the better the score. The best possible *FOH* is one, the worst possible score is zero.

4 Scores Computed for Specific Forecast Elements

4.1 Temperature, Wind Speed and Direction, and Wave Height

Scores frequently computed for forecasts of temperature, wind speed and direction, and wave height include:

a. **Mean Error *(ME)*** indicates whether collective forecast values were too high or too low. This is also called the mean algebraic error.

$$ME = \left(\frac{1}{N}\right)\sum_{i=1}^{N}(f_i - o_i) \quad ,$$

where:

f_i = forecast for the *ith* case,
o_i = observation for the *ith* case,
N = the number of cases.

b. **Mean Absolute Error *(MAE)*** measures error without regard to the sign (whether positive or negative).

$$MAE = \left(\frac{1}{N}\right)\sum_{i=1}^{N}|f_i - o_i| \quad ,$$

where:

f_i = forecast for the *ith* case,
o_i = observation for the *ith* case,
N = the number of cases.

c. **Root Mean Square Error *(RMSE)*** weights large errors more than the MAE.

$$RMSE = \sqrt{\left[\frac{\sum_{i=1}^{N}(f_i - o_i)^2}{N}\right]} \quad ,$$

where:

f_i = forecast for the *ith* case,
o_i = observation for the *ith* case,
N = the number of cases.

d. **Measuring Errors Against Some Standard**

The above measures of accuracy *(ME, MAE, RMSE)* may also be computed for some forecast standard, such as Model Output Statistics *(MOS)* guidance, climatology *(CLI)*, or persistence *(PER)*. For example, the *MAE* for *MOS* guidance forecasts (m_i) is

$$MAE_{MOS} = \frac{1}{N}\sum_{i=1}^{N} |m_i - o_i| \ ,$$

where:

m_i = forecast MOS for the *ith* case,
o_i = observation for the *ith* case,
N = the number of cases.

Forecast skill may be determined by measuring the improvement of forecasts over some forecast standard, such as MOS, climatology or persistence. For example, the *MAE* may be used to compute the percent improvement of forecasts over MOS.

4.2 Probability of Precipitation

Scores typically computed for probability of precipitation verification include:

a. **Brier Score** measures the mean square error of all PoP intervals forecast. The standard NWS Brier score, defined below, is one-half the original score defined by Brier (1950).

$$BS = \frac{1}{N}\sum_{i=1}^{N} (f_i - o_i)^2 \ ,$$

where:

f_i = forecast probability for the *ith* case,
o_i = observed measurable precipitation for the *ith* case (no = 0, yes = 1),
N = the number of cases.

b. **Climatological Brier Score** is an application of the Brier score to climatic relative frequencies.

$$BS_{CLI} = \frac{1}{N}\sum_{i=1}^{N} (c_i - o_i)^2 \ ,$$

where:

c_i = the climatic relative frequency,
o_i = observed measurable precipitation for the *ith* case (no = 0, yes = 1),

N = the number of cases.

c. MOS Brier Score

$$BS_{MOS} = \frac{1}{N}\sum_{i=1}^{N}(m_i - o_i)^2 \ ,$$

where:

m_i = MOS guidance probability for the *ith* case. These are forecast to the nearest 0.01; however for NWS PoP verification, the m_i values are rounded to one of the following values: 0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0, and

o_i = observed measurable precipitation for the *ith* case (no = 0, yes = 1), N = the number of cases.

d. **Improvement over Climate, MOS, or Persistence Based on Brier Score** measures the improvement gained from actual forecasts versus some standard measure, such as climatology, the MOS forecast, or persistence. For example:

$$I(BS)_{MOS} = \left(\frac{BS_{MOS} - BS}{BS_{MOS}}\right)100 \ ,$$

where:

BS = Brier score calculated from a set of local forecasts, and

BS_{MOS} = Brier score calculated from a set of MOS products matched in time and space to the set of local forecasts.

e. Relative Frequency of an Event is the fraction of the time an event occurred.

$$RF = \frac{1}{N}\sum_{i=0}^{N} o_i \ ,$$

where:

o_i = observed measurable precipitation for the *ith* case (no = 0, yes = 1), N is the total number of events.

f. Reliability, a measure of bias, compares the number of forecasts of an event with the observed relative frequency of the event. The reliability may be determined overall or by forecast interval, e.g., 10 percent PoP intervals or (0, 5, 10, 20, 30, . . ., 80, 90, 100.

$$\frac{1}{N}\sum_{i=1}^{N} f_i \quad \text{compared with} \quad \frac{1}{N}\sum_{i=1}^{N} o_i \ ,$$

where:

f_i = forecast probability for the *ith* case,
o_i = observed measurable precipitation for the *ith* case (no = 0, yes = 1),
N = the total number of events or the number of events in the interval.

If the number of forecasts of the event or interval is larger (smaller) than the observed relative frequency of the event or interval, the event or interval was over-forecast (under-forecast).

4.3. QPF

a. Bias, Threat Score, *POD*, and *FAR*

$$B = \frac{F}{O} \ ,$$

$$TS = CSI = \frac{H}{F + O - H} \ ,$$

where:

B = Bias,

TS = Threat Score,

F = the number of points forecast to have at least a certain amount (a determined threshold) of precipitation, e.g., one inch,

O = the number of points observed to have at least the threshold amount of precipitation,

H = the number of points with correct forecasts for the threshold amount of precipitation,

When the bias is less [greater] than unity for a given threshold, the forecast is under- [over-] forecasting the areal coverage for that amount.

Geometrically, the threat score for a given threshold amount represents the ratio of the correctly predicted area to the threat area. Threat area is defined as the envelope of forecast and observed areas for that threshold. A perfect forecast results in a threat score equal to one, and a forecast with no areas correctly predicted receives a zero. The threat score, therefore, provides a measure of how accurately the location of precipitation is forecast within the valid period of the forecast. To receive a high threat score, forecast precipitation is accurate—both

spatially and temporally. For example, if a 1.00-inch isohyet is forecast, and all the observed rainfall within that area ranges from 0.8 to 0.99 inch, the forecaster's 1.00-inch threat score would be zero. However, the 0.8 to 0.99 inch area would favorably affect the 0.5-inch threat score. Also, a forecast area that is adjacent to an observed area with no overlap produces a zero threat score, and forecasts that are incorrect by just a couple of hours may receive little or no credit. Closely related to the threat score are *POD* and *FAR*, which are expressed as:

$$POD = \frac{H}{O} \ ,$$

$$FAR = \frac{F - H}{F} \ .$$

b. Equitable Threat Score (Messinger 1996):

$$ETS = \frac{H - E}{F + O - H - E} \ ,$$

where:

F = the number of points forecast to have at least a certain amount (a determined threshold) of precipitation, e.g., one inch,

O = the number of points observed to have at least the threshold amount of precipitation,

H = the number of points with correct forecasts for the threshold amount of precipitation,

$E = \frac{FO}{N}$,

N = the number of points verified.

E is substantial for low precipitation categories, i.e., 0.10 inch or less in 24 hours, small at intermediate categories, and negligible for high categories, i.e., 1 inch or more in 24 hours.

4.4 Ceiling Height and Visibility

The Log Score is used for verifying ceiling height and visibility forecasts. It emphasizes accuracy in the more critical lower ceiling height and visibility ranges.

$$LS = \left(\frac{50}{N}\right) \sum_{i=1}^{N} \left| Log_{10} \frac{f_i}{o_i} \right| ,$$

where:

f_i is the category of the *ith* forecast,
o_i is the category of the *ith* observation.

Note, f_i and o_i may also be used to represent the actual respective forecast and observed values of the element (i.e., ceiling height in feet, visibility in statute miles). Persistence is often used as the reference standard for evaluating ceiling height and visibility forecasts. The last hourly observation available to the forecaster before dissemination of the terminal aerodrome forecast defines the persistence forecasts of ceiling height and visibility to which the TAFs are compared.

4.5 Aviation Weather Center (AWC) Verification Statistics

The following statistics are used for verifying AWC forecasts:

a **Probability of Detection** (See section 3a of this appendix).

b. **False Alarm Ratio** (See section 3b of this appendix).

c. **Probability of Detection of "No" Observations** is an estimate of the proportion of null events ("no" observations) that were correctly forecast (i.e., PIREPs include reports such as negative icing or negative turbulence). An alternative name for this statistic is the probability of null events *(PON)*. Using Table 1-2 (located in section 3 of this appendix),

$$POD[N] = PON = \frac{X}{X + C} ,$$

where:

X = the number of correctly forecast null events, and
C = the number of incorrectly forecast null events.

d. **Percent Area** is the percentage of the forecast domain's area where the forecast variable is expected to occur. It is the percent of the total area with a YES forecast.

e. **Percent Volume** is the percentage of the forecast domain's volume where the forecast variable is expected to occur. It is the percent of the total volume with a YES forecast.

5 Ruth-Glahn Convergence Score

This section was taken from Ruth et. al, 2009. This score does not measure accuracy. Accuracy is measured by other scores, such as the mean absolute error. While forecast accuracy is

important, forecast continuity from cycle-to-cycle is also important. Forecasts that exhibit continuity and change relatively little for a given forecast valid time or period, from the time they are first issued (e.g., seven days in advance) until the last time they are issued (e.g., 12 hours in advance), are easier for forecast users to interpret than forecasts that jump back and forth from cycle to cycle. The Ruth-Glahn forecast convergence score (FCS) measures forecast continuity. When considering *n* forecasts made over a number of days for subsequent forecast cycles that decrease in forecast projection until the valid time of the forecast is reached, the FCS is defined as follows:

$$FCS = \frac{T_1 + T_2}{T_3 + T_4}.$$

The first term (T_1) is the number of forecasts that changed insignificantly (less than a given threshold) from the previous forecast F_{i-1} or that moved closer to the next forecast F_{i+1}, where i varies from 2 to n. When $i = n$, the observation is used as the next forecast F_{i+1}:

$$T_1 = \sum_{i=2}^{n} \begin{cases} 1 \text{ if } | F_i - F_{i-1} | < \text{ threshold} \\ 1 \text{ if } | F_i - F_{i+1} | < |F_i - F_{i-1}|. \\ 0 \text{ if neither of the above} \end{cases}$$

The second term (T_2) is the difference between the first and last forecasts scaled by the significance threshold:

$$T_2 = \frac{| F_n - F_1 |}{\text{threshold}}.$$

The third term term (T_3) is the number of possible forecast changes:

$$T_3 = n - 1.$$

The fourth term (T_4) is the sum of the forecast changes scaled by the significance threshold:

$$T_4 = \sum_{i=2}^{n} \frac{| F_i - F_{i-1} |}{\text{threshold}}.$$

The T_1 and T_3 terms account for the actual and possible numbers of swings, respectively. The T_2 and T_4 terms account for the magnitudes of the swings. The significance threshold specifies the minimum change necessary to count as a swing.

The FCS ranges from near zero (many large forecast swings) to 1.0 (no swings). Repeatedly forecasting the same value will yield a perfect FCS of 1.0. This score says nothing about forecast accuracy, as accuracy may be sacrificed for forecast continuity, or vice versa.

6 References

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1-3.

Burroughs, L.D., 1993: National marine verification program - verification statistics. *OPC Technical Note/NMC Office Note No. 400,* National Weather Service, NOAA, U.S. Dept. of Commerce, 48 pp.

Flueck, J.A., 1987: A study of some measures of forecast verification. *Preprints, 10th Conference on Probability and Statistics in the Atmospheric Sciences.* Edmonton, AB, Canada, American Meteorological Society.

Gandin, L.S., and A.H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Monthly Weather Review,* **120**, 361-370.

Gerrity, J.P., 1992: A note on Gandin and Murphy's equitable skill score. *Monthly Weather Review,* **120,** 2709-2712.

Hanssen, A.W. and W.J.A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *Mededeelingen en Verhandelingen,* Royal Netherlands Meteorological Institute, **81.**

Jolliffe, I.T. and D.B. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, Ltd., 240 pp.

Livezey, R.E., 2003: Categorical events (chapter 4). *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* Edited by I.T. Jolliffe and D.B. Stephenson, John Wiley and Sons, Ltd., 240 pp.

Messinger, F., 1996: Improvements in precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bulletin of the American Meteorological Society,* 77, 2637-2649.

Peirce, C.S., 1884: The numerical measure of the success of predictions. *Science,* **4,** 453-454.

Ruth, D.P., B. Glahn, V. Dagostaro, and K. Gilbert: The performance of MOS in the digital age. *Wea. Forecasting,* **24,** 504-519.

Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences, Volume 100, Third Edition (International Geophysics Series).* Elsevier (Academic Press), Oxford, 676 pp.

APPENDIX B – GLOSSARY OF CONTRACTIONS AND TERMS

AOP	Annual Operating Plan
CFW	Coastal Flood Warning
C-MAN	Coastal-Marine Automated Network (hourly weather observations)
CONUS	Contiguous United States
CPC	Climate Prediction Center
CPHC	Central Pacific Hurricane Center
CSI	Critical Success Index, see Appendix A, Section 3.3.
CWF	Coastal Waters Forecast
GFS	Global Forecast System Model
EF Scale	Enhanced Fujita Scale
ESS	Equitable Skill Score
FAR	False Alarm Ratio, see Appendix A, Section 3.2
FFW	Flash Flood Warning
FOH	Frequency of Hits
FLW	Flood Warning
GLF	Great Lakes Open Lake Forecast
GSS	Gerrity Skill Score
HRAP	Hydrologic Rainfall Analysis Project (frequently used 4-km grid)
HSS	Heidke Skill Score
LAMP	Local AWIPS MOS Program
LST	Local Standard Time
MAE	Mean Absolute Error, see Appendix A, Section 4.1
METAR	Meteorological Aviation Reports
MDL	Meteorological Development Laboratory
ME	Mean Error (algebraic), see Appendix A, Section 4.1
MOS	Model Output Statistics
MVF	Marine Verification Forecast (coded)
NAM	North American Mesoscale Model
NDFD	National Digital Forecast Database
NFDRS	National Fire Danger Rating System
NFWOC	National Fire Weather Operations Coordinator
NHC	National Hurricane Center
NPMC	National Performance Management Committee
NPVU	National Precipitation Verification Unit
NSH	Near Shore Forecast (Great Lakes)
OFF	Offshore Forecast
OPC	Ocean Prediction Center
PFM	Point Forecast Matrix (coded public forecast at points)
POD	Probability of Detection, see Appendix A, Section 3.1
PoP	Probability of Precipitation
PROB	Probabilistic Forecast in a TAF
PSS	Peirce Skill Score
QPE	Quantitative Precipitation Estimate (past precipitation analysis)

QPF	Quantitative Precipitation Forecast
RFC	River Forecast Center
RFW	Red Flag Warning
RMSE	Root Mean Square Error, see Appendix A, Section 4.1
RTMA	Real Time Mesoscale Analysis, provided hourly for select elements
SMW	Special Marine Warning
SPECI	Special Aviation Weather Reports
SVR	Severe Thunderstorm Warning
TAF	Terminal Aerodrome Forecast
TCM	Tropical Cyclone Forecast/Advisory
TOR	Tornado Warning
TEMPO	Temporary Forecast Conditions in a TAF
VCTS	Thunderstorms in the vicinity (within a 5- to 10-mile radius) of the aerodrome
VTEC	Valid Time and Event Code
WMO	World Meteorological Organization
WPC	Weather Prediction Center (formerly Hydrometeorological Prediction Center)

Change – For Terminal Aerodrome Forecasts (TAF), this term is used to describe observation variability, with regard to scoring the Operational Impact Forecast and TEMPO evaluation. For specifics regarding each element, see section 6.1.6 a.

Area Corrected Lead Time – For flash flooding, the warned area lead time is multiplied by the percentage of the area warned.

Hit – A correct forecast, as defined by a contingency table or some forecast error threshold value.

Lead Time – The amount of advance notice provided by a watch or warning concerning some operationally significant or life-threatening weather phenomenon. Negative lead times (when the warning is issued after the event is first observed) are recorded as zero.

Operational Impact Forecast – A two-step process for determining whether to base the verification of a TAF on (a) the prevailing forecast or (b) the TEMPO or PROB forecast, whichever may be in effect at a given projection time. For more detail, see Appendix A, section 6.1.7.2.

Percentage of the Area Warned – With flash flood warnings, the area of reported flash flooding and the forecast area of the warning are overlaid to compute the percentage of the event area that was warned.

Storm Data - NOAA's official publication which documents the occurrence of storms and other significant weather phenomena having sufficient intensity to cause loss of life, injuries, significant property damage, disruption to commerce, and other noteworthy meteorological events.

Timing Error – In warning verification, the timing error is defined as the event beginning time

minus the forecast start time in the warning.

Warned Area Lead Time – For a flash flood event, the warned area lead time is calculated by subtracting the warning issuance time from the time when the event began. Negative lead times are set to zero.