A CROWDSOURCED HAIL DATASET: POTENTIAL, BIASES, AND INACCURACIES

by

Joseph R. Pehoski

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Mathematics

at

The University of Wisconsin-Milwaukee

December 2013

ABSTRACT

A CROWDSOURCED HAIL DATASET: POTENTIAL,
BIASES, AND INACCURACIES

by

Joseph R. Pehoski

The University of Wisconsin-Milwaukee, 2013
Under the Supervision of Dr. Kyle Swanson

Hail is a substantial severe weather hazard in the USA, with significant damage to property and crops occurring annually. Traditional methods of forecasting hail size have limited accuracy, and despite improvements in remote sensing of precipitation, the fall characteristics of hail make quantification of hail imprecise. Research into hail is ongoing, but traditional hail datasets have known biases and low spatiotemporal resolution. The increased usage of smartphones creates the opportunity to use a crowdsourced dataset provided by the Precipitation Identification Near the Ground (PING) program, a program developed by the National Severe Storms Laboratory. PING data is compared to approximate ground truth in the form of preliminary Severe Prediction Center (SPC) hail reports, and National Weather Service (NWS) issued severe warning polygons. Biases and inaccuracies in the dataset are also explored through exploratory data analysis.

While PING reports did not suffer from biases based on time of day or day of week, the location of PING reports was found to have a heavy bias towards high population density areas compared to SPC reports. Skill scores of PING

reports, compared to SPC reports, were low, with a remarkably high False Alarm Rate (FAR), indicating false reports being a problem in the PING dataset. Comparing PING reports to severe polygons did not substantially improve the skill scores. The low number of severe PING reports prevented any meaningful analysis of size accuracy. While the number of SPC reports were mostly correlated with the number of warning polygons issued by each Weather Forecast Office, the PING reports were not well correlated, with an anomalously high number of reports in the Oklahoma City region. The inaccuracy of PING reports and strong population bias suggest that the PING hail database may not have high utility, and should only be used in conjunction with other databases in order to ensure quality.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Thanks go to Dr. Kyle Swanson for advising me in the writing of this thesis, and to Drs. Clark Evans and John Kahl for reviewing the completed thesis. Thanks also go to Dr. Kimberly L. Elmore for help with the PING dataset, and the American Geographical Society Library at UWM for procuring the US Census Data.

1. Introduction

Hail is a significant severe weather phenomenon affecting people, crops, and property throughout North America and the world. Hail storms occurring in the US during the 1990s caused an average of $2.4 billion a year in damage to property and crops (Changnon 1999). Strong and Lozowski (1977) found that crop damage increases logarithmically with kinetic energy, and kinetic energy of a hailstone is exponentially related to its diameter (Brimelow and Reuter, 2002). Thus, determination of maximum hail size (defined as the diameter of the major axis) is important when considering severe weather risk. More accurate forecasts of hail size are of interest in climatologically high risk areas. Hail research necessitates a high quality hail dataset to use as “ground truth”. Specifically, a country-wide, high resolution ground truth dataset is desirable, due to the highly variable spatiotemporal nature of hail, and the difference in hail climatology in the U.S (Cintineo et al. 2012; Witt et al. 1998b). Such a dataset is practically difficult to construct due to the highly variable spatiotemporal distribution of hail, and the current inability to accurately remotely sense hail. With the rise of social media and smartphone ownership, new methods of data collection have become possible. This thesis explores the accuracy of one particular crowdsourced dataset, Precipitation Identification Near the Ground (PING), and the potential for current or future use of this data in subsequent hail studies.

Hail studies use a variety of different datasets as ground truth, but one that is accurate over a large domain is unavailable. Studies relying on high

quality hail data generally have to choose between a spatially large domain and low resolution data, or a spatially small domain and high resolution data (Witt et al. 1998b*)*. Field projects are the traditional source of highly accurate, high resolution data. The development of Hailcast, a popular numerical hail model, used observations from the Alberta Hail Project, which took place in the plains of southern Alberta. The study area comprised 33,700km^2, and used telephone surveys, requests for hail reports sent through the mail, and hailpads for an approximate observer density of one per three square kilometers (Brimelow and Reuter 2002). The National Center for Atmospheric Research's Real-time Analysis and Prediction of Storms (RAPS) project used two chase vehicles in the central Colorado plains (Kessinger and Brandes 1995). The scope of field projects is usually limited, due to the expense involved. The Alberta Hail Project lasted from 1957 to 1985, while the RAPS covered June and July of 1992 and 1993 (Brimelow and Reuter 2002; Kessinger and Brandes 1995*).* As social media has become more popular, hail surveys using video and photo sharing websites have proved useful (Blair and Leighton 2012). There is an ongoing effort to collect high resolution severe data using telephone surveys over the entire CONUS through the Severe Hazards Analysis and Verification Experiment (SHAVE) (Ortega et al. 2009). Results from SHAVE have already been used to determine the skill of a multi radar multi sensor algorithm, and to find the optimal threshold for severe hail in "An Objective High Resolution Hail Climatology of the Contiguous United States" (Cintenio et al. 2012). These methods of collecting a

reports based dataset are promising, but expensive, and not typically comprehensive over the entire USA.

Hail studies often utilize the *Storm Data* severe storm database as verification of ground truth (Witt et al. 1998a; Witt et al. 1998b; Blair and Leighton 2012). Storm Data is a database maintained by the National Climate Data Center, and includes hail reports dating back to 1955 (Hales 1993). Storm Data is collected from the various local NWS offices, which have over time used various criteria for collecting reports (Doswell et al. 2005). Verification of observer reports began in 1980, but different local NWS offices have different criteria for recording reports, and different rates of storm spotter deployment. (Hales 1993, Doswell et al. 2005). Different NWS districts show a marked dissimilarity in report density (Hales 1993). Reports collected by National Weather Service offices are used primarily for warning verification, so report density is too low for useful hail research purposes (Blair and Leighton 2012). SPC recommends that non-tornadic severe weather reports be separated by at least 10 miles and 15 minutes, but this recommendation is not typically followed (Doswell et al. 2005). Observations also demonstrate a very problematic population bias, particularly at the lower end of the severe scale. The effect is not as apparent when considering larger hail sizes (Hales 1993; Cintenio et al. 2012).

Various methods have been attempted to deal with the problems in Storm Data. While generally accepted that population bias is a problem in severe weather report datasets, there are some studies indicating the opposite (Doswell et al. 2005; King 1997. Davis and LaDue 2004). This may indicate that population

bias itself is not a factor of only population, but of other (possibly sociological) factors. Cintineo et al. (2012) attempts to tune a multiradar multisensor algorithm in order to develop a remotely sensed dataset of any-hail and severe-hail over the entire Nexrad period. Elsner et al. (2013) describes a method which weights tornado reports by their distance from population centers. Doswell et al. 2005, which discusses the climatology of nontornadic severe weather, smooths the data temporally and spatially in order to wash out any details. Such methods, while acceptable for climatological studies, are not appropriate for hail studies in which higher resolution data is required.

Remote sensing of hail size has low accuracy, and research into improving hail size estimation is ongoing (Straka et al. 2000; Edwards and Thompson 1998). Indeed, the promise of remote sensing of hail, and the current low skill, is the impetus behind many high resolution observation datasets. A variety of radar algorithms for detecting hail exist, usually placing it into two categories: any hail, and severe hail. Severe hail is defined as hail greater than 1 inch in diameter (quarter sized), although it was formerly defined as hail with a diameter greater than .75 inches (penny sized), before 2010. More basic hail detection techniques rely only on radar reflectivity, and not on any polarimetric or derived products. One technique uses the height of the 50 dBz level above the freezing level, which is analogous to the amount of supercooled water in the updraft. (Richter and Deslandes 2007; Donavon and Jungbluth, 2006). Hail can be inferred from the presence of a three-body scatter spike (TBSS), although this is not necessary for the presence of large hail. High reflectivities resulting in the

Mie scattering which cause a TBSS can be due to wet, non-severe hail (Richter and Deslandes 2007). Similarly, a bounded weak echo region (BWER) is operationally used as an indicator of large hail, but it is caused by presence of a mesocyclone, not by hail, and indicates only high updraft velocities.

More quantitative radar derived estimates of hail size are calculated from a number of different algorithms. Vertically Integrated Liquid (VIL) based algorithms have been used since the early 1970s (Greene and Clark 1972). VIL, by itself, has proven to have no skill as an indicator of hail size. By examining temperatures aloft or other meteorological parameters, efforts to produce a "VIL of the day" which results in better warning verification, have been attempted (Amburn and Wolf 1997; Donavon and Jungbluth 2006). VIL of the day considers the current atmospheric temperature profile, and attempts to determine a minimum VIL for the conditions that will result in hail. This method is fraught with problematic assumptions, and has, in practice, proved a poor indicator of hail size (Amburn and Wolf 1997). More recent non-VIL based methods take advantage of advances in computing power. The Probability of Severe Hail (POSH), Severe Hail Index (SHI), and Maximum Expected Size of Hail (MESH) are commonly used algorithms, and adopted by the National Weather Service (Kessinger and Brandes 1995; Witt et al. 1998a). Parameters for these algorithms are by necessity tuned using observer based datasets. A review of these algorithms show that the variance in observed hail size for a given algorithm estimated hail size is quite large, and they thus cannot be used as a one-to-one hail size indicator (Wilson et al. 2009; Cintenio et al. 2012). While

beneficial for producing climatologies, where the error due to nonmeteorological factors may outweigh the error in the radar algorithm used, greater accuracy is desired for many research applications (Doswell et al. 2005, Cintenio et al. 2012).

Polarimetric radar has the potential to improve upon the hail size prediction skill of single pole radar, but research in the area is ongoing. The nationwide upgrade of the Next Generation Radar (NEXRAD) network to polarimetric radar has necessitated development of hydrometeor type algorithms, requiring tuning of the parameters. Unfortunately, while polarimetric radar is useful in differentiating most hydrometeor types and drop size distribution, the irregular shape of hail makes quantification of hail size a difficult problem (Zrnic et al. 1993; Straka et al. 2000). Hail tends to tumble as it falls through a storm, allowing it to be assumed to be spherical, and identified using the coincidence of high reflectivity and low Differential Reflectivity (ZDR). However, the dual-pol hail signals change based on the wetness of hail, isotropy, and surface characteristics, making accurate size detection problematic (Zrnic et al. 1993). Additionally, the fall characteristics of hail are in reality more complicated than other hydrometeors. Evidence exists both for hail falling oriented along its minor axis, as well as falling oriented along its major axis, not tumbling as is normally assumed (Straka et al. 2000). Hail size estimation by any remote sensing method remains crude at best, and thus cannot be used for research purposes where an accurate size of hail, and not just its presence, is a consideration.

The upgrade of the NEXRAD system to polarimetric radar led to the creation of the PING program by the National Severe Storms Laboratory, in order to collect precipitation type data. As the penetration of smartphones into rural markets increases, apps like PING have the potential to greatly decrease the difficulty of reporting, thus increasing the number of hail reports. There remain questions on how accurate the reports are, and how much the PING dataset could enhance more traditional datasets. This thesis is a first attempt at addressing those questions, in an attempt to reveal how useful projects like PING may prove.

The outline of the thesis is as follows: A description of the data, hail climatology, and methodology used is discussed in section 2. Section 3 will explore the data in further detail and discuss the results of the experiments performed. Section 4 provides a conclusion.

2. Methods and Data

2.1 Hail Climatology

Hail forms in the updraft of a strong thunderstorm in the presence of supercooled liquid water (Rogers and Yau 1989*)*. It has long been established that the vertical temperature profile and updraft strength are directly related to hail size. Highly unstable atmospheres are usually required for a strong updraft, which can support a large hailstone with a high terminal velocity for longer before it falls out of the storm. In practice, the height of high radar reflectivities is a good indicator of updraft strength in a storm. The height of the melting layer, and the

regime in which supercooled water can exist, is also important, with this data acquired traditionally through a numerical model or radiosonde. Several detection methods rely on the height of high reflectivities; For example, “Evaluation of a Technique for Radar Identification of Large Hail across the Upper Midwest and Central Plains of the United States” describes several methods relying on the height of high radar reflectivities above the freezing level (Donavon and Jungbluth 2007).

Hail in the USA falls mostly east of the Rocky Mountains (Cintenio et al. 2012). The maximum hail threat is in the southern Great Plains, with a lower threat along the entire lee of the Rockies. In general, changes in the vertical temperature profile in different regions of the US can explain differences in hail prevalence. In the Great Plains, near the Rockies, higher midlevel lapse rates are likely responsible for the increase in prevalence. The high number of hail reports extends into the plains of Western Canada, where the Calgary-Edmonton corridor sees a very high number of hail days per year. Because hail growth is a function of both supercooled liquid water accretion and sublimation, and because of this dependence on the environment’s temperature and humidity (Rogers and Yau 1989), hail growth regimes in tropical locations cannot be expected to be similar to that in mountainous regions. In more tropical regimes (warmer and more humid), hailstones can melt faster, due to less evaporative cooling in humid air (Cintineo et al. 2012). Thus, a VIL of the day equation for one region cannot be expected to be accurate for other regions (Paxton and Shepherd 1993; Donavon and Jungbluth 2006). A national, rather than a local dataset, is

important to account for climatological variations in the USA. (Edwards and Thompson 1998). In “An Objective High-Resolution Hail Climatology of the Contiguous United States”, different Heideke Skill Scores were achieved using the same methodology for the entire United States, with a significantly low score in the Northeast. Although climatology may have been the cause, a lack of reports in the Northeast was also identified as a potential contributor. Because of different airmass characteristics in different temporal and regional regimes in the US, it is reasonable to expect hail growth to work differently.

2.2 Description of the Data

PING is a crowdsourced dataset that is intended to improve the Hydrometeor Classification Algorithm (Park et al. 2008). Crowdsourcing, in general, can be described as the allocation of tasks to a large group of individuals with unknown expertise. PING is a crowdsourced dataset in that the task of collecting weather reports, traditionally performed by professional meteorologists, is instead performed by anyone with a smartphone. The PING program was initially launched in 2006, and allowed anyone with an internet connection to input a latitude, longitude, time, and precipitation type. The program changed in February of 2013 with the public release of the mPING mobile application. Anyone with access to an Android or iOS smartphone can easily input a precipitation type, with the application automatically submitting time, and GPS determined latitude and longitude. The interface is shown in figure 1. When the option for hail is chosen, the user is asked to input the size, from a

default size of .25 inches, increasing in increments of .25 up to 10.0 inches. Standard object sizes are also named at appropriate diameters. Users must have an active internet connection, either through Wi-Fi or wireless broadband, to use the app. The app is not able to cache an observation and submit it later, when an internet connection is available. Thus, an observer has to both be able to observe an event, and also be in an area with an internet connection. In terms of hail, this means they must be located in the location where the hail fell. The PING database records the latitude and longitude to one one-hundredth of a decimal degree, which is equivalent to about 1 kilometer in the midlatitudes (Cintenio et al. 2012). Observations can only be submitted once every 30 seconds. This was formerly limited to an observation every 5 minutes in order to dismiss microscale characteristics of storms. After the app was expanded on May 2, 2013 to allow observations of meteorological phenomena aside from hydrometeor type, the submission delay was reduced. Report times are rounded to the nearest minute.

As described earlier, there is no single best national reports based hail dataset. For this analysis, SPC preliminary hail reports are used as the ground truth dataset. Reports are considered preliminary because they have not yet been reviewed by the NCDC, which typically takes between 60 and 90 days to occur (NCDC website). Because of the temporal domain of the PING reports being analyzed, and in the interest of consistency in the dataset, the SPC preliminary reports will be used. SPC reports are collected from a number of different sources that are either professionally confirmed or come from professional sources. Methods of collection include storm spotters, media,

pictures on social networking sites, county, state and federal emergency management officials, local law enforcement officials, skywarn spotters, NWS damage surveys, newspaper clipping services, and the insurance industry (National Climatic Data Center 2013). The reports contain a latitude and longitude to one one-hundredth of a degree, a size to the nearest quarter inch, and a UTC time to the nearest minute.

The SPC dataset used here only contains reports of severe hail, that is, hail of size .75 inches (penny sized) or greater. Because this dataset is used primarily to verify warnings, reports of nonsevere hail are not included in the set. Severe thunderstorm warnings verify when a single severe hail report occurs in the warning polygon. If hail occurs, but it is below the severe threshold, it is not counted, but such a report may be useful in later research.

There exist entries in the PING dataset that are exact duplicates of one another (hail of the same size occurring at the exact same time and location), as well as hail of two different sizes at the same location and time. This is possible because if two users less than one kilometer apart report hail within the same 30 second period, the reports are identical, given the resolution of the dataset (Kim Elmore, personal communication). For this analysis, exact duplicates have been removed. When two reports were identical except for hail size, the report with the smaller size was removed. This is in keeping with the typical hail report methodology of reporting the largest of all hail falling in the observation area. In all, 237 exact duplicates were found, and 3 duplicates with different sizes were

found. Because an SPC report date begins at 1200Z and continues until 1159Z the next UTC day, PING report dates were changed to this format.

Population data was acquired from the United States Census Bureau. The data is based on the 2010 census, and gives the population density per census block. A census block is the smallest unit the U.S. Census subdivides, and is thus best representative of the population density where hail fell. The census data is joined to the hail reports using ArcGIS (ESRI, accessed September 30, 2013.

Hail is a unique weather variable, because an observer typically needs to be at the exact location where hail fell. Because of this, population density is a more serious problem in hail datasets (Doswell et al. 2005; Hales 1993). The number of hail reports submitted can be expressed as a function of whether hail fell, population density, interest of the observer, and ease of reporting (Schaefer et al. 2004; Elsner et al. 2013; Witt et al. 1998b; Davis and LaDue 2004; Hales 1993). The mPING app greatly increases the ease of reporting, possibly increasing the number of hail reports. While the interest of an observer who takes an interest in the PING program can be expected to be higher than average, it may not be as high as the observers involved in SPC observation collection. The PING dataset provides a unique opportunity to interrogate some sociological factors involved in weather observations. Exploratory data analysis will be used to examine possible factors

The interest variable encompasses a number of sociological concerns which an increase or decrease in hail reports can be attributed to. Population

density is not the only factor which can increase the number of reports. The engagement of an observer is also important. For example, if hail falls during the early morning, despite population density being high, fewer people are able to observe the hail. Daytime or nighttime differences in warning verifications were observed in Davis and LaDue 2004. A number of factors may change this interest variable, depending on the dataset. In terms of PING data, knowledge of the mPING app is a very important part of this variable. In the manner of Eisner et al. 2013, figure 2 shows the number of Google search results for “mPING” by month. After peaking at 100 hits in February, when the app was publicly released, the data shows a decline until 18 hits are registered by October. This is evidence for a decrease in interest as time goes on. The remarkableness of a storm can also be considered in here. It may be expected that people are more likely to be engaged in a severe storm if large hail is falling.

Ease of reporting can be described as how difficult it is for an observer to get their observation into a database. In the case of the SPC dataset, this can be expected to be difficult, unless Weather Forecasting Offices are actively looking for verification. This has been alleviated somewhat with the rise of social media. It is fairly simple for an observer to take a picture and upload it to the social media page of the local severe weather office. In the case of PING, it is very easy for a user to input the hailsize using the mPING app, but only as long as the user has an internet connection at the time and place of observation.

The number of reports received is important to make a hail dataset useful, but he accuracy of PING data is, of course, a concern. Accuracy of a hail report

is made up of three measures: Spatial accuracy, temporal accuracy, and size accuracy. A major difference, and major concern, of the PING dataset is the diversity of the attitudes in the observers. Outright incorrect reports, reported for no reason in particular, can be one source of incorrect reports. Less deliberate misuse of the mPING app, such as using it to submit a report at a time or place aside from where hail is observed, is also a concern. An observer reporting an incorrect size during a hail storm may also be expected (Doswell et al. 2005). Trapp et al. (2005) discusses the inaccuracy of observer based high wind reports. Hail size can more easily be measured or estimated than windspeed, but some error between categories is to be expected. The list of common size comparison objects in the mPING app may be expected to limit this type of error.

Another error which may be demonstrated in the dataset is confusion of hail with sleet. Such weather phenomena are unlikely to be confused by an experienced weather observer, and would not normally make their way into the SPC observation database. However, the public's unfamiliarity with hail, a rather rare event, results in sleet observations being included as small hail. By only using May through August hail reports in this study, these errors can be eliminated.

The spatial domain of the PING program encompasses the entire continental United States. In this thesis, the entire spatial domain will be used, but only PING reports from May through August of 2013 are used, because it is the only complete season after the introduction of the mPING app.

Figure 1. The mPING mobile app interface.

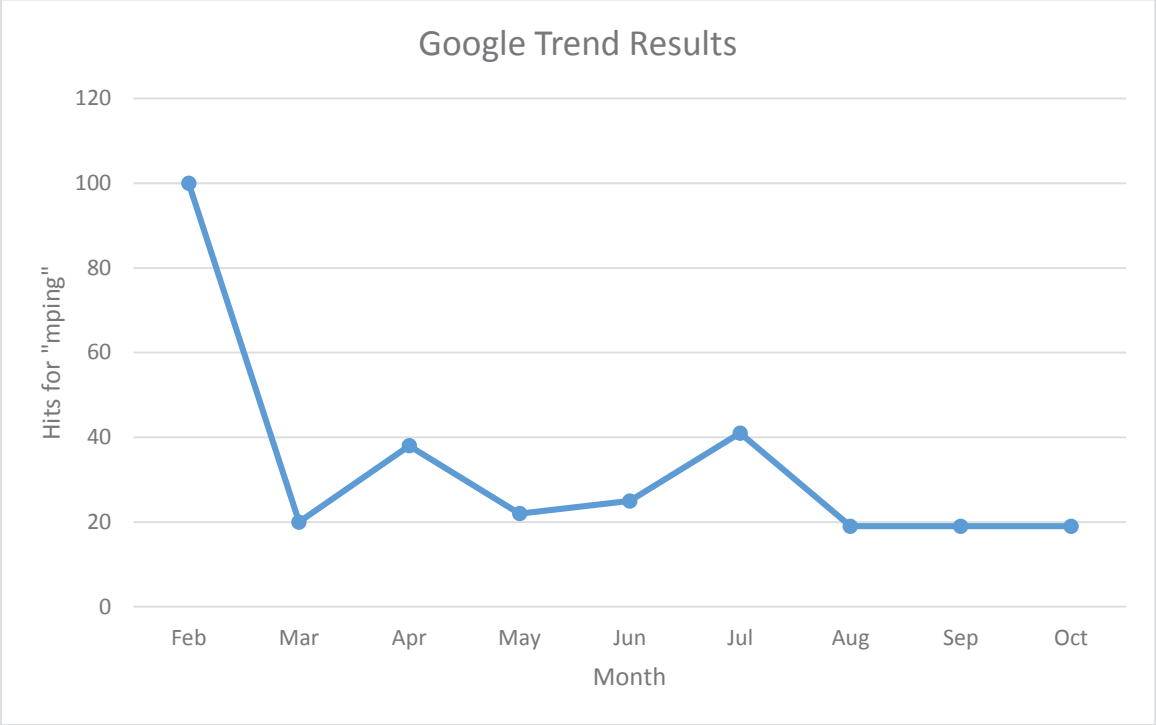


Figure 2: Google Trend results for “mping” from the public release of the app in February 2013, to October 2013.

2.3 Methodology

The number of factors affecting the accuracy of the data is large, and it is difficult to isolate any one of them. Ultimately, the question being asked is whether the PING dataset represents a subset of ground truth, and if that ground truth is able to augment a traditional dataset. The only way to definitively determine this is to compare the PING dataset and the SPC dataset to a completely accurate, high definition hail dataset. A perfect hail dataset would consist of the size and strike time of every piece of hail that falls to the ground within the given domain and resolution. Such a dataset is purely theoretical. For this study, SPC observations will be used as ground truth.

In this study, temporal variation will be mostly neglected. All severe reports in an entire day (1200Z to 1159Z the next day) will be compared. Both location and size accuracy will be determined using a contingency table, in the manner of Davis and LaDue, 2004. The False Alarm Rate (FAR), Probability of Detection (POD), and Critical Success Index (CSI) will be computed using multiple different search radii. The correspondence of hail reports with warning polygons will be determined, from which information on the accuracy of PING reports will be derived.

The skill scores are defined here as in Davis and Ladue (2004):

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

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

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

where A is the number of correct positives, B is the number of false nulls, and C is the number of false positives.

The standard search radius used will be 10km. Points will thus said to be concurrent if they are separated by 10km or less, and if they occurred on the same day. This distance is used because it falls in the middle of the meso-gamma scale. However, hail does not typically fall out of an entire thunderstorm, making the chosen radius larger than what might realistically be expected. A radius closer to 1km is more desirable, as this is approximately the resolution of both datasets (due to Latitude and Longitude being rounded to the nearest hundredth degree). However, due to the sparsity of reports in all databases used in the analyses, the number of concurrent points becomes too low to be useful if the search radius is set to a more reasonable value of 1km. The response of the skill scores to different search radii will be discussed.

Because of the problems with the methodology of using points as ground truth, a method based on Severe Thunderstorm Warning polygons will also be used. Such warnings are issued by Weather Forecasting Offices when hail is expected to reach one inch or greater, or high winds or a tornado are expected. It is assumed that inside a warning polygon, small hail at least can be expected. Maps of number of severe warning polygons (National Weather Service, accessed October 28, 2013) by county warning area (National Weather Service, accessed September 6, 2013) are examined.

Tornado warning polygons, while also expected to contain large hail, are not used in this analysis. This is primarily because tornado warnings have unique

biases not found in severe warnings. With public perception of tornado warned storms as much more dangerous than severe warned storms, countering against the increase in storm chasing in recent years, it is difficult to say how interest in gathering hail reports might change. Issuance of tornado warnings is a more critical problem, with more dire consequences for mistakes in timing and shape of the polygon. A lead time that is too long will result in increased risk to human life, while a short lead time will do the same (Simmons and Sutter 2007). Tornado warnings, because of these biases, must be considered separately from severe polynomials, and using this method is beyond the scope of this thesis.

Section 3: Results and Discussion

3.1 A closer look at the data

A map displaying all SPC and PING hail reports is given in figures 3 and 4 respectively. The report density is noticeably higher in the SPC dataset, as evinced in figures 5 and 6, with a different spatial distribution than the PING dataset. The location of hail reports in the SPC reports follows what may be expected based on climatology; the highest report density is between the Rocky Mountains and the western Great Plains. Even at this zoom level, population effects on observer density are obvious. Several large metropolitan areas are clear, as are some major highways. High density is also obvious in the Appalachians, and Northeast Ohio into upstate New York, which cannot be instantly explained by a population bias. These locally high values were also noted in Cintenio et al. 2012, when they found that the number of hail

observations in these locations significantly exceeded what would be expected from their objective hail climatology. Other factors must be used to explain this. While the distribution of these points seem to indicate a possible lake effect cause, another likely explanation is the effort with which weather forecast offices attempt to verify warnings (Hales 1993).

The map of all PING observations (figure 4) clearly shows population centers, and in fact looks very much like a population density map. The low number of reports in the PING dataset compared to SPC is partially responsible for this, allowing the population bias to be clearer. Figures 7 and 8 display cumulative distributions of reports based on population density. 94.9% of all SPC reports occur in areas with a population density below 3300 people per square mile, while 81.1% of all PING reports occur within that population density range.

The fraction of PING reports by location stays somewhat constant, with SPC observations becoming more dominant at higher latitudes and farther East (Figures 9 and 10). This is evidence of an increase in Weather Forecast Office warning verification efforts in the Northeast, along with a decrease in usage of the mPING app. Areas west of the Rocky Mountains have few hail reports in either dataset, so little insight can be gained in those regions. An outlier is clear at latitude 35N and longitude 97W, where the PING fraction increases greatly. This is the location of Oklahoma City, where the mPING program was developed at NSSL and University of Oklahoma, thus interest in the mPING program can be expected to be high.

The total number of observations for every day of the study period, overlayed with the fraction that are PING observations, are given in figures 11 through 14. PING fraction remains consistently low through the entire study period, aside from days when very few hail observations occurred. This could indicate that PING observations are subject to false positives, but this may also be the result of PING reports being submitted in unremarkable storms that are either unwarned, or for which a Weather Forecast Office does not put effort into collecting hail reports. Examining figure 15, it appears the usage of PING gradually declined as time progressed, which may be expected as interest in the program declined after the initial launch of the mPING app.

The dependence of reports on time of day is given in figures 16 and 17. The distribution of reports in both datasets follow what may be expected from climatology. The greatest number of reports occur in the late afternoon into the evening, with a lull overnight to noon. It may have been expected that the fraction of PING reports would decrease overnight, however, that is not the case.

Day of week should have no climatological difference in hail producing storms, so given a long enough period of time, the number of hail producing storms occurring each day of the week should be the same. However, changing behavior of people based on day of the week may affect the number of hail events submitted. Figure 18 shows the percentage of hail reports within the respective datasets that fell on every day of the week. There are no great outliers, with the PING fraction on Saturday farthest away from the expected value of .14, at .19.

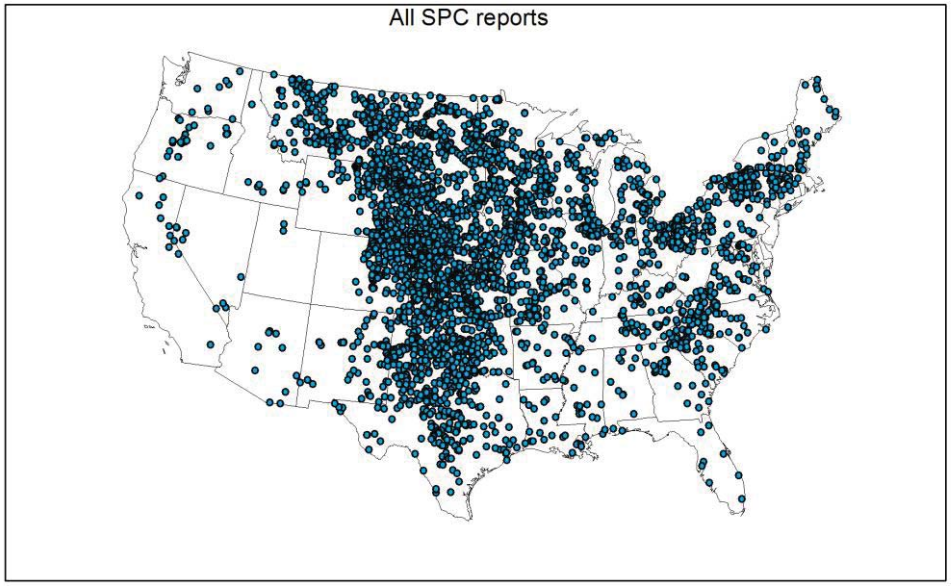


Figure 3: All SPC hail reports from May 1, 2013 through August 31, 2013

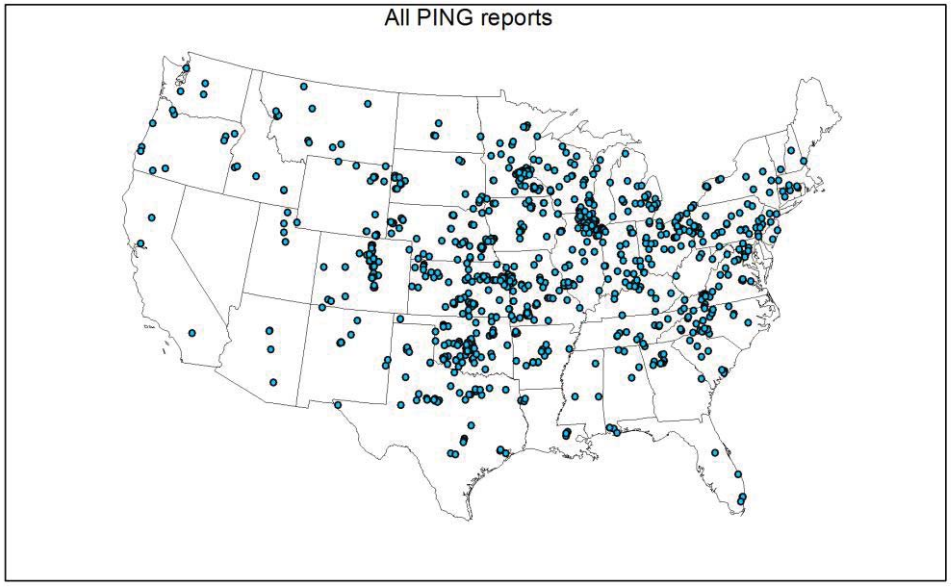


Figure 4: All PING hail reports from May 1, 2013 through August 31, 2013

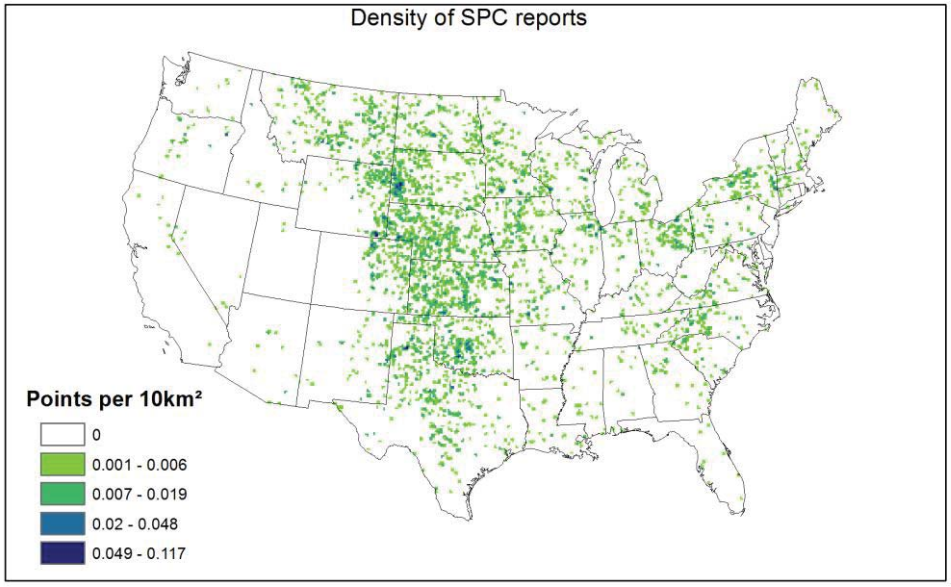


Figure 5: SPC report density.

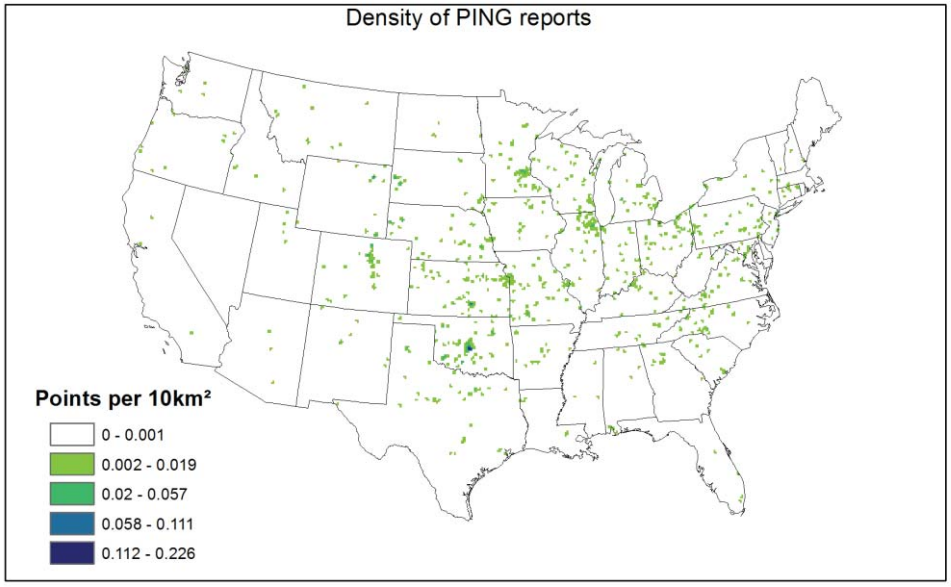


Figure 6: PING report density.

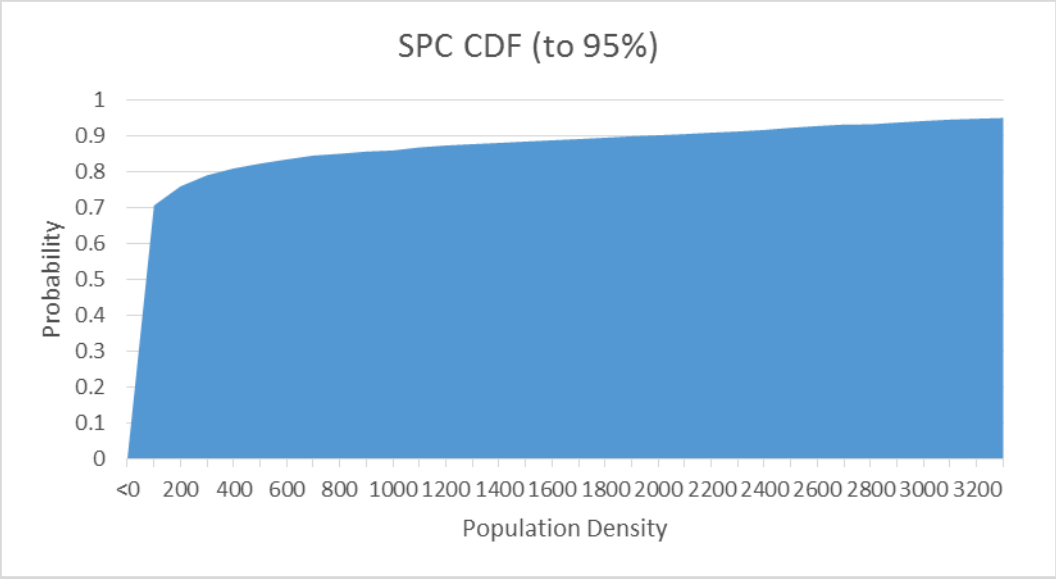


Figure 7: Population distribution of SPC hail reports, in people per square mile, shown for 95% of hail reports.

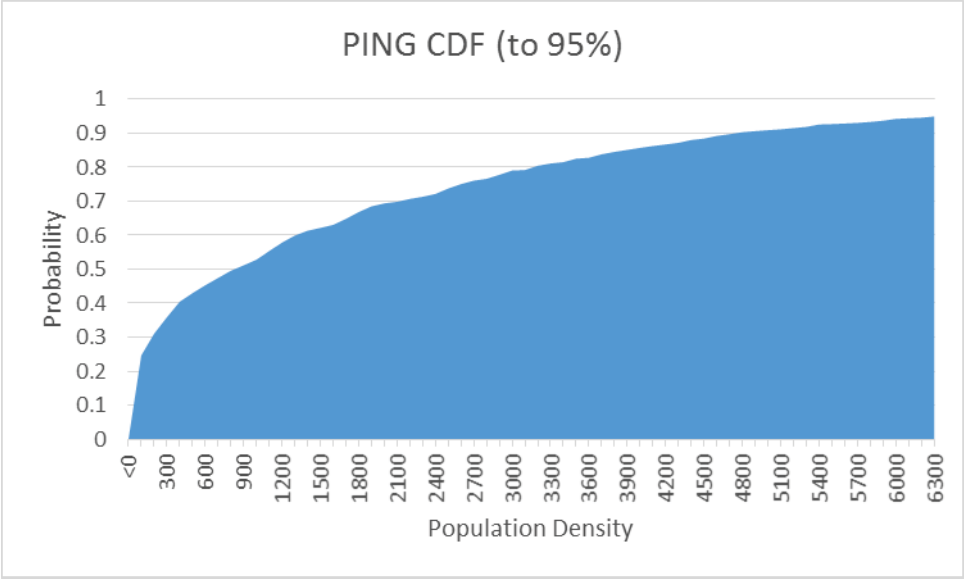


Figure 8: Population distribution of PING hail reports, in people per square mile, shown for 95% of hail reports.

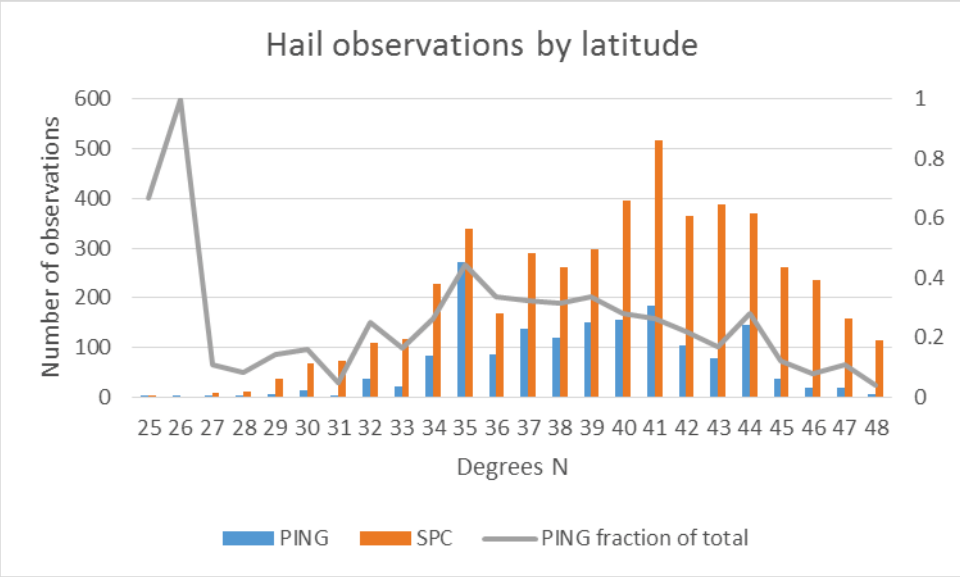


Figure 9: Number of hail observations by latitude for both PING and SPC, with the PING fraction of total hail reports for that latitude overlayed.

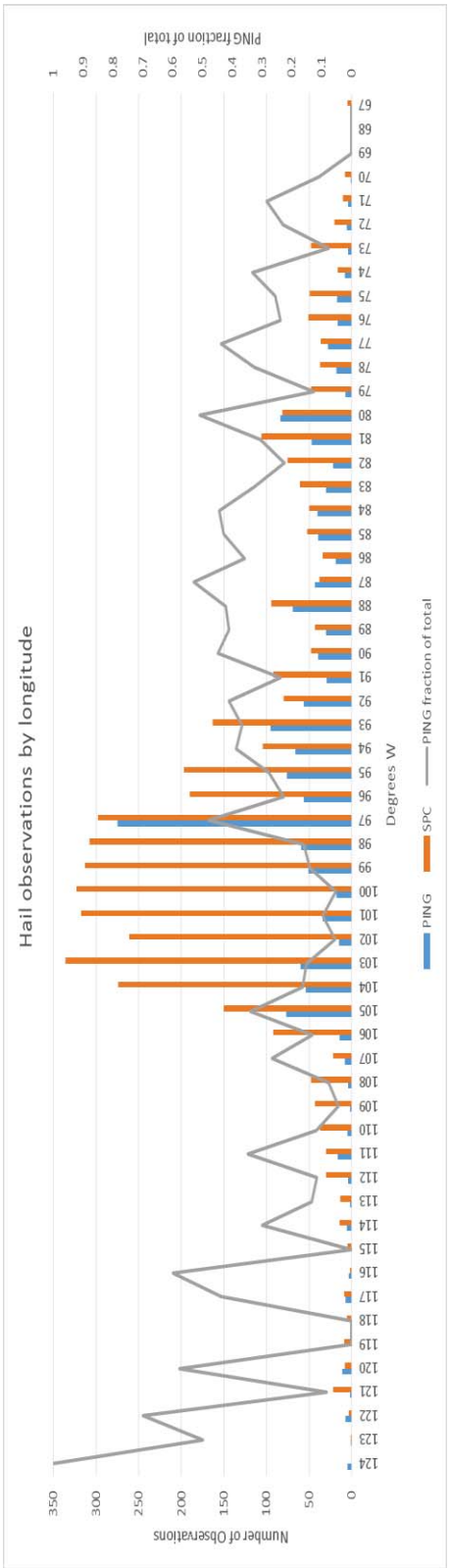


Figure 10: Number of hail observations by longitude for both PING and SPC, with the PING fraction of total hail reports for that longitude overlayed.

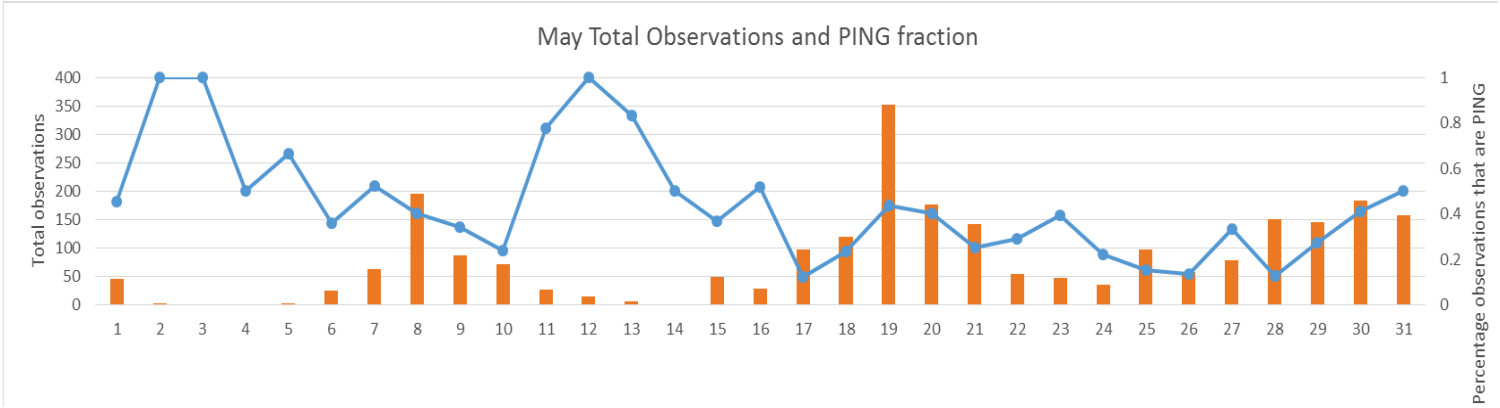


Figure 11: Sum of PING and SPC observations for the month of May, with the fraction of PING reports overlayed.

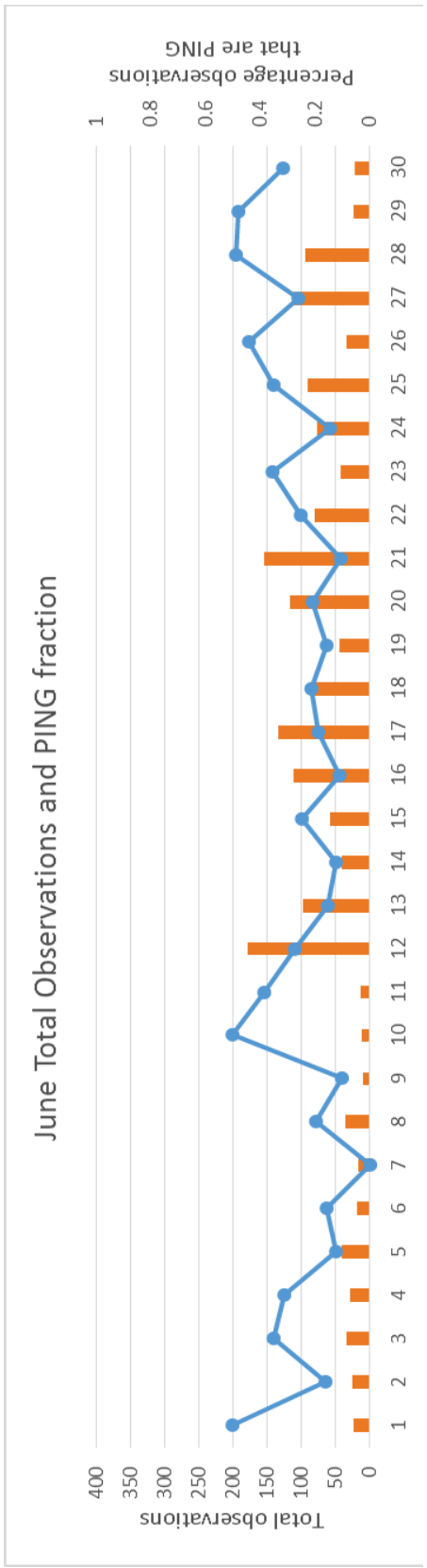


Figure 12: Sum of PING and SPC observations for the month of June, with the fraction of PING reports overlayed.

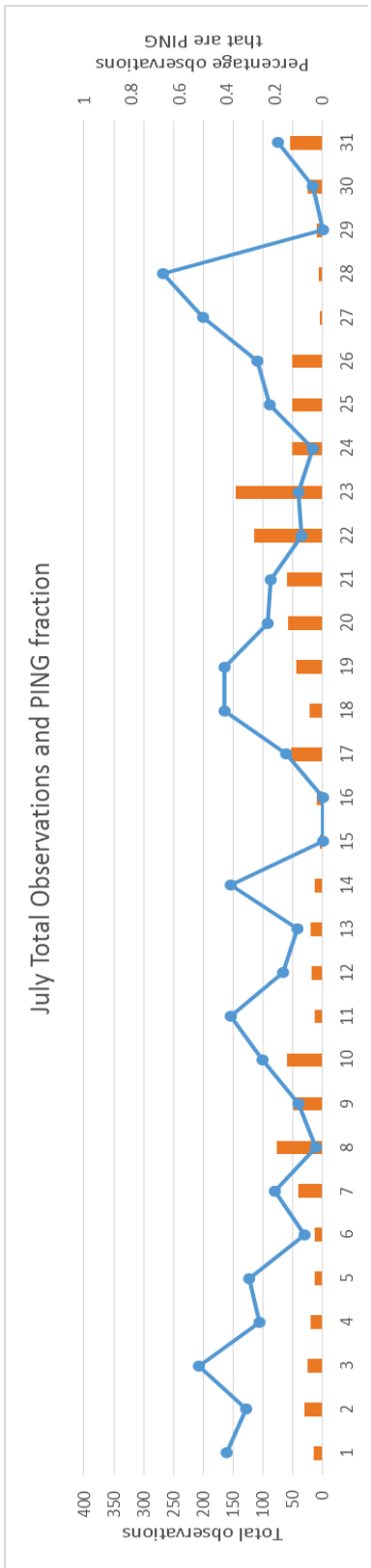


Figure 13: Sum of PING and SPC observations for the month of July, with the fraction of PING reports overlayed.

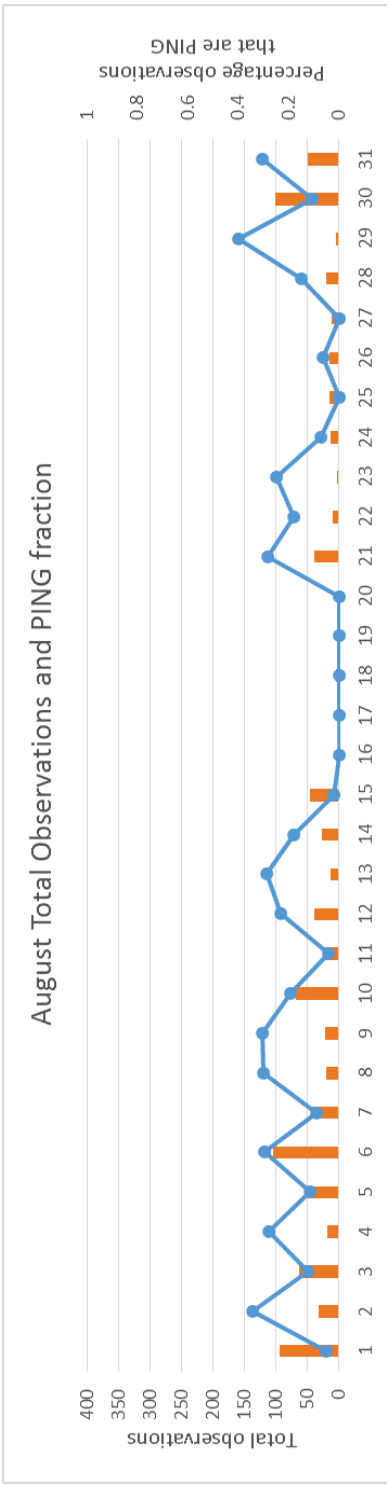


Figure 14: Sum of PING and SPC observations for the month of August, with the fraction of PING reports overlayed.

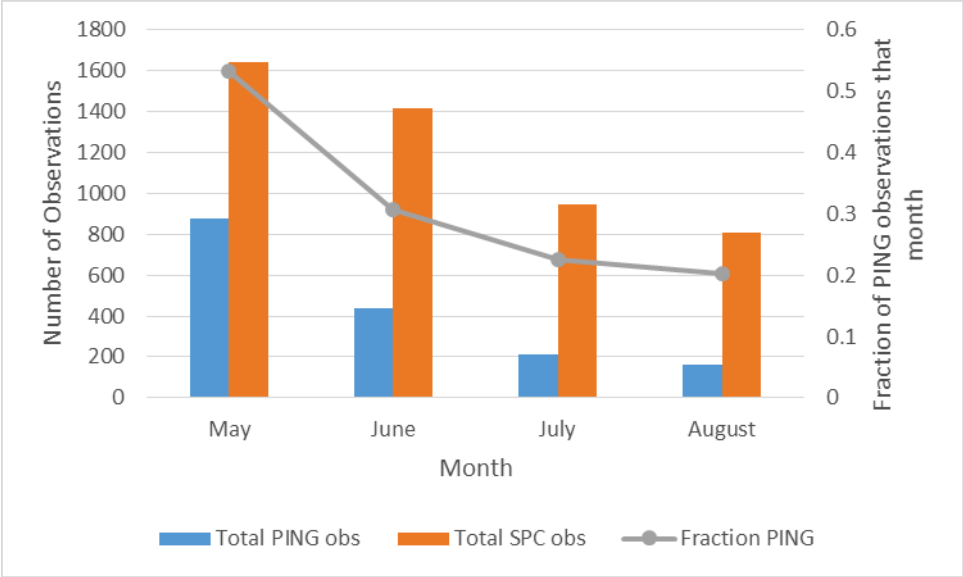


Figure 15: Number of PING and SPC observations by month, overlayed with fraction of PING observations by month.

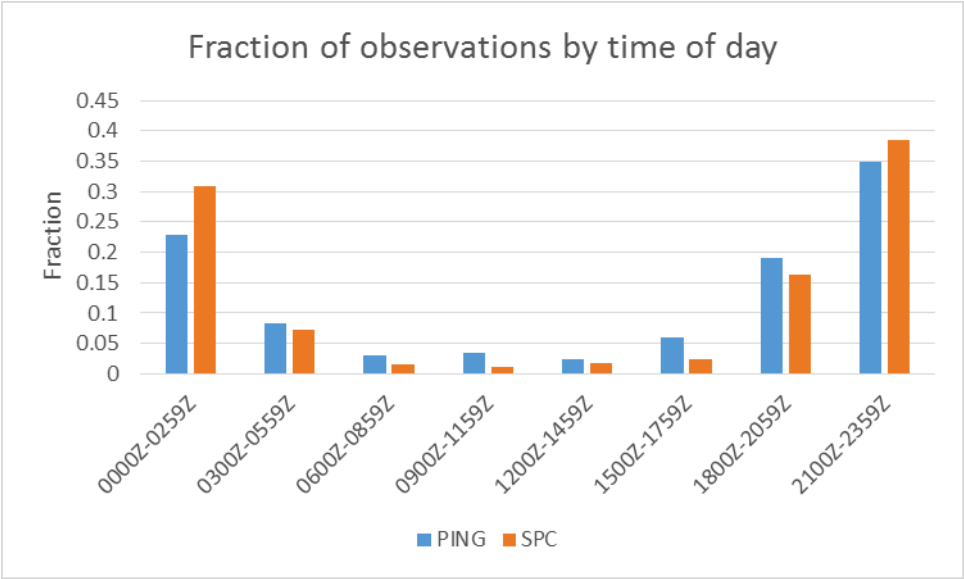


Figure 16: Fraction of reports in both the SPC and PING datasets that occurred at the corresponding times.

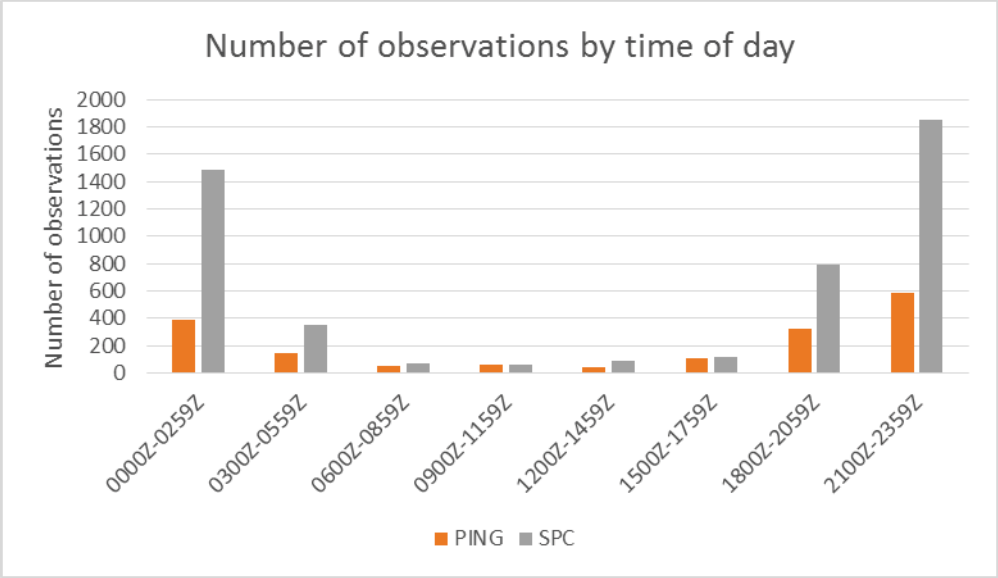


Figure 17: Number of observations in both the SPC and PING datasets that occurred at the corresponding times.

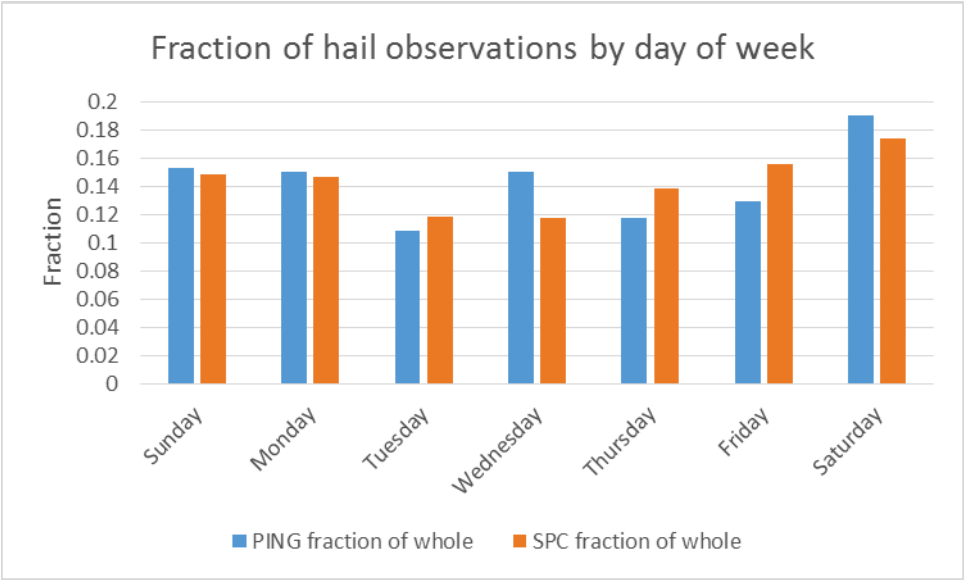


Figure 18: Fraction of reports in both the SPC and PING datasets that occurred on the corresponding days of the week.

3.2 Contingency Tables

Contingency tables are a common method of determining skill of forecasts. Davis and Ladue 2004 used the approach to analyze factors that bias warning confirmation, and a similar approach will be used here. In this section, spatial distribution of PING reports will be compared to the ground truth, SPC reports. A correct positive is defined as a PING report occurring within 10km of an SPC report, a false positive is defined as a PING report falling more than 10km from any SPC reports, and a false null is defined as an SPC report which does not have a PING report within 10km of it. The responses of the scores to different distances are also evaluated. From this data, we can compute the Probability of Detection, False Alarm Ratio, and Critical Success Index of PING hail observations in confirming SPC reports.

Because the number of SPC reports is much greater than the number of PING reports, a few conjectures about the scores can be made before they are computed. The number of false nulls can be expected to be very high, lowering the Critical Success Index and False Alarm Ratio. In the case where a theoretical perfect hail dataset is used, the Critical Success Index would be expected to approach zero, and the False Alarm Ratio can be expected to approach one, due to the large number of observations. The probability of detection can be expected to be low due to the demonstrated population bias in the PING dataset. If participation in the PING program were much greater, there would be more duplicate observations in the dataset (that is, multiple observations the same day, within 10km of one another). Two PING observations very close to each

other give more confidence that the observations are correct. Thus, clusters of observations are also compared to each other. The standard 10km distance is used, but a 5km, 15km, and 20km clustering distance is also examined. An increase in distance may be undesirable due to the decrease in resolution, and the approach of the limits of the meso-gamma scale. A decrease in clustering distance can be expected to result in fewer clusters, and thus fewer matches between points.

The results of the different runs are shown in figures 19-21. Skill scores do improve with increasing search radius, but remain low. One exception is the CSI and POD score for the severe clustered case. This decreases when the search radius goes from 10km to 15km, which is an artifact of the small number of severe PING clusters. From these results, it is clear that it is most accurate to compare individual PING points to the SPC dataset. While a cluster of PING reports may be expected to be more indicative of an accurate observation, considering only clusters comes at the cost of reducing the POD and CSI. The false alarm rate actually increases when considering only clusters. In the most basic way, this can be described as “Areas where several people report PING observations are not very likely to have several SPC observations in them”. It is possible that this is due to the quality control on SPC reports. Areas that receive many PING reports, that is, near population centers, may not receive nearly as many SPC reports due to removal of duplicates. The false alarm rate, high for the sets that contain non-severe reports, is remarkably low in the severe-only

clustered and non-clustered cases. This is due to the large number of SPC reports and clusters compared to the PING reports.

While the skill scores appear low, a few problems exist when viewing them without comparison. Because the number of PING reports is less than the number of SPC reports, PING cannot achieve a perfect POD or CSI. Also, it is not known what constitutes a “good” skill score. By taking random permutations of the SPC dataset, the same size as the PING dataset, and comparing it to the remaining SPC points, we can attempt to come to some kind of benchmark. Table 1 shows the results for 10 random selections of SPC points, compared to the remaining SPC points. Only the point-to-point, 10km search radius case is presented, because it was previously found to have a better score than the clustering method. The average POD of the random perturbations is .16, the FAR .63, and CSI .13. However, even if there were no false nulls, the POD and CSI would rise to only .43. Thus, the random subset of the SPC set attains a POD and CSI of 37% and 30% of optimal. Considering again the PING reports, the POD and CSI of .102 and .085 are 24% and 20% of optimal. While skill is lower in the PING point to point comparison, it still has sufficiently high value to have merit, compared to the SPC results.

The assumptions necessary for this kind of analysis limit the usefulness of these contingency tables. It is assumed that hail producing storms will always be well resolved by SPC reports. Due to the high spatial variability of hail reports, other hail may have fallen nearby, but not near places where the SPC hail was observed. While a nearby report of large hail by SPC would indicate a PING

report is valid, the high number of reports below .75 inches in the PING database means most PING reports were observed in areas where non-severe hail was falling (assuming the PING report was accurate). Given that SPC reports are mainly collected to confirm severe warnings, it is possible that areas away from SPC reports could contain smaller hail reports. An analysis that deals with the possibility of using warning polygons to confirm accuracy of PING reports is discussed in the next section.

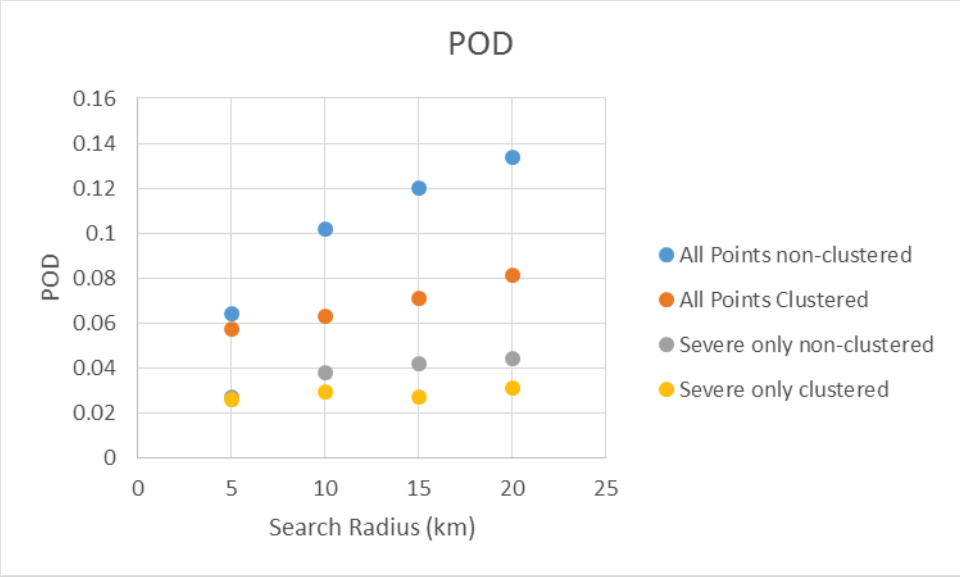


Figure 19: Probability of Detection for the four different verification methods, for all four search radii. From top to bottom: All PING reports compared to all SPC reports; clusters of PING reports compared to clusters of SPC reports; only severe PING reports compared to SPC reports, and only severe PING clusters compared to SPC clusters.

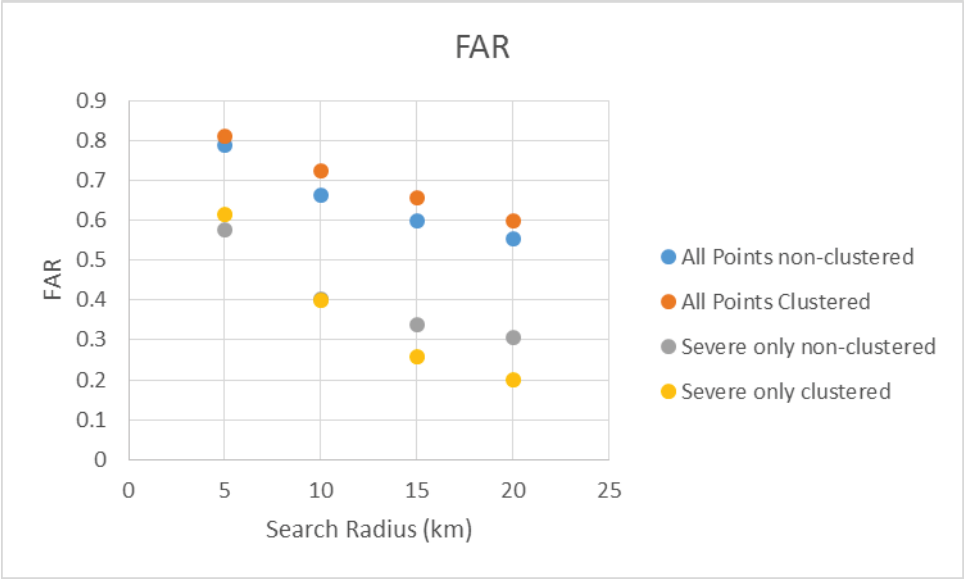


Figure 20: As in figure 18, but for the False Alarm Rate

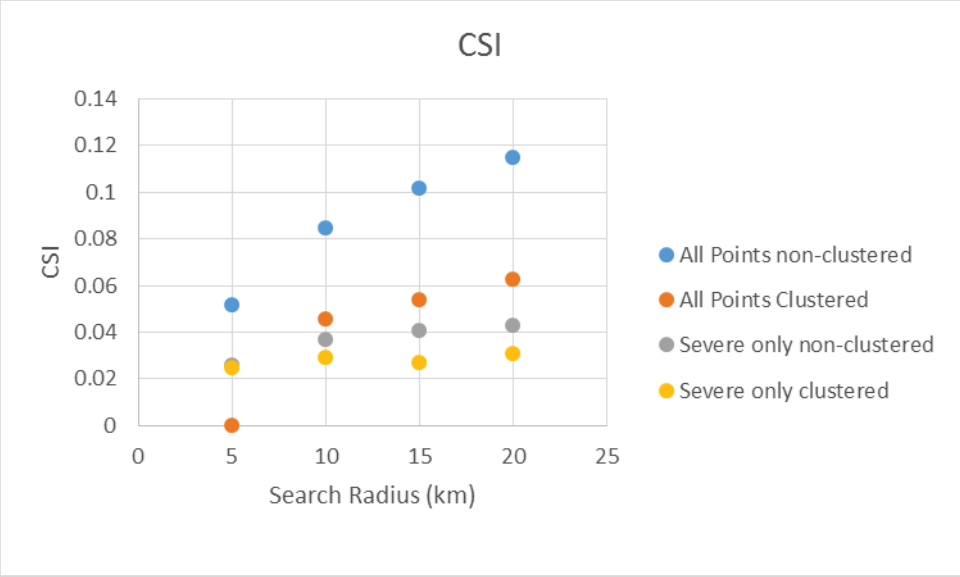


Figure 21: As in figure 17, but for the Critical Success Index.

3.3 Comparison to Severe Warnings

As stated in Doswell 2005, Point reports of hail represent only a small part of the area affected, limiting the effectiveness of using points for verification. False positives detected above may not all be false positives. In order to determine whether PING points that do not align with SPC points may be spatially accurate, an analysis of the isolated points that fall within warning polygons is presented. While a PING point may not be near an SPC point, if the PING point is not severe, and falls within a warning polygon, it is reasonable that non-severe hail could have fallen in that area. Likewise, PING points that have neither a matching SPC report, and that fall outside of a warning polygon, can reasonably expected to be false.

A correct positive is defined as a PING report that occurs within a severe polygon, a false positive is defined as a PING report that falls outside of a severe polygon, and a false null is defined as a severe polygon not containing a PING report. Only 81 PING reports without a corresponding SPC report fell inside of a warning polygon. Of the 959 total false positives from the point to point analysis, 878 can reasonably be expected to be truly inaccurate reports, giving a more reasonable estimate of the POD of .08, FAR of .61, and CSI of .07

In order to put those scores into perspective, benchmark skill scores must be determined for the polygon method of confirmation as well, in much the same manner as with point reports. Ten different random selections of the SPC point dataset, all the same size as the PING dataset, are compared to the polygons in order to determine the polygons' skill. SPC reports are primarily used to confirm

severe warnings, so a join of the SPC reports to the severe polygons can be expected to have high skill. However, by downsampling the SPC reports to only 1449 reports, it is obvious that not every one of the 7256 severe polygons can be confirmed. Even the size of the SPC hail dataset, at 4815 observations, cannot confirm every polygon, resulting in many false nulls.

A false null in the warning polygon test is a somewhat specious concept. It is defined here as an event where a severe polygon exists with no hail reports inside it. Because hail is not expected in every severe polygon, but only a certain fraction of severe polygons, a false null total of zero cannot be expected to be obtained, even if every storm in the polygons was closely observed on the ground. This is a limitation of the methodology used, but false nulls, when working with any incomplete hail datasets, are problematic. Comparing any hail dataset to a perfect theoretical dataset, the number of false nulls would be extremely high. Thus, the number of false nulls taken by itself cannot be considered very useful. However, if one assumes that a constant fraction of severe polygons are expected to contain hailfall, the false null numbers are useful in comparison to one another.

Table 2 shows the permutation tests performed on the polygons. Because SPC reports are primarily collected to confirm severe warnings, high skill can be expected to be demonstrated. Optimal values of POD and CSI are .20, and the average of the permutations show a POD and CSI .02 below that score. The false alarm rate is also low, at .09 compared to an optimal score of 0. This results in a POD and CSI of 90% of optimal for the average of the perturbations.

Returning to the PING dataset, the POD score is 39% of optimal, and CSI 35% of optimal, with an FAR of .91. These scores are better than those achieved using the point to point analysis, but much lower than the SPC scores. Such a result indicates that there may be high spatial inaccuracy in the PING dataset.

Figure 22 shows the total number of warning polygons issued for all Weather Forecast Offices. The highest values occur in the Rapid City, SD and Norman, OK forecast regions. The number of warning polygons per office are generally correlated with the number of SPC reports per office (Figure 23) in the central region, aside from an anomalous small number of issuances in the Sioux Falls, SD district. Such an anomaly may be expected to be caused by the warning issuing style of that particular office, and indeed, the number of SPC hail reports in that district is similar to those surrounding it. The high number of warnings issued in the Rapid City and North Platte warning areas are generally in line with the expected climatologies for those areas as described in Cintenio et al. 2012, and have a high number of confirmed SPC reports there.

The corresponding figures of PING reports (Figure 24) display the population bias of this report dataset. Anomalously high numbers of reports exist in the Minneapolis and Denver forecast areas. Figures 25a and 25b show the number of severe polygons that did not contain an SPC report, but did contain a PING report of any size. The Rapid City and Norman warning areas have a high number of reports confirmed in this manner. Not only are the Black Hills and Oklahoma climatologically high risk hail areas, these regions also have high population densities. Thus, these areas can be expected to have a high number

of reports. Other County Warning Areas contain similar rates of PING reports, excepting the Chicago and Cleveland areas. It was considered earlier that the large number of reports on the south side of Lake Erie was possibly due to aggressive report collecting on the part of the WFO. However, because the PING reports also show an anomalously high number of reports here, it is likely that the 2013 season had a genuinely high rate of hail producing storms in this area.

A clear population bias is less apparent when the County Warning Areas are examined. Figure 26 is a chart similar to figure 6, but contains only PING points which confirmed Severe Polygons that were unconfirmed by SPC reports. The population distribution of PING reports that confirm severe polygons is very similar to the population distribution of all PING reports. The similarity in results indicates that the population bias is no different when considering all PING reports or a subset of PING reports that are not near SPC reports. This indicates the correlation between the placements of SPC reports confirming polygons at the same time as PING reports is low. Thus, in polygons that do not contain a standard SPC hail report, PING reports can be used to confirm the polygon, but the population bias still remains high. The hail size distribution is also remarkably similar to the entire PING dataset, which means reported sizes are biased heavily towards the non-severe end.

Considering only severe PING reports that fall in polygons without an SPC report is the most practical use of the PING dataset. However, the number of such reports is very low, at only 42. With the very heavy bias of the PING reports

towards nonsevere, the use of the dataset is best for when non-severe sized hail is important.

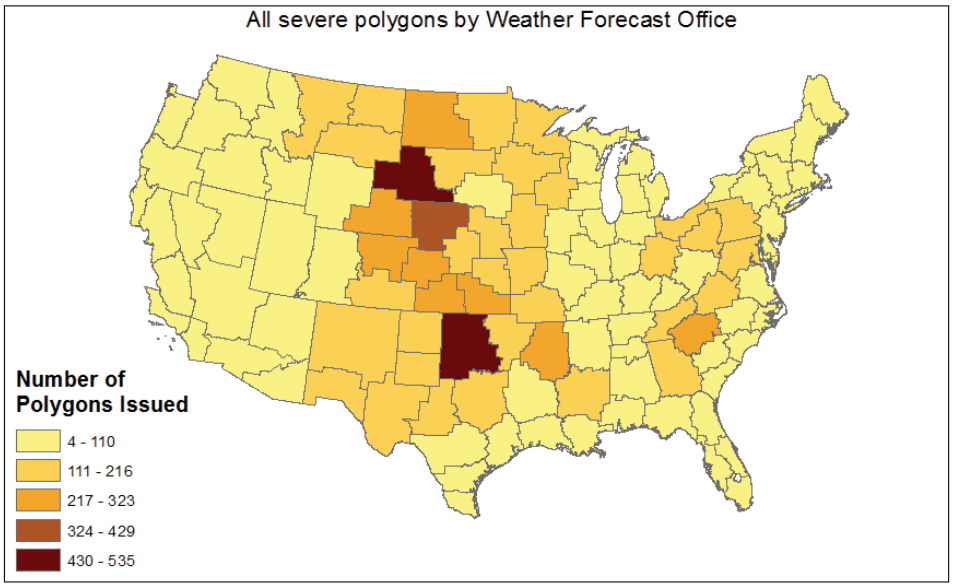


Figure 22: Number of warning polygons issued by each Weather Forecast Office from May 1 through August 31, 2013.

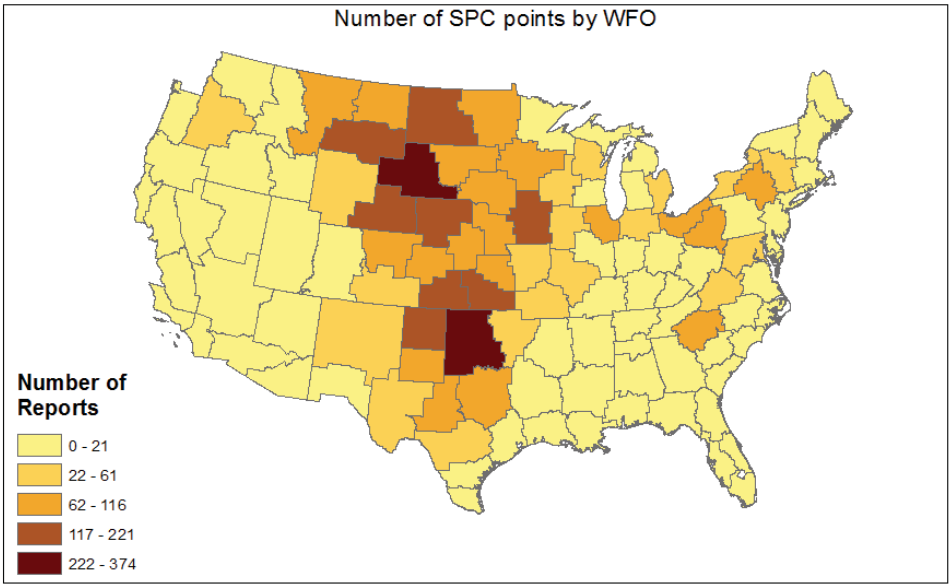


Figure 23: Number of SPC reports by Weather Forecast Office.

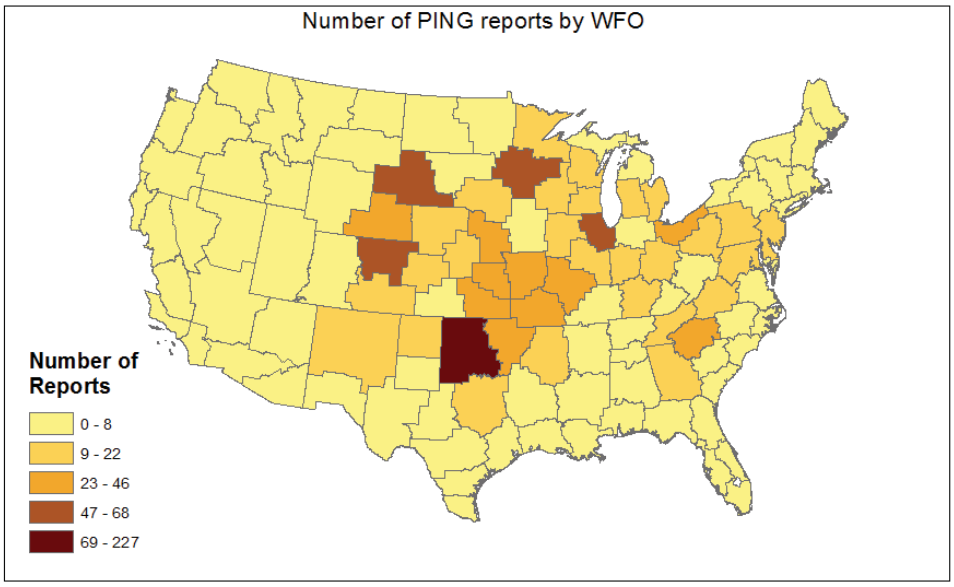
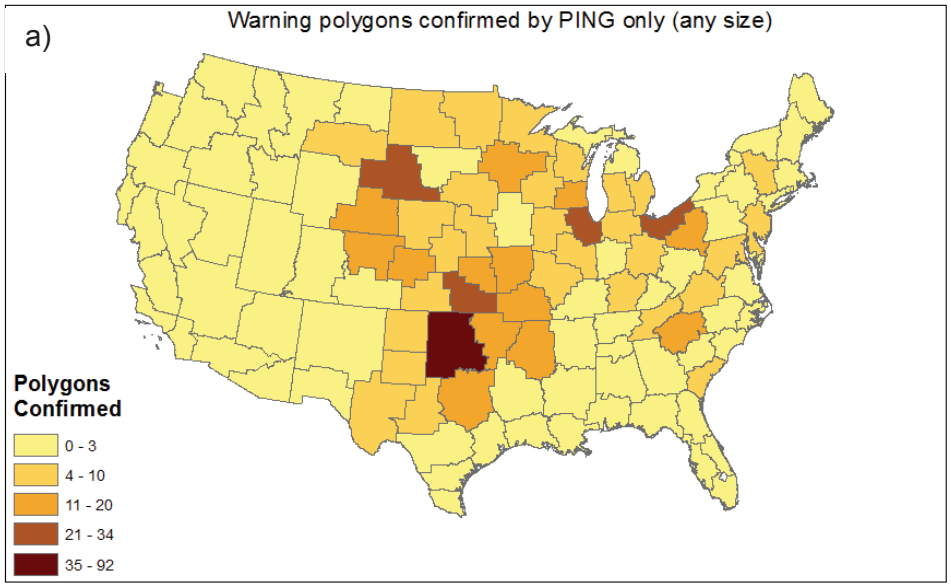
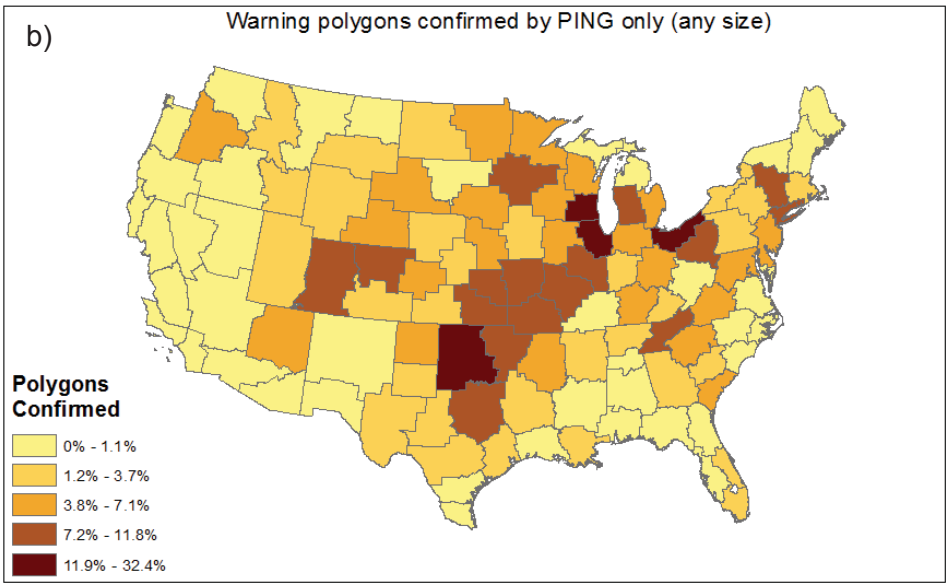


Figure 24: Number of PING reports by Weather Forecast Office.





Figures 25: a. Number of warning polygons confirmed by a PING report and not by an SPC report by Weather Forecast Office, and b. Percent of all warning polygons by office confirmed by a PING report and not by an SPC report.

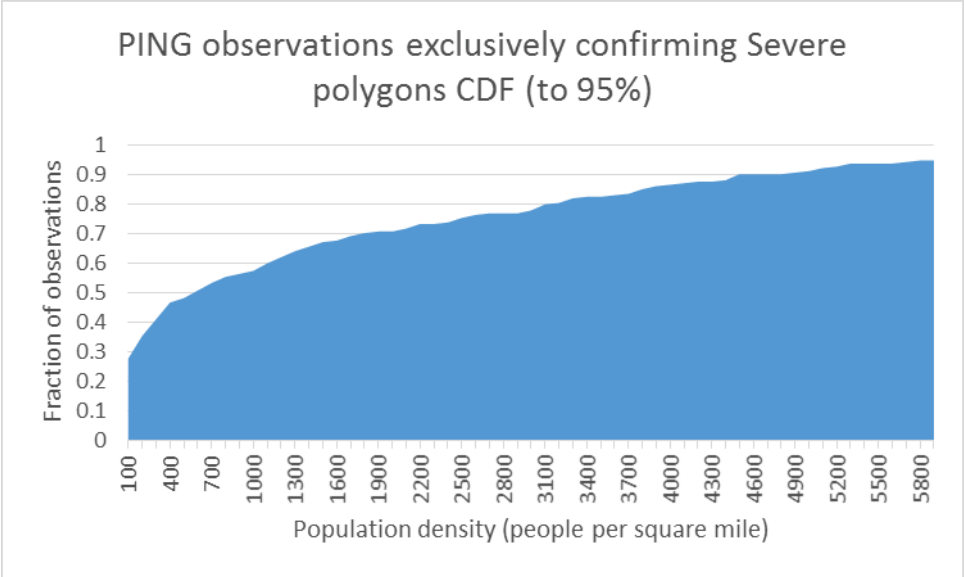


Figure 26: Population density of PING reports which confirmed a warning polygon without a corresponding SPC report, expressed in people per square mile.

4. Conclusion

While the number of PING reports by WFO is proportional to the number of SPC reports, this does not hold for every office, due to the strong population bias of PING reports. The number of warning polygons confirmed by severe PING reports, which would otherwise remain unconfirmed, is too low to be useful in most situations. Even if this number were high, the risk of using an incorrectly submitted PING report is great without another method to confirm the observation. A consensus of PING reports may be used if the density of PING reports is high enough, but the number of clusters found in this analysis is low.

As the National Weather Service continues to publicize the mPING app, and with increasing smartphone use, the rate of PING reporting will likely increase. With more data, more work can be done using the dataset. In particular, more severe PING reports must be collected in order for reasonable analysis of significant hail to be done. An analysis of the number and accuracy of PING reports in discrete population density areas, compared to another dataset, would help to isolate non-population density and non-meteorological factors. Large amounts of data would enable different regions to be analyzed separately (as in Cintineo et al. 2012), and allow the domain to be constrained for more thorough analysis.

In its current form, the PING dataset could be used to create a high resolution hail dataset in some heavily populated areas. The greatest utility may still lie in the unique method of data collection this dataset utilizes, which allows

examination of sociological factors affecting other datasets. However, a rigorous quality control method must be developed to eliminate false PING reports.

References

Amburn, S. A., and P. L. Wolf, 1997: VIL Density as a Hail Indicator. *Wea. Forecasting,* **12,** 473-478.

Blair, S. F., and J. W. Leighton, 2012: Creating high-resolution hail datasets using social media and post-storm ground surveys. *Electronic J. Operational Meteor.*, **13** (3), 32-45.

Brimelow, J.C., and G.W. Reuter, 2002: Modeling Maximum Hail Size in Alberta Thunderstorms. *Wea. Forecasting,* **17,** 1048-1062

Changnon, S. A., 1999: Data and Approaches for Determining Hail Risk in the Contiguous United States. *J. Appl. Meteor. Climatol.,* **68,** 1730-1739.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An Objective High-Resolution Hail Climatology of the Contiguous United States. *Wea. Forecasting,* **27**, 1235-1248.

Davis, S. M., and J. LaDue, 2004: Nonmeteorological factors in warning verification. Preprints, *22nd Conf. on Severe Local Storms, Hyannis,* MA, Amer. Meteor. Soc., P2.7. [Available online at http://ams.confex.com/ams/pdfpapers/81766.pdf.]

Donavon, R. A., and K. A. Jungbluth, 2007: Evaluation of a Technique for Radar Identification of Large Hail across the Upper Midwest and Central Plains of the United States. *Wea. Forecasting,* **22,** 244-254.

Doswell, C. A. III, H.E. Brooks, and M.P. Kay, 2005: Climatological Estimates of Daily Local Nontornadic Severe Thunderstorm Probability for the United States. *Wea. Forecasting,* **20**, 577-595.

Edwards R., and R.L. Thompson, 1998: Nationwide Comparisons of Hail Size with WSR-88D Vertically Integrated Liquid Water and Derived Thermodynamic Sounding Data. *Wea. Forecasting,* **13**, 277-285.

Elsner, J. B., L. E. Michaels, K. N. Scheitlin, I. J. Elsner, 2013: The Decreasing Population Bias in Tornado Reports across the Central Plains. *Wea. Climate Soc.,* **5,** 221-232.

Greene, D. R., and R. A. Clark, 1972: Vertically integrated liquid water—A new analysis tool. Mon. Wea. Rev., 100, 548–552.

Hales, J. E., Jr., 1993: Biases in the severe thunderstorm data base: Ramifications and solutions. Preprints, *13th Conf. on Weather Analysis and Forecasting, Vienna,* VA, Amer. Meteor. Soc., 504–507.

Kessinger, C. J., and E. A. Brandes, 1995: A comparison of hail detection algorithms. Final Rep. to the FAA, 52 pp. [Available from NCAR, P.O. Box 3000, Boulder, CO 80307.]

King, P., 1997: On the Absence of Population Bias in the Tornado Climatology of Southwestern Ontario. *Wea. Forecasting,* **12,** 939-946.

National Climatic Data Center, cited 2013: Storm Data FAQ Page. [Available online at http://www.ncdc.noaa.gov/stormevents/faq.jsp.]

National Severe Storms Laboratory, cited 2013: mPING Project:FAQ. [Available online at http://www.nssl.noaa.gov/projects/ping/faq.php.]

National Weather Service: Historical Shapefiled for Events and polygon Warnings, 2013 Warnings through August, September 6, 2013, http://www.nws.noaa.gov/regsci/gis/historical_shapefiles/

National Weather Service: National Weather Service Warning Area Shapefile, October 28, 2013, http://www.weather.gov/geodata/catalog/wsom/html/cwa.htm.

National Weather Service Warning Decision Training Branch, cited 2013: Dual-Pol Applications: Hail Detection. [Available online at http://www.wdtb.noaa.gov/courses/dualpol/Applications/Hail/player.html.

Ortega, K. L., T. M. Smith, K. L. Manross, K. A. Scharfenberg, A. Witt, A. G. Kolodziej, and J. J. Gourley, 2009: The Severe Hazards Analysis and Verification Experiment. *Bull. Amer. Meteor. Soc.,* **90** (10), 1519-1530.

Park, H., A. V. Ryzhkov, D. S. Zrnic, K. E. Kim, 2008: The Hydrometeor Classification Algorithm for the Polarimetric WSR-88D: Description and Application to an MCS. Wea. Forecasting, **24** (3), 730-748.

Paxton, C. H., and J. M. Shepherd, 1993: Radar diagnostic parameters as indicators of severe weather in central Florida. NOAA Tech. Memo. NWS SR-149, 12 pp. [Available from the National Weather Service Southern Region Headquarters, 819 Taylor Street, Room 10A26, Fort Worth, TX 76102.]

Richter, H. and R. B. Deslandes, 2007: The four large hail assessment techniques in severe thunderstorm warning operations in Australia. *33rd Conference on Radar Meteorology, Cairns, Australia, Amer. Meteor. Soc. and Australian Bureau of Meteorology Research Center.* [Available online at https://ams.confex.com/ams/pdfpapers/123766.pdf.]

Rogers, R. R., Yau, M. K., 1989: *A Short Course in Cloud Physics*. 3rd ed. Butterworth-Heinemann, 290 pp.

Schaefer, J. T., J. J. Levit, S. J. Weiss, and D. W. McCarthy, 2004: The frequency of large hail over the contiguous United States. Preprints, *14th Conf. Applied Climatology, Seattle,* WA, Amer. Meteor. Soc., 3.3. [Available online at http://ams.confex.com/ams/pdfpapers/69834.pdf.]

Simmons, K. M., D. Sutter, 2007: Tornado Warnings, Lead Times, and Tornado Casualties: An Empirical Investigation. Wea. Forecasting, **23**, 246-258.
Straka, J. M., D. S. Zrnic, and A. V. Ryzhkov, 2000: Bulk Hydrometeor Classification and Quantification Using Polarimetric Radar Data: Synthesis of Relations. *J. Appl. Meteor. Climatol.,* **39,** 1341-1372

Strong, G.S., and E. P. Lozowski, 1977: An Alberta study to objectively measure hailfall intensity. *Atmos.–Ocean*, **15**, 33–53.

Trapp, Robert J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and Ray Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting,* **21,** 408-415.

ESRI: USA Block Group Boundaries, September 30, 2013, http://www.arcgis.com/home/item.html?id=1c924a53319a491ab43d5cb1d55d8561

Wilson, C. J., K. L. Ortega, and V. Lakshmanan, 2009: Evaluating multi-radar, multi-sensor hail diagnosis with high resolution hail reports. Preprints, *25th Conf. on Interactive Information Processing Systems, Phoenix,* AZ, Amer. Meteor. Soc., P2.9. [Available online at http://ams.confex.com/ams/pdfpapers/146206.pdf.]

Witt, A., M. D. Eilts, G. J. Stumpf, J.T. Johnson, E. D. Mitchell, and K. W. Thomas, 1998a: An Enhanced Hail Detection Algorithm for the WSR-88D. *Wea. Forecasting,* **13,** 286-303.

Witt, A., M. D. Eilts, G. J. Stumpf, E. D. Mitchell, J.T. Johnson, and K.W. Thomas, 1998b: Evaluating the Performance of WSR-88D Severe Storm Detection Algorithms. *Wea. Forecasting,* **13,** 513-518.

Zrnic, D. S., V. N. Bringi, N. Balakrishnan, K. Aydin, V. Chandrasekar, and J. Hubbert, 1993: Polarimetric measurements in a severe hailstorm. *Mon. Wea. Rev.*, **121**, 2223-2238.

Appendix

Permutation number	Number of correct positives	False Positives	False Nulls	POD	FAR	CSI
1	540	909	2826	0.160	0.627	0.126
2	580	869	2786	0.172	0.600	0.137
3	540	909	2826	0.160	0.627	0.126
4	544	905	2822	0.162	0.625	0.127
5	568	881	2798	0.169	0.608	0.134
6	533	916	2833	0.158	0.632	0.124
7	512	937	2854	0.152	0.647	0.119
8	542	907	2824	0.161	0.626	0.127
9	550	899	2816	0.163	0.620	0.129
10	522	927	2844	0.155	0.640	0.122
Average	543.1	905.9	2822.9	0.161	0.625	0.127
Max	580	937	2854	0.172	0.647	0.137
Min	512	869	2786	0.152	0.600	0.119
Best theoretically possible	1449	0	1917	0.430	0.000	0.430

Table 1: Skill scores for 10 downsamplings of the SPC observations to 1449 observations, compared to the rest of the SPC dataset.

Permutation number	Number of correct positives	False Positives	False Nulls	POD	FAR	CSI
1	540	909	2826	0.160	0.627	0.126
2	580	869	2786	0.172	0.600	0.137
3	540	909	2826	0.160	0.627	0.126
4	544	905	2822	0.162	0.625	0.127
5	568	881	2798	0.169	0.608	0.134
6	533	916	2833	0.158	0.632	0.124
7	512	937	2854	0.152	0.647	0.119
8	542	907	2824	0.161	0.626	0.127
9	550	899	2816	0.163	0.620	0.129
10	522	927	2844	0.155	0.640	0.122
Average	543.1	905.9	2822.9	0.161	0.625	0.127
Max	580	937	2854	0.172	0.647	0.137
Min	512	869	2786	0.152	0.600	0.119
Best theoretically possible	1449	0	1917	0.430	0.000	0.430

Table 2: Skill scores for 10 downsamplings of the SPC observations to 1449 observations, compared to the 7256 warning polygons.