MAXIMUM HAIL SIZE PREDICTION

Evangelos Tsagalidis[1*], Kyriakos Tsitouridis[1], Georgios Evangelidis[2], and Dimitris Dervos[3]

(1) Hellenic Agricultural Insurance Organization, Meteorological Applications Centre, International Airport Makedonia, 55103 Thessaloniki, GREECE, {e.tsagalidis, k.tsitouridis}@elga.gr
2) Department of Applied Informatics, University of Macedonia, 156 Egnatia Street, 54006, Thessaloniki, GREECE, gevan@uom.gr
3) Information Technology Dept., T.E.I., P.O. BOX 141, 57400 Sindos, GREECE, dad@it.teithe.gr

ABSTRACT. We examine the possibility of building a meteorological prediction tool using data from the Greek National Hail Suppression Program. More specifically, we focus on maximum hail size prediction from operational meteorological radar and/or sounding data. Factor analysis and linear regression are applied in order to identify the optimum number of independent variables and the sequence to build the corresponding meteorological tool. A significant linear relationship is discovered for non-seeded storms relating hail size to various radar parameters, such as the Reflectivity or the group of Vertically integrated liquid density and Cloud top. A relationship for predicting hail size using radar parameters for seeded storms failed to be statistically significant.

1. INTRODUCTION

The Hellenic Agricultural Insurance Organization (ELGA) is a public organization and the main insurance carrier of the agricultural production in Greece. The Meteorological Applications Centre (KEME) is the section of ELGA that since 1981 conducts the Greek National Hail Suppression Program (GNHSP) using airborne seeding. The Program aims at reducing insurance payments due to hail damage and is being applied in Central Macedonia and Thessaly, covering an area of 5,000 square kilometres, during the April to September period. The cloud seeding is performed by three aircraft releasing AgI in developing hail-bearing clouds as indicated by radar (Tzoumaki *et al.* 2006).

In this study, we explore the possibility of building a prediction tool using the GNHSP data. More specifically, we focus on maximum hail size estimation and prediction from operational meteorological radar and/or sounding data. We apply factor analysis in a pre-processing phase to identify the optimum number of independent variables, and, subsequently, linear regression to build a simple, yet effective, meteorological tool. A meteorologist can easily use the tool to quickly map radar and atmospheric measurements to possible hail size on the ground.

**Corresponding author address:* Evangelos Tsagalidis, Hellenic Agricultural Insurance Organization, Meteorological Applications Centre, International Airport Makedonia, 55103 Thessaloniki, Greece; e-mail: e.tsagalidis@elga.gr

The remainder of the paper is organized as follows: Section 2 describes the dataset we used to build the prediction tool, and Section 3 presents the adopted methodology. In Section 4, we present the results we obtained by experimenting with the chosen techniques, and, finally, Section 5 concludes the paper.

2. DATASET

The analysis utilizes radar data of the EEC S-band meteorological radar installed at Airport "Macedonia" of Thessaloniki. Data recorded by the Thunderstorm Identification, Tracking, Analysis and Nowcasting system (TITAN) (Dixon and Wiener 1993), are further analyzed to create a sample of Storm Cell Complexes (SCC) which is actually a structured form of the initial data and represents the storm characteristics data (Tsagalidis and Tsitouridis 2000, Tsagalidis et al. 2006). The data were recorded during the storm activity from April to September 1999, 2000, 2001 and 2005 in the protected area of Central Macedonia. The SCC structured data represent the values of each hailstorm attribute, and, more specifically, the type, reflectivity, cloud top, vertically integrated liquid water content (VIL) and vertically integrated liquid water density (VIL density).

The structure of cloud systems and their classification in different categories follows the classification of SCC (Tsagalidis and Tsitouridis 2000, Tsagalidis et al. 2006). The classes are represented by the values of the Type attribute, where "S" is used for unicellular storms of a single ordinary cell, "SU" for Unicellular storms of a supercell, "M" for multicell storms, and, "L" for line storms. During the entire lifetime of the SCC, reflectivity in dBz is the maximum radar reflectivity at the -50ºC level or higher and the cloud top in km is the maximum height. The VIL in $kg \cdot m^{-2}$ is the integration from the echo base to the echo top of the liquid water content and is estimated using a mathematical function between liquid water content and radar reflectivity (Greene and Clark 1972). Our VIL parameter is the maximum value recorded during the entire lifetime of the SCC. The VIL Density is simply the VIL divided by the echo top (m) and multiplied by 1000 in order to express the result in $g \cdot m^{-3}$ (Amburn and Wolf 1997). In Table 1, we show the mean, standard deviation, minimum and maximum values of Reflectivity, Cloud top, VIL and VIL Density in our data.

Table 1: Mean, standard deviation, minimum and maximum values of Reflectivity, Cloud top, VIL and VIL Density.

	Refl. (dBz)	C. top (km)	VIL ($kg \cdot m^{-2}$)	VIL Density ($g \cdot m^{-3}$)
Mean	52.2	10.4	22	2.2
St.dev.	5.3	1.4	12.1	1.1
Min.	40	7	3.8	0.5
Max.	69	14.5	55.8	5.1

Furthermore, during the analysis, meteorological parameters are examined from the sounding data of the Upper Air Observation Station of Thessaloniki, which relate to the hail size on the ground too. The meteorological station is located close to the project area of Central Macedonia and the calculated values of the atmospheric parameters, such as wet bulb zero (WBZ) and mean temperature are associated with the SCC environment. The WBZ is the height in km of the wet bulb temperature 0ºC level, corresponding to the melting level of the hailstone during its fall to the ground, whereas, the mean temperature in Kelvin of the layer between that level and the ground is the mean temperature. These parameters were calculated using the most representative sounding related to the occurrence time of each SCC. In Table 2 we show the mean, standard deviation, minimum and maximum values of WBZ and Mean temperature in our data.

Table 2: Mean, standard deviation, minimum and maximum values of WBZ and Mean temperature.

	WBZ (m)	*Mean temp*. (K)
Mean	3123	288
St.dev.	419	2.1
Min.	2118	282
Max.	4182	293

The WBZ values associated with hail days are bounded in a specific range of values, because low WBZ values imply stable air conditions, not sufficient for hailstorms, and high values an increasing possibility that the hailstones will melt before reaching the ground (Tsagalidis 1996). During the preprocessing phase, we made the appropriate transformations of the WBZ values to '1', '2' and '3' values using the method of Z-score normalization. The '2' value corresponds to WBZ Z-scores between -1 and 1, the '1' to less than -1 and the '3' to greater than 1. Similarly, the mean temperature values have been transformed to '1', '2' and '3' values, where the '1' value represents a relatively cold air layer and the '3' value a relatively warm air layer.

Each SCC is identified as a hailstorm using the data from the GNHSP hailpad network. These data include values of maximum hail diameter in mm, called hailsize, for each one hailstorm (SCC). In our sample, it was not the case that two or more SCC passed over a hailpad before changing the hailpad the next day with a new one.

In addition, due to potential seeding effect on hail size during the GNHSP operation, the insertion of the SCC seed attribute has been considered as crucial. Analyzing the operational radar data for each SCC, the value 'yes' or 'no' was given to the seed attribute. The 'yes' value represents an acceptable SCC seeding operation according to GNHSP seeding criteria, and the 'no' value a non-acceptable seeding operation or the case of a non-seeded SCC. During the preprocessing phase, we used the values '1' and '0' in the place of 'yes' and 'no' respectively. Examples of non-acceptable seeding operations were delayed or corrupted seeding, or the seeding far away from the targets, and, in general, cases where the experienced meteorologist analyst believed that there was not a seeding effect in a particular SCC.

The values of the above parameters belonging to the groups of radar, sounding, seeding and hailpad network data comprise for each one SCC one record. We obtained 74 records for the 74 SCCs that were identified on radar and had hail records on the hailpad network. According to the ‘yes’ and ‘no’ values of the seed variable our dataset is split into two subsets having 32 and 42 records respectively.

3. METHODOLOGY

A lot of research work deals with the problems of detecting hail or estimating the probability of hail and the hail size within the cloud or on the ground (Waldvogel et al. 1979, Witt et al. 1998, Foote et al. 2005, Auer 1994, Greene and Clark 1972, Amburn and Wolf 1997). In Waldvogel et al. (1979) the authors relate the height difference between the top of the 45 dBZ echo in the storm and the freezing level to the probability of hail, with S-band radar returns validated against a surface hailpad network. Witt et al. (1998) present an enhanced hail detection algorithm, which estimates the probability of hail (any size), probability of severe-size hail (diameter ≥ 19mm), and maximum expected hail size for each detected storm cell, and in addition the severe hail index (SHI) which is the primary predictor variable for severe-size hail. Foote et al. (2005) discuss the sensitivity and variation with time of several radar hail parameters computed using the TITAN system, including probability of hail, hail mass aloft, vertical integrated hail mass, hail kinetic energy flux, and the FOKR index. The FOKR index (Foote-Krauss) is a hail storm classification system that uses the maximum reflectivity in the storm and the difference between the height of the top of the 45 dBZ echo in the storm and the height of the 0ºC isotherm. Auer (1994) describes a technique whereby the radar reflectivity can be combined with cloud-top temperature, from either satellite imagery and/or sounding analysis, to provide a reliable discrimination between heavy rain and/or hail in convective clouds. In addition, hail sizing is also possible. Greene and Clark (1972) introduced the VIL, whereas, Amburn and Wolf (1997) propose VIL density as a useful indicator for assessing hail potential in thunderstorms.

The aim of this study is the estimation and prediction of maximum hail size associated with a SCC, using radar or/and sounding parameters. Our dataset has tuples consisting of the variables type, reflectivity, cloud top, VIL, VIL density, WBZ, mean temperature, seed and the values of the observed hailsize.

The problem of predicting hail size or the corresponding hail size classes from our database is a typical classification problem. In Tsagalidis et al. (2008), we attempted to predict the hail size class using supervised classification techniques, such as the decision tree-based algorithm of C4.5, a widely-used decision tree algorithm for classification in the field of data mining, and the Bayes classifier. In this study, we chose statistical linear regression to perform classification in our dataset having as dependent variable the numerical hail size (Hailsize).

Statistical regression is a supervised technique that generalizes a set of numeric data by creating a mathematical equation relating one or more input attributes to a single output attribute. A linear regression equation is of the form:

$$f(x_1,x_2,\dots,x_n) = b_1x_1 + b_2x_2 + \cdots + b_nx_n + c$$

where $x_1,x_2,\dots,x_n$ are independent variables and $b_1,b_2,\dots,b_n$ and c are constants. $f(x_1,x_2,\dots,x_n)$ represents the dependent variable. In general, linear regression is appropriate when the relationship between the dependent and independent variables is nearly linear.

4. ANALYSIS AND RESULTS

4.1 Data reduction

In a step prior to linear regression and in order to achieve data reduction by identifying representative variables from our dataset, we apply factor analysis (Hair et al. 2005). Factor analysis provides insight into the interrelationships among variables and the underlying structure of the data and is an excellent starting point for many other multivariate techniques, such as regression. For data reduction, we use the SPSS statistical software (SPSS) and we chose the principal component analysis extraction method with eigenvalues greater than 1 and the Varimax rotation method to construct a solution.

Table 3 shows the rotated component matrix that helps to determine what the 3 extracted components represent. The boldface cells show the significant loadings greater than the absolute value of 0.65. This threshold is chosen due to the dataset size consisting of 74 observations (Hair et al. 2005). We remark on the absence of cross-loadings and that the first component is most highly correlated to VIL, VIL density and reflectivity (group of SCC intensity variables), the second component to WBZ and mean temperature (group of atmospheric variables) and the third component to seed (operational variable). Additionally, the cloud top (SCC attribute) is correlated to both the first and third components in a remarkable level of 0.59 and 0.47 respectively.

Table 3: Factor analysis, rotated component matrix

	Component			*Comm.*
	1	2	3	
Type	0.57	-0.26	0.41	0.55
Cloud Top	0.59	0.28	0.47	0.65
Reflectivity	**0.84**	0.05	0.13	0.73
VIL	**0.91**	0.22	-0.13	0.89
VIL Density	**0.88**	0.10	-0.23	0.85
WBZ	0.07	**0.83**	-0.15	0.71
Mean temp.	0.14	**0.81**	0.30	0.76
Seed	0.10	-0.05	**-0.75**	0.57

The communalities, the estimation of the variance in each variable accounted for by the components, are all high, having values between 0.55 and 0.89. This indicates that the extracted components represent the variables well. Finally, we mention that in this procedure we used the values '1', '2' and '3' for both the WBZ and mean temperature variables. In addition, we experimented using their numerical values and we had the same results.

4.2 Hail size prediction

The factor analysis outcome during the data reduction procedure shed light on the dataset and highlighted the predictors. Interpreting the third component, where the operational variable *seed* is designated and taking into account the possible effect of seeding operations to hail size on the ground, we divide our sample based on the value 'yes' or 'no' of that variable. Similarly, the first component dictates the use of predictors from the group of SCC intensity variables, such as the *VIL Density*, *VIL* and *reflectivity*, whereas the second component is based on the group of atmospheric variables, such as *WBZ* and *mean temperature*.

The SPSS software package was used to apply linear regression on our dataset, where the dependent variable was hailsize (Hair et al. 2005, SPSS, Neter et al. 1996). Choosing different predictors, many trials were made in order to build an effective model, at least in the case of the non-seeded SCC. The results showed that the atmospheric variables *WBZ* and *mean temperature* do not contribute to the prediction. On the contrary, the *reflectivity* or the *VIL density*, especially when combining with cloud top, can give an acceptable model. The exploitation of *WBZ* or *mean temperature* variables that express the second component could be accomplished by furthermore dividing our already split dataset according to their values. Taking into account the small size of our dataset we prefer not to consider these atmospheric variables.

In the following, we use the notation regression (a) and regression (b) to refer to the two modes of the regression application, where in the first case the predictor is the *reflectivity* and in the second one the *VIL density* and *cloud top*. The assumptions of the linear regression model were checked and we can accept them as valid in the cases of the non-seeded hailstorms. Tables 4 and 5 show the corresponding coefficients of the regression lines.

Table 4: Coefficients of regression (a)

	B	S.Error	Beta	t	Sig.
Seed=no					
Const. Refl.	-24.93 0.738	5.46 0.10	 0.75	-4.57 7.07	0.00 0.00
Seed=yes					
Const. Refl.	2.35 0.206	8.72 0.17	 0.22	0.27 1.25	0.79 0.22

Table 5: Coefficients of regression (b)

	B	S.Error	Beta	t	Sig.
Seed=no					
Const.	-6.56	4.46		-1.47	0.15
VIL D.	1.876	0.57	0.41	3.29	0.002
C. top	1.53	0.43	0.43	3.49	0.001
Seed=yes					
Const.	-2.44	6.73		-0.363	0.719
VIL D.	0.273	0.86	0.06	0.316	0.754
C. top	1.463	0.71	0.39	2.075	0.047

In the case of non-seeded hailstorms, the hail size in mm is equal to

$$0.738 * Reflectivity - 24.93 \qquad \text{Eq. (1)}$$

or

$$1.876 * VILDensity + 1.53 * Cloudtop - 6.56 \qquad \text{Eq. (2)}$$

For seeded hailstorms neither model is reasonable.

In Tables 6 and 7, we show the ANOVA for testing the acceptability of the models from a statistical perspective.

Table 6: ANOVA of Regression line (a)

	Sum of Squares	df	Mean Square	F	Sig.
Seed=no					
Regression	553.7	1	553.7	50.02	0.00
Residual	442.8	40	11.1		
Total	996.5	41			
Seed=yes					
Regression	42.7	1	42.7	1.55	0.22
Residual	823.5	30	27.5		
Total	866.2	31			

Table 7: ANOVA of Regression line (b)

	Sum of Squares	df	Mean Square	F	Sig.
Seed=no					
Regression	444	2	222	15.67	0.00
Residual	552.5	39	14.2		
Total	996.5	41			
Seed=yes					
Regression	147.4	2	73.7	2.97	0.067
Residual	718.8	29	24.8		
Total	866.2	31			

We notice that only for the case of non-seeded hailstorms the significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance. Also, in the case of seeded hailstorms, the values of the sum of squares of the regression are low compared to the total.

In Table 8, we show the multiple correlation coefficient R, the coefficient of determination R square, the adjusted R square and the standard error of the estimate for the two regression application modes.

Table 8: Regression coefficients

	R	R Square	Adj. R Sq	S.Error
Regression (a)				
Seed=no	0.745	0.556	0.545	3.33
Seed=yes	0.222	0.049	0.018	5.24
Regression (b)				
Seed=no	0.668	0.446	0.417	3.76
Seed=yes	0.413	0.17	0.113	4.98

The values of the multiple correlation coefficient R for the cases of non seeded hailstorms are relatively large, especially for regression (a) (0.745). This indicates a strong relationship between the observed and model-predicted values of the *Hailsize*. The mean value of *hailsize* is equal to 13.5mm and the standard deviation to 4.93mm. Comparing the latter with the standard error of the estimate we observe that it is greater only in the cases of the non seeded hailstorms.

5. CONCLUSION

The present study refers to the specific system of GNHSP, where we examine the prediction of hail size that is associated with a SCC (hailstorm), using radar and /or sounding parameters. The factor analysis and the classification technique of statistical linear regression are used to build a meteorological prediction tool.

Applying linear regression, only in the case of non-seeded SCC (hailstorms) is there a strong relationship using as independent variables the *reflectivity* or the group of *VIL density* and *cloud top*. Hailsize from the non-seeded hailstorms shows a positive linear relationship with *reflectivity* or *VIL* or *VIL density*, whereas this relationship is violated in the case of seeded hailstorms.

Our study is a first attempt to build a meteorological prediction tool and it is limited by the hailstorm sample size. In the future, the examination of additional cases will improve the proposed tool as a robust element for decision-making during the GNHSP seeding operations.

6. REFERENCES

Amburn, S.A. and P.L. Wolf, 1997: VIL density as a hail indicator. *Wea. Forecasting*, **12**, 473-478.

Auer, A.H., 1994: Hail recognition through the combined use of radar reflectivity and cloud-top temperatures. *Mon. Wea. Rev.*, **122**, 2218–2221.

Dixon, M. and G. Wiener, 1993: TITAN: thunderstorm identification, tracking, analysis, and nowcasting – a radar based methodology. *J. Atmos. Ocean. Technol.*, **10 (**6), 785-797.

Foote, G.B., T.W. Krauss, and V. Makitov, 2005: Hail metrics using conventional radar. *Proc., 16th Conference on Planned and Inadvertent Weather Modification,* Amer. Meteor. Soc., Boston.

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

Hair, J.F., Jr., W.C. Black, B.J. Babin, R.E. Anderson, and R.L. Tatham, 2005: *Multivariate Data Analysis*, 6th Edition, Pearson-Prentice Hall, New Jersey, 928 pgs.

Neter, J., M.H. Kutner, C.J. Nachtsheim, and W. Wasserman, 1996: *Applied Linear Regression Models*, Irwin.

SPSS. http://www.spss.com. Statistical Analysis Software.

Tsagalidis, E., 1996: Ύψος της μηδενικής θερμοκρασίας του υγρού θερμομέτρου (WBZ) και η συσχέτισή του με την εμφάνιση ή μη χαλαζιού στο έδαφος κατά την περίοδο Απριλίου-Σεπτεμβρίου 1993" (The height of the wet bulb zero temperature (WBZ) and its relationship with hail on the ground during the period April-September 1993), 3rd Hellenic Conference of Meteorology, Climatology and Atmospheric Physics, Athens, 117-122, (Presentation).

Tsagalidis, E. and K. Tsitouridis, 2000: Συμπλέγματα Κυττάρων Καταιγίδας. Τύποι, σύγκριση των χαρακτηριστικών τους και περιοχές εμφάνισης" (Storm Cell Complexes: Types, features comparison and regions of appearance), 5th Hellenic Conference of Meteorology, Climatology and Atmospheric Physics, Thessaloniki, (Presentation).

Tsagalidis, E., E. Chatzi, and D. Boucouvala, 2006: Comparison of the hailstorm characteristics between two different areas in Greece. *J. Wea. Mod.*, **38**, 11-15.

Tsagalidis, E., K. Tsitouridis, G. Evangelidis, and D. Dervos, 2008: Hail Size Estimation and Prediction using Data Mining Techniques. *ERAD 2008 – The 5th European Conference on Radar in Meteorology and Hydrology*, Helsinki, Finland, (Poster).

Tzoumaki, S., E. Tsagalidis, E. Chatzi, and S. Dimoutsi, 2006: Seeding operations in the Greek national hail suppression program. *J. Wea. Mod.*, **38**, 16-22.

Waldvogel, A., B. Federer, and P. Grimm, 1979: Criteria for the Detection of Hail Cells. *J. Appl. Meteor.*, **18**, 1521-1525.

Witt, A., M.D. Eilts, G.J. Stumpf, J.T. Johnson, E.D. Mitchell, and K.W. Thomas, 1998: An enhanced hail detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 286-303.

APPENDIX I: Acronyms definition

Hellenic Agricultural Insurance Organization	ELGA
Meteorological Applications Centre	KEME
Greek National Hail Suppression Program	GNHSP
Enterprise Electronics Corporation	EEC
Thunderstorm Identification, Tracking, Analysis and Nowcasting system	TITAN
Storm Cell Complexes	SCC
Vertically Integrated Liquid Water content	VIL
Vertically Integrated Liquid Water Density	VIL Density
Unicellular storms of a Single ordinary cell	S
Unicellular storms of a Supercell	SU
Multicell storms	M
Line storms	L
Wet Bulb Zero	WBZ
Severe Hail Index	SHI