Estimating the Probability of Rain in an SSM/I FOV Using Logistic Regression

DAVID S. CROSBY * AND RALPH R. FERRARO

NOAA/NESDIS/Office of Research and Applications, Camp Springs, Maryland

HELEN WU

American University, Washington, D.C.

(Manuscript received 30 June 1994, in final form 1 May 1995)

ABSTRACT

The SSM/I has been used successfully to estimate precipitation and to determine the fields of view (FOV) that contain precipitating clouds. The use of multivariate logistic regression with the SSM/I brightness temperatures to estimate the probability that it is raining in an FOV is examined. The predictors used in this study are those that have been evaluated by other investigators to estimate rain events using other procedures. The logistic regression technique is applied to a matched set of SSM/I and radar data for a limited area from June to August 1989. For this limited dataset the results are quite good. In one example, if the predicted probability is less than 0.1, the radar data shows only 2 of 340 FOVs have precipitation. If the predicted probability is greater than 0.9, the radar data shows precipitation in 748 of 774 FOVs. These probabilities can be used for both instantaneous and climate timescale retrievals.

1. Introduction

In this paper we will estimate the probability of rain in the field of view (FOV) of the Special Sensor Microwave/Imager (SSM/I) from the Defense Meteorological Satellite Program (DMSP). We will focus on using the scattering technique. The scattering technique is applicable over land and ocean and measures the scattering of upwelling radiation by precipitation size ice particles in the rain layer. Since the amount of scattering detected increases in proportion to the observation frequency, the higher-frequency channels (e.g., the 85-GHz channels on the SSM/I) are best suited for this approach. Rain rate can be derived indirectly based on the relationship between the volume of ice in the rain layer to the actual rainfall on the surface. Several investigators have developed instantaneous rain-rate retrieval algorithms using this approach. We use the channel combinations used by Spencer et al. (1989) and Grody (1991).

In this paper we examine the use of logistic regression techniques to estimate the probability of rain events using the SSM/I brightness temperatures over ocean regions. The physical procedures described in this paper are not new. As noted above, other papers have described techniques for using the SSM/I instrument to determine rain events. Here we use a statistical modification of threshold techniques for determining if it is raining in the SSM/I field of view. Instead of giving a fixed value for a function of the radiances above which (or below which) it is determined to be raining, the logistic regression technique gives an estimated probability $\hat{p}$ that is raining for all values of the function. This probability of rain can be used to map regions of very likely rain; that is, $\hat{p}$ is close to 1.0, possible rain and very unlikely rain, $\hat{p}$ is close to 0.0.

Most instantaneous SSM/I rain-rate retrieval techniques make a decision as to whether rain is detected in the satellite footprint through a series of computations based on physical constraints and/or empirical relationships. There is a tendency for these algorithms to determine raining pixels where they do not exist (e.g., false signatures), or conversely, to detect rain events only above a certain intensity, which minimizes the false signatures at the expense of not detecting all the rain events. Logistic regression gives the predicted probability of rain in a pixel. In many applications of satellite data it is important to have a measure of the reliability of the measurement. Techniques may require that the data are from a pixel that has no rain. This, for example, is the case for certain temperature sounding techniques. If it is almost certain that it is not raining in a pixel, then the sounding produced from that data will have more reliability than a sounding from a pixel where it may or may not be raining.

Additionally, the uses of this parameter should benefit climatological estimates of rainfall, which is a goal

* Permanent affiliation: American University, Washington, D.C.

Corresponding author address: Dr. David S. Crosby, NOAA/NESDIS, Rm. 810/E/RA 14:DC, 5200 Auth Road, Camp Springs, MD 20746.

of the Global Precipitation Climatology Project (GPCP) (Arkin and Xie 1994). Because of the rain/no-rain decisions made for the monthly rainfall retrievals, some information on the certainty of rain would benefit the global monthly rainfall distributions. For example, these rainfall estimates could be generated using different probability thresholds, with the resulting rain fields being stand-alone products.

The technique is applied to a dataset of matched SSM/I and radar data. For this limited dataset the results are quite good. Using one pair of channels, the actual occurrence of rain is less than one percent when the predicted probability is less than 0.1, and the actual occurrence is 97% when the predicted probability is above 0.9. We compare the results from three different sets of channels from the SSM/I.

2. Logistic regression

Logistic regression is used to predict dichotomous events. What follows will be a very brief outline of multivariate logistic regression. For a relatively complete and elementary introduction to this procedure, we recommend Neter et al. (1989).

In our case we are given the brightness temperatures from the SSM/I. For simplicity we will label them TB_1, TB_2, ..., TB_k. Let $p(TB_1, TB_2, \ldots, TB_k)$ be the probability that it is raining given $TB_1, TB_2, \ldots, TB_k$. We assume that this function p is of the form

$$p(TB_1, TB_2, \cdots, TB_k) = \frac{\exp(f)}{1 + \exp(f)}, \quad (1)$$

where

$$f = B_0 + \Sigma B_j TB_j. \quad (2)$$

The function given in (1) is called a logistic response function. It is a monotonic function of f, and its values are restricted to be between 0.0 and 1.0. We then define a dichotomous variable Y:

$$Y = \begin{cases} 0, & \text{not raining} \\ 1, & \text{raining.} \end{cases}$$

Given a set of vectors

$$\mathbf{X}'_i = (Y_i, TB_{1i}, TB_{2i}, \cdots, TB_{ki}) \quad i = 1, \cdots n,$$

the problem then becomes that of estimating the parameters $B_0, B_1, \ldots, B_k$.

To estimate the B's, we use the following procedure: let

$$P(Y_i = 1.0) = P_i = \frac{\exp(f)}{1 + \exp(f)} \quad (3)$$

and

$$P(Y_i = 0.0) = 1 - P_i,$$

where f is defined in (2). The joint distribution of Y_1, $Y_2, \ldots, Y_n$ is given by

$$g(y_1, y_2, \cdots, y_n) = \prod_{i=1}^{n} P_i^{y_i}(1 - P_i)^{1-y_i}. \quad (4)$$

We wish to maximize this function. This is the same as minimizing the negative of the log of g. This gives a loss function of

$$\text{loss} = -[(\Sigma Y_i \log(p_i) + \Sigma(1 - Y_i)\log(1 - p_i)], \quad (5)$$

where p_i is defined by (1). The minimization of this loss function requires a nonlinear procedure. We used the SYSTAT package on a VAX system to find the estimates of the parameters. The search procedure is a quasi-Newton method that estimates the first and second derivatives of the loss function and then does a directed search. The documentation for the SYSTAT package is in Wilkinson (1990).

It should be noted that the simple functional form of the TBs that is used in (2) could be modified. That is, there could be powers of the TBs or ratios or other nonlinear terms. The logistic technique could still be applied. The technique allows for very general forms of this function. However, since the technique uses numerical search procedures, the more complex the functional form the more difficult it may be to find a solution.

3. Description of matched dataset

a. Radar data and characteristics

The Japanese Meteorological Agency (JMA) maintains a high-resolution network of rain gauges supplemented by a weather radar network that gives complete coverage of Japan and adjacent coastal waters. The rain gauge network is part of the Automated Meteorological Data Acquisition System (AMeDAS). The radar observation network consists of 20 land-based and 2 ship-based radars. The radar data are routinely digitized to a 500 km × 500 km grid, with individual grid boxes being 5 km. Hourly rainfall analyses are created from a composite of the AMeDAS automated rain gauge data and radars at 0.05° latitude by 0.0625° longitude resolution. A Z–R relationship of $Z = 200R^{1.6}$ is used for the radar data.

b. SSM/I data and characteristics

The first SSM/I instrument was launched on 19 June 1987 aboard the DMSP block 5D-2 *F-8* satellite. Other instruments have been launched into orbit aboard the *F-10* (1 December 1990), *F-11* (28 November 1991), and *F-12* (August 1994) satellites. The SSM/I is a seven-channel radiometer measuring earth-emitted radiation at 19.35, 22.235, 37.0, and 85.5 GHz. All of the measurements are dual polarization except at

22.235 GHz, where only vertical polarization is measured. For the sake of simplicity, all of the subsequent referrals to the SSM/I channels will truncate the decimal (e.g., 85 GHz instead of 85.5 GHz) and will refer to the vertical and horizontal polarizations as V and H, respectively. More detail can be found in Hollinger et al. (1987).

For this study, data from the *F-8* satellite were used. The 85V channel became unusable in May 1988. A stepwise linear regression scheme developed by the Hughes Aircraft Company to "synthesize" the 85V channel using the other six channels was used in this investigation. For a discussion of the linear regression scheme see Ferraro and Marks (1995).

c. *Methodology*

Cases of rain systems where coincident radar and SSM/I observations existed were identified and processed in the following manner. All of the radar observations within a 7.5-km radius of the SSM/I field of view center were linearly averaged. This effectively averaged the radar data to the 85-GHz FOV. Additionally, data consisting of rain-free areas were identified and assembled using both radar and GMS satellite imagery. The period of study was June–August 1989, which corresponds to the Global Precipitation Climatology Project, First Algorithm Intercomparison Project period of study. High-quality radar data were made available for this project by JMA (Arkin and Xie 1994). It should be noted that there are many sources of error that occur when matching satellite and ground-based radar data. Detailed discussions are provided by Petty and Katsoras (1992) and Ferraro and Marks (1995).

4. Results

To test the use of logistic regression, we have selected data only from ocean areas where the effects on surfaces due to rainfall are minimized (Grody 1991). A subset of the dataset described above was selected. It consisted of 3432 matched points of radar and SSM/I data over the open oceans. The data were divided into two groups using the radar signal. If it was indicated that there was less than 1 mm h^{-1} of rain, the point was labeled as no rain. The other points were labeled as having rain. Of the 3432 points 1274 were placed in the no-rain category; the remaining 2158 were placed in the rain category. In addition, the 3432 points were divided into two datasets using a random number generator. There were 1699 in the first dataset and 1733 in the second. For all the results given in this section the parameters were estimated using the first dataset and then applied to the second.

Various models could be used for the function given in (1). The models given here were selected because they have appeared in the literature or were a simple extension of these.

The first model used the 85V and 85H channels. This is similar to the polarization-corrected temperature (PCT) approach described by Spencer et al. (1989). For simplicity we will give the linear form, which is in the exponential in (1). The coefficients for each of the models were found using a nonlinear search procedure in SYSTAT.

The f from (1) for this model is

$$f = 64.358 - 0.4985(85V) + 0.2696(85H). \quad (6)$$

We used this f in (1) to predict p for each of the points in the second dataset; that is,

$$\hat{p} = \frac{\exp(f)}{1 + \exp(f)}. \quad (7)$$

It is of interest to note that the logistic regression model has selected coefficients comparable to those selected by Spencer et al. (1989). In this paper they report using coefficients of −1.0 for the 85V channel and 0.45 for the 85H channel. The ratio of the coefficients given in (6) is approximately the same. This is especially interesting since the logistic function given in (7) is approximately linear in f for p values between 0.2 and 0.8 (Neter et al. 1989).

The $\hat{p}$ values were then placed in the bins 0.0–0.1, 0.1–0.2, . . . , and 0.9–1.0. The results of this model are given in Table 1, which gives $\hat{p}$ against the no-rain/rain values in frequency and the relative frequency of rain.

The logistic regression technique allows flexibility in deciding between areas of indicated rain or no rain. If we use the decision rule that no rain is predicted when the estimated value of $\hat{p}$ is less than 0.5 and rain when it is greater than 0.5, we get the following table. This is a technique suggested by Neter et al. (1989). Other values of $\hat{p}$ could be used to construct such a decision rule. This would depend on the risk of making incorrect decisions.

We see from Table 2 that when we predicted no rain we were correct 86% of the time and when we predicted rain we were correct 93% of the time. There are other techniques that can be used to make these types of

TABLE 1. Relationship between $\hat{p}$ values using 85V and 85H channels and the measured incidence of rain.

$\hat{p}$	No rain	Rain	Rain/total
0.0–0.1	338	2	0.006
0.1–0.2	77	11	0.125
0.2–0.3	79	26	0.248
0.3–0.4	41	19	0.317
0.4–0.5	21	22	0.512
0.5–0.6	16	39	0.709
0.6–0.7	16	48	0.750
0.7–0.8	9	57	0.864
0.8–0.9	5	132	0.964
0.9–1.0	26	748	0.966

TABLE 2. Satellite prediction of no-rain/rain against radar measurements using the 85V and 85H channels.

	Radar measurement		
Satellite prediction	No rain	Rain	Rain/total
No rain	556	80	0.144
Rain	72	1025	0.934

decisions. However, we see from Table 1 that if $\hat{p}$ is less than 0.1, then for this dataset less than 1% of the FOVs will have precipitation in them. We will be almost certain that it is not raining in a FOV where $\hat{p}$ is less than 0.1. It is basically this type of information that gives the advantage of logistic regression over a simple threshold technique.

The second model uses the 19V, the 22V, and the 85V channels to predict p. These are the same channels as reported by Grody (1991). The f from (2) is given by

$$f = 7.0866 + 0.02285(19V) + 0.1838(22V) - 0.22087(85V). \quad (8)$$

The results for this model are given in Table 3.

For model 3 the predictors 19V, 22V, 85V, and 85H were used. The f in (2) is given by

$$f = 52.2227 - 0.04021(19V) + 0.10420(22V) - 0.47913(85V) + 0.22731(85H). \quad (9)$$

The results for model 3 are given in Table 4.

A plot of the relative frequency of rain against the estimated value of p using the logistic regression for the three models is given in Fig. 1. The results are similar for all three models. Although, there are some differences. The additional predictor for model 3 versus those for model 1 do not seem to make significant differences in the ability to predict rain/no-rain events. There are still open questions about the interpretation of the results shown in Fig. 1.

TABLE 3. Relationship between $\hat{p}$ values using 19V, 22V, and 85V channels and the measured incidence of rain.

$\hat{p}$	No rain	Rain	Rain/total
0.0–0.1	248	3	0.012
0.1–0.2	123	3	0.023
0.2–0.3	97	10	0.093
0.3–0.4	58	32	0.357
0.4–0.5	35	32	0.468
0.5–0.6	16	56	0.778
0.6–0.7	6	53	0.898
0.7–0.8	15	73	0.830
0.8–0.9	6	134	0.957
0.9–1.0	24	715	0.968

TABLE 4. Relationship between $\hat{p}$ values using 19V, 22V, 85V, and 85H channels and the measured incidence of rain.

$\hat{p}$	No rain	Rain	Rain/total
0.0–0.1	359	3	0.008
0.1–0.2	50	5	0.091
0.2–0.3	77	19	0.198
0.3–0.4	39	22	0.361
0.4–0.5	35	19	0.351
0.5–0.6	18	35	0.660
0.6–0.7	14	61	0.813
0.7–0.8	7	70	0.909
0.8–0.9	5	126	0.962
0.9–1.0	24	745	0.968

5. Conclusions

We have examined the possibility of using logistic regression to estimate the probability of rain in an SSM/I FOV. It was not our purpose to provide complete answers to the questions of using the SSM/I to predict rain events. We selected a situation where there was good data and where there were proven rain retrieval techniques. Our purpose was to provide an example of the use of logistic regression. The use of this statistical technique clearly extracts more information in an easily usable form from the data than does a simple threshold procedure. Tables 1, 3, and 4 show that the technique gives good results on the limited dataset we used for this paper. Figure 1, which compares the results for the three algorithms tested, raises more questions than it answers. It may be possible to use this type of information to gain more insight into the physics of the problem.

If the application only requires in each case a simple decision that it is raining or not, then this technique is probably no better than a simpler method. However, in situations where certain conditions are

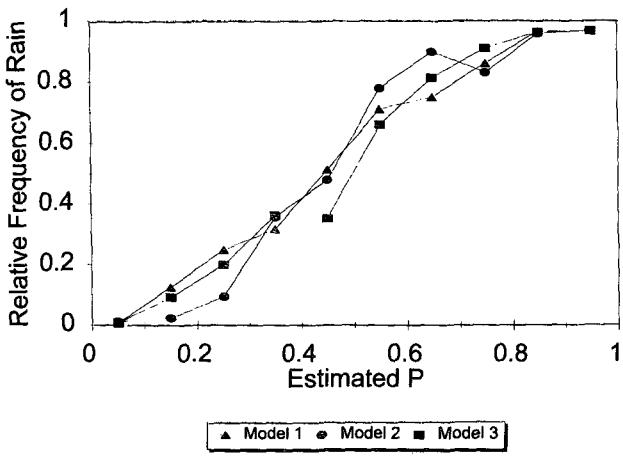

FIG. 1. The relative frequency of rain from the radar data against the predicted probability of rain using the SSM/I radiances for the three different sets of predictors.

satisfied by the model and where it is useful to have a measure of the certainty of the decision, then logistic regression should be considered as an alternative method. This information may be useful for both instantaneous and climate timescales. It is simple to carry out, and there are a number of statistical packages that will do logistic regression with a minimal amount of difficulty.

Acknowledgments. Part of this work was carried out while D. S. Crosby was a visiting scientist at National Institute of Water and Atmospheric Research Ltd., Wellington, New Zealand. We wish to thank the anonymous reviewers whose suggestions greatly improved the exposition in this paper.

REFERENCES

Arkin, P. A., and P. Xie, 1994: The global precipitation climatology project: First algorithm intercomparison project. *Bull. Amer. Meteor. Soc.*, **75,** 402–420.

Ferraro, R. R., and G. F. Marks, 1995: The development of SSM/I rain-rate retrieval algorithms using ground-based radar measurements. *J. Atmos. Oceanic Technol.*, **12,** 755–770.

Grody, N. C., 1991: Classification of snow cover and precipitation using the Special Sensor Microwave/Imager (SSM/I). *J. Geophys. Res.*, **96,** 7423–7435.

Hollinger, J., R. Lo, G. Poe, R. Savage, and J. Pierce, 1987: Special Sensor Microwave Imager User's Guide, Naval Research Laboratory, Washington, DC, 120 pp.

Neter, J., W. Wasserman, and M. H. Kutner, 1989: *Applied Linear Regression Models.* Irwin, 667 pp.

Petty, G. W., and K. B. Katsaros, 1992: *Nimbus-7* SMMR precipitation observations calibrated against surface radar during TAMEX. *J. Appl. Meteor.*, **31,** 489–505.

Spencer, R. W., H. M. Goodman, and R. E. Hood, 1989: Precipitation retrieval over land and ocean with the SSM/I. Part I: Identification and characteristics of the scattering signal. *J. Atmos. Oceanic Technol.*, **6,** 254–273.

Wilkinson, L., 1990: SYSTAT: *The System for Statistics.* SYSTAT, Inc., 677 pp.