2008
20th International Lightning Detection Conference
21-23 April • Tucson, Arizona, USA
2nd International Lightning Meteorology Conference
24-25 April • Tucson, Arizona, USA

Potential use of a Bayesian network for discriminating flash type from future GOES –R Geostationary Lightning Mapper (GLM) data

Richard Solakiewicz
NASA Postdoctoral Program (administered by ORAU), Huntsville, Alabama, USA

William Koshak
Earth Science Office, VP61, NASA Marshall Space Flight Center, Huntsville, Alabama, USA

1. Introduction

The future GOES-R Geostationary Lightning Mapper (GLM) will have many noteworthy characteristics including high detection efficiency, continuous monitoring of both ground and cloud flashes, and a wide (nearly hemispheric) field-of-view coverage. By contrast, the National Lightning Detection Network™ (NLDN) system has exceptional detection efficiency only for ground flashes over the continental US [Cummins et al., 2006; Murphy et al., 2006].

In this work, a fundamental question is asked: *Can GLM space-based optical measurements be used to discriminate ground flashes from cloud flashes?* Continuous knowledge of the *ratio* of cloud flashes to ground flashes (or "Z-ratio") derived from GLM data would provide a better understanding of thunderstorm dynamics, intensification, and evolution, and would improve the value-content of GLM data for severe weather warning. Knowledge of the Z-ratio for lumped regional flashes would also benefit lightning climate studies.

Statistical evidence suggests that the optical parameters of ground and cloud flashes are, on average, significantly different [Koshak 2007]. While no specific statements can be made from GLM data alone, it may be possible to give the *probability* that a given flash is a ground or cloud flash, conditioned on the evidence provided by GLM and additional GOES-R observations. Furthermore, it would be desirable to improve the flash type discrimination methodology by validating flash type results with NLDN data; this allows the methodology to "learn."

Following Koshak [2007], a Bayesian Network (BN) is suggested for use as a flash type discriminator. In a BN, graphical methods are employed to determine conditional independencies, and using these, a convenient form for the joint probability distribution is derived. Certain conditional probabilities must be known to execute BN calculations. These can be estimated using Optical Transient Detector (OTD) or Lightning Imaging Sensor (LIS) data.

In this work, we conducted initial tests using conditional probabilities for five OTD flash-level parameters (radiance, area, duration, # of optical groups, # of optical events) obtained from Koshak [2007]. These tests indicated that there is too much overlap in the conditional probability distributions between ground and cloud flashes to effectively discriminate flash type for individual flashes based on these five parameters. Additional data mining efforts are needed to intercompare additional ground and cloud flash characteristics. However, because of the central limit theorem, the overlap problem is substantially reduced by converting the conditional probability distributions into distributions of the means. This allows us to estimate the fraction of ground flashes in a large sample of flashes (for example, as accumulated in a single latitude/longitude bin for climate studies).

Finally, we note several advantages of the BN analyses. For example, message passing algorithms exist that allow for efficient calculation of the required probabilities [Pearl 1988]. Comparisons between proposed distributions and between structures are facilitated by distance measures [Jensen 1996]. The probabilities of each piece of evidence may be obtained by marginalizing. Correct findings should be positively correlated. The joint probability of all of the evidence should exceed the product of probabilities of independent findings. This provides a way of detecting conflicting data.

In this effort, we adopt the familiar convention that random variables are denoted by upper case letters. Their lower case counterparts will represent particular states of the associated random variables. Vectors and matrices will be denoted by bold roman letters. Probability parameters will generally be written as θ.

2. Prior and Posterior Probabilities

In Koshak [2007], the average parameters: radiance (J/m^2/ster/µm), area (km^2), duration (sec), # groups (integer) and # events (integer) are compared for ground and cloud flashes. Each average ground flash parameter was roughly numerically twice that of its cloud flash counterpart. Standard *z* scores ranged from about 40 to 65 indicating very high confidence for rejection of any hypotheses involving equality of the means. Very large sample sizes, ranging from tens to hundreds of thousands of lightning flashes, taken over several years were used. It was suggested that the probability that a flash is a ground flash could be updated given one of the observed optical parameters. In other words, the probability that flash type is a ground flash ($T = g$) given an observed radiance $R = r$ could be written as

$$P(g \mid r) = \frac{P(r \mid g)P(g)}{\left[P(r \mid g)P(g) + P(r \mid c)P(c)\right]}, \tag{1}$$

where $T = c$ indicates a cloud flash. Here, $P(g)$ is called the prior probability and can be estimated using the Z-ratio results obtained in Boccippio [2001], $P(g \mid r)$ is referred to as the posterior probability. The suggestion in (1) is generalized in the following section.

3. Bayesian Networks

In a BN, the variables are represented graphically as nodes. Arrows indicate the parentage (causes) for each node. From the graph, one is able to determine conditional independencies, and calculate the joint probability distribution. The output will be a conditional probability similar to that given in (1), but conditioning will be on a vector of optical parameters rather than just radiance. The required inputs consist of conditional probability tables supplying the probabilities of the evidence given the causes.

Advantages of Bayesian updating are given in Hopgood [2001]: (i) The technique is based on a proven statistical theorem. (ii) The result is expressed as a conditional probability which has a clearly defined and familiar meaning. (iii) The probability of a hypothesis can be updated in response to more than one observed variable. Disadvantages include: (i) The prior probability of an assertion must be known or estimated. (ii) Conditional probabilities must be known or estimated. (ii) Certain assumptions of independence may be unfounded. The large amount of data available and the strong correlations it exhibits allow for good estimates of the priors. Disadvantages (i) and (ii) are not expected to present difficulties. As for disadvantage (iii), all models depend on assumptions. The conditional independencies assumed here appear most reasonable.

Again, flash type will be represented by *T*. This is a binary variable which can take on states *c* and *g* for a cloud and ground flash, respectively. Provision will be made for variables that we may need to incorporate later; i.e., such as observations from the GOES-R Advanced Baseline Imager (ABI), or other GLM products. All of these will comprise the latecomers **L.** The remaining variables will be those quantities that are measured, **M**. The random vector **M** is given by

$$\mathbf{M} = (I, O, R, A, D, U, V), \tag{2}$$

where *I* and *O* are ice content and optical depth as measured by ABI, and *R*, *A*, *D*, *U*, *V* represent radiance, area, duration, number of groups, and number of events respectively. Each is a random variable which can take on a finite number of mutually exclusive states. Unlike in elementary probability, the variables are not events (subsets of the sample space). The probability of $R = r$ and $A = a$ will be written $P(r, a)$. The output desired of the network will be the entries in the table

$$P(T, \mathbf{L}, \mathbf{M}) \tag{3}$$

for all of the states of each of the variables. By marginalizing, we can determine $P(T \mid \mathbf{M})$**.**

The definition for conditional probability [Hogg and Tanis, 1997]

$$P(X \mid Y) = P(X, Y) / P(Y), \tag{4}$$

may be used to obtain the chain rule of probability

$$
\begin{aligned}
P(X_1, X_2, \ldots, X_n) &= P(X_1 \mid X_2, \ldots, X_n) P(X_2, \ldots, X_n) \\
&= P(X_1 \mid X_2, \ldots, X_n) P(X_2 \mid X_3, \ldots, X_n) \ldots P(X_n).
\end{aligned} \tag{5}
$$

Conditional independencies can be used to simplify the result.

There are 3 possible connections in a network, serial, diverging, and converging. See Fig. 1

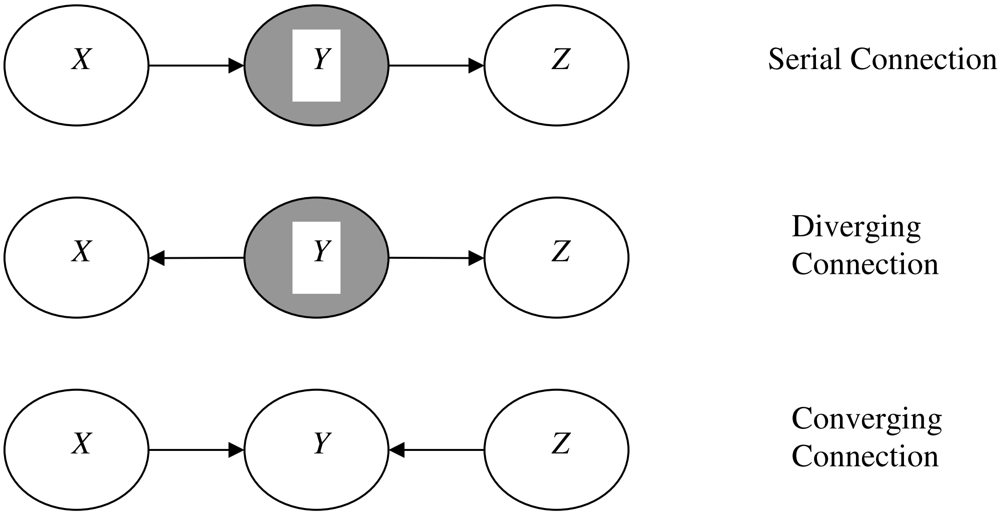
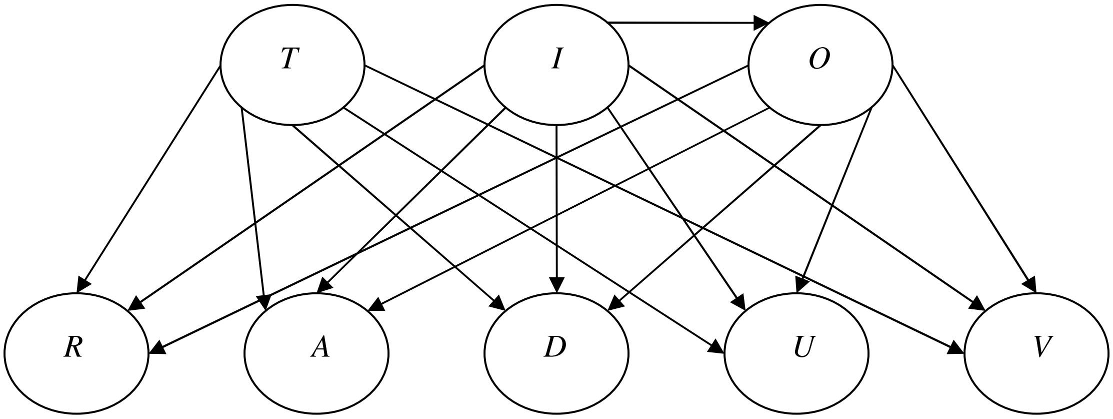


Fig. 1

The shaded variables are observed. In the serial and diverging connections, *X* and *Z* are independent given *Y*; $P(X \mid Y, Z) = P(X \mid Z)$. For a converging connection, *X* and *Z* are independent (i.e., $P(X, Z) = P(X)P(Z)$) unless *Y* or one of *Y*'s children are observed. Generalization to connections with more than 2 parents or children is immediate.

Presently, we assume the following structure for the network.



Directed Acyclic Graph
Fig. 2

Explicitly, from the chain rule we have

$$\begin{aligned} &P(T,I,O,R,A,D,U,V) \\ &= P(T)P(I\,|\,T)P(O\,|\,T,I)P(R\,|\,T,I,O)P(A,|\,T,I,O,R)P(D\,|\,T,I,O,R,A) \\ &\times P(U\,|\,T,I,O,R,A,D)P(V\,|\,T,I,O,R,A,D,U). \end{aligned} \tag{6}$$

Conditional independencies may be read directly from the Directed Acyclic Graph (DAG). The result is

$$\begin{aligned} &P(T,I,O,R,A,D,U,V) \\ &= P(T)P(I)P(O\,|\,I)P(R\,|\,T,I,O)P(A\,|\,T,I,O)P(D\,|\,T,I,O) \\ &\times P(U\,|\,T,I,O)P(V\,|\,T,I,O). \end{aligned} \tag{7}$$

We are only interested in a particular conditional probability in order to be able to discriminate cloud and ground flashes. It will turn out that we will not need *P*(*I*, *O*);

$$\begin{aligned} P(T\,|\,I,O,R,A,D,U,V) &= \frac{P(T,I,O,R,A,D,U,V)}{P(I,O,R,A,D,U,V)} \\ &= \frac{P(T)P(R\,|\,T,I,O)P(A\,|\,T,I,O)P(D\,|\,T,I,O)P(U\,|\,T,I,O)P(V\,|\,T,I,O)}{\sum_{T=g,c} P(T)P(R\,|\,T,I,O)P(A\,|\,T,I,O)P(D\,|\,T,I,O)P(U\,|\,T,I,O)P(V\,|\,T,I,O)}. \end{aligned} \tag{8}$$

The sum consists of only 2 terms. The tables might become large. Most are 4-dimensional arrays. For each lightning flash, the probability is obtained with only 12 operations (multiplications and additions).

The calculations above assume that exactly one state will be known for each of the measured variables in **M**. Evidence can consist of any number of states. In order to compress the notation, let M_j denote the *j*th component of **M** and $m_j^k, k = 1, 2, \ldots, t_j$ be the *k*th state of variable M_j. Each variable is allowed to have a different number of states. Evidence on M_j will be denoted by $\varepsilon_j(M_j) = m_j^{e_{j1}} \vee m_j^{e_{j2}} \vee \ldots \vee m_j^{e_{jq_j}}$. The $\vee's$ are logical symbols for "or". Each e_{jk} is an integer in $\{1, 2, \ldots, t_j\}$. It represents the *k*th state of M_j allowed by the evidence. It was assumed that the evidence allows q_j states of M_j. The set of all of the evidence vectors will be written as

$\boldsymbol{\varepsilon}(\mathbf{M})$. It is expected that the evidence will consist of exactly one state of M_j for each *j*. However, it is not difficult to allow for several states. Passing to the limit allows for direct extensions from discrete states to intervals for continuous variables.

From elementary probability theory, $P(x \vee y) = p(x) + P(y) - P(x \wedge y)$. Since the variables can only take on mutually exclusive states, the last term is zero and we have

$$P\left[\varepsilon_j\left(m_j\right)\right] = P\left(m_j^{e_{j1}} \vee m_j^{e_{j2}} \vee \ldots \vee m_j^{e_{jq_j}}\right) = \sum_{k=1}^{q_j} P\left(m_j^{e_{jk}}\right). \tag{9}$$

For the probability of *T* given evidence, we have

$$\begin{aligned} P(T \mid \boldsymbol{\varepsilon}) &= P(T \mid m_1^{e_{11}} \vee \ldots \vee m_1^{e_{1q_1}}, \ldots, m_n^{e_{n1}} \vee \ldots \vee m_n^{e_{nq_n}}) \\ &= \frac{\sum_{k_1}^{q_1} \cdots \sum_{k_n=1}^{q_n} P(T, m_1^{e_{1k_1}}, \ldots, m_n^{e_{nk_n}})}{\sum_T \sum_{k_1}^{q_1} \cdots \sum_{k_n=1}^{q_n} P(T, m_1^{e_{1k_1}}, \ldots, m_n^{e_{nk_n}})} \\ &= \frac{\sum_{\boldsymbol{\varepsilon}} P(T, \mathbf{M})}{\sum_{T,\boldsymbol{\varepsilon}} P(T, \mathbf{M})} = \frac{\sum_{\mathbf{L},\boldsymbol{\varepsilon}} P(T, \mathbf{L}, \mathbf{M})}{\sum_{T,\mathbf{L},\boldsymbol{\varepsilon}} P(T, \mathbf{L}, \mathbf{M})}. \end{aligned} \tag{10}$$

By taking maxima over the variables instead of sums, one can obtain a most probable configuration of variables given the evidence. This may be useful if some nuisance variables appear in **L** that cannot be measured or otherwise estimated.

4. Result #1: Determining Flash Type of a Single Flash

Since ABI data is not yet available, tests have been performed using (8) and the DAG in Fig. 2 with the *I* and *O* nodes omitted. The required conditional probabilities were obtained by using OTD data. When a prior probability of a lightning flash being a ground flash of 0.25 was used, the average values of *R*, *A*, *D*, *U*, *V* for a ground flash yielded a posterior probability of about 0.32. Using the average values for a cloud flash the posterior probability was decreased to about 0.22. While the change in the prior is significant for both cases, it is impractical to use the results to discriminate individual flashes since the results are not near unity (ground flash) or zero (cloud flash). This ambiguity is fundamentally linked to the fact that the parameter distributions for ground and cloud flashes overlap to a significant degree. Therefore, additional data mining of the optical characteristics of flashes will be needed (and possibly ABI data) to see if the BN can discriminate flash type on a flash-by-flash basis. Of course any such discrimination is fundamentally statistical.

5. Result #2: Determining the Fraction of Ground Flashes in a Set of Flashes

Rather than interrogating individual flashes, we can consider a set of flashes and then estimate what fraction α of these are ground flashes. The same method applies; one only need change *T* to α and replace the *R*, *A*, *D*, *U*, *V*, by their average values for the set of lightning,

$$P(\alpha \mid \mathbf{L}, \overline{\mathbf{M}}) = P(\alpha \mid \mathbf{L}) P(\overline{\mathbf{M}} \mid \alpha) \Big/ \sum\nolimits_{\alpha'} P(\alpha' \mid \mathbf{L}) P(\overline{\mathbf{M}} \mid \alpha'). \tag{11}$$

For initial testing, we removed *I* and *O*, and we eliminated **L**. If we assume that all values of α are equally likely, an assumption known as the *ignorance prior*, the result is the simple formula

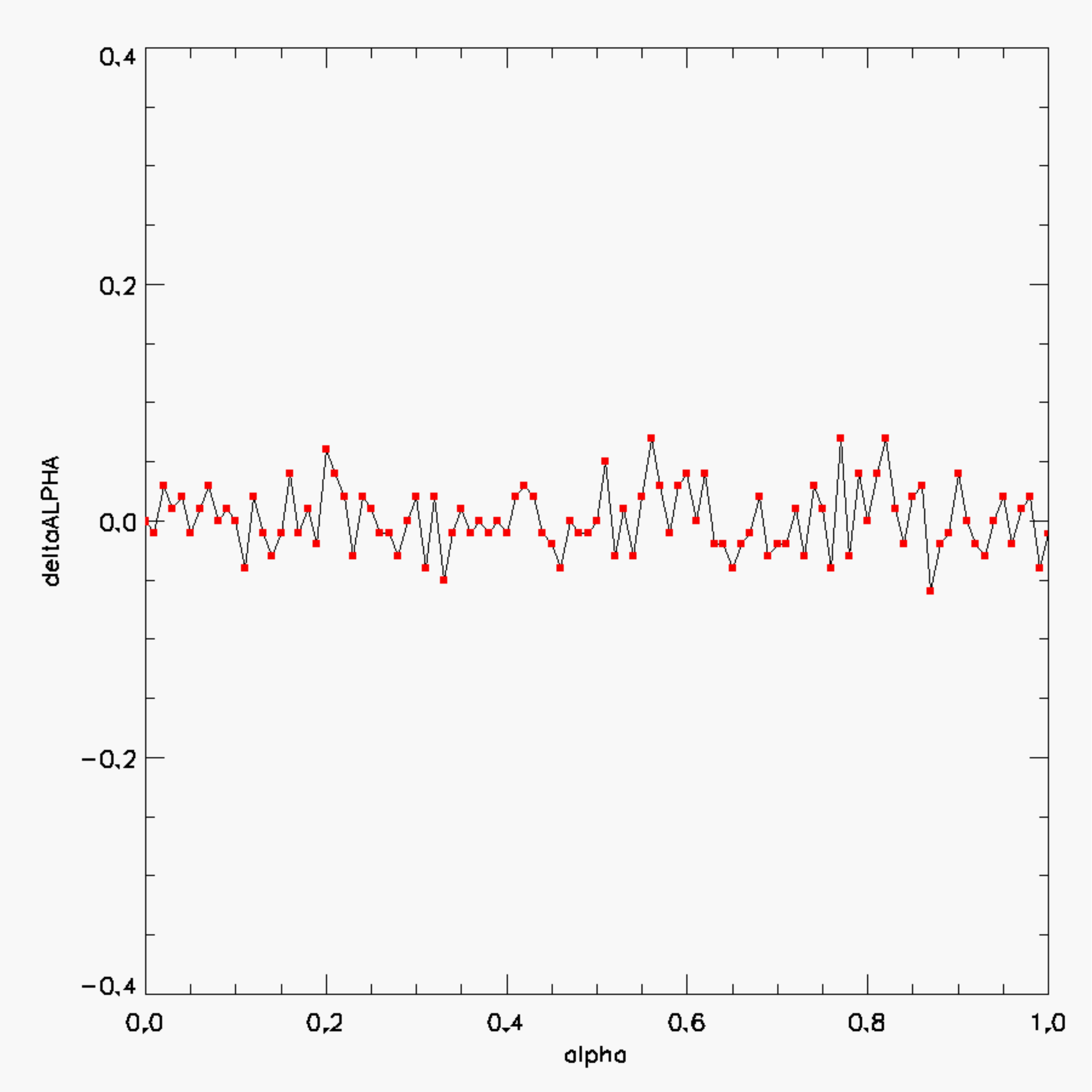
$$P(\alpha \mid \overline{\mathbf{M}}) = P(\overline{\mathbf{M}} \mid \alpha) \Big/ \sum\nolimits_{\alpha'} P(\overline{\mathbf{M}} \mid \alpha'). \tag{12}$$

Conditional probabilities for the average parameters $P(\overline{\mathbf{M}}\,|\,\alpha)$ may be estimated by resampling the OTD data for averages. It is also possible, in the case of large sets of lightning, to approximate these probabilities using normal distributions by appealing to the central limit theorem. Using a theorem for statistics, the sample mean for any of the average variables is distributed with mean and variance given by

$$\mu = (1-\alpha)\mu_c + \alpha\mu_g\,, \quad \sigma^2 = \frac{(1-\alpha)\sigma_c^2 + \alpha\sigma_g^2}{n}, \tag{13}$$

where μ_c, μ_g, σ_c^2, σ_g^2 denote the means and variances for a variable associated with pure cloud flashes and pure ground flashes respectively, and n is the sample size.

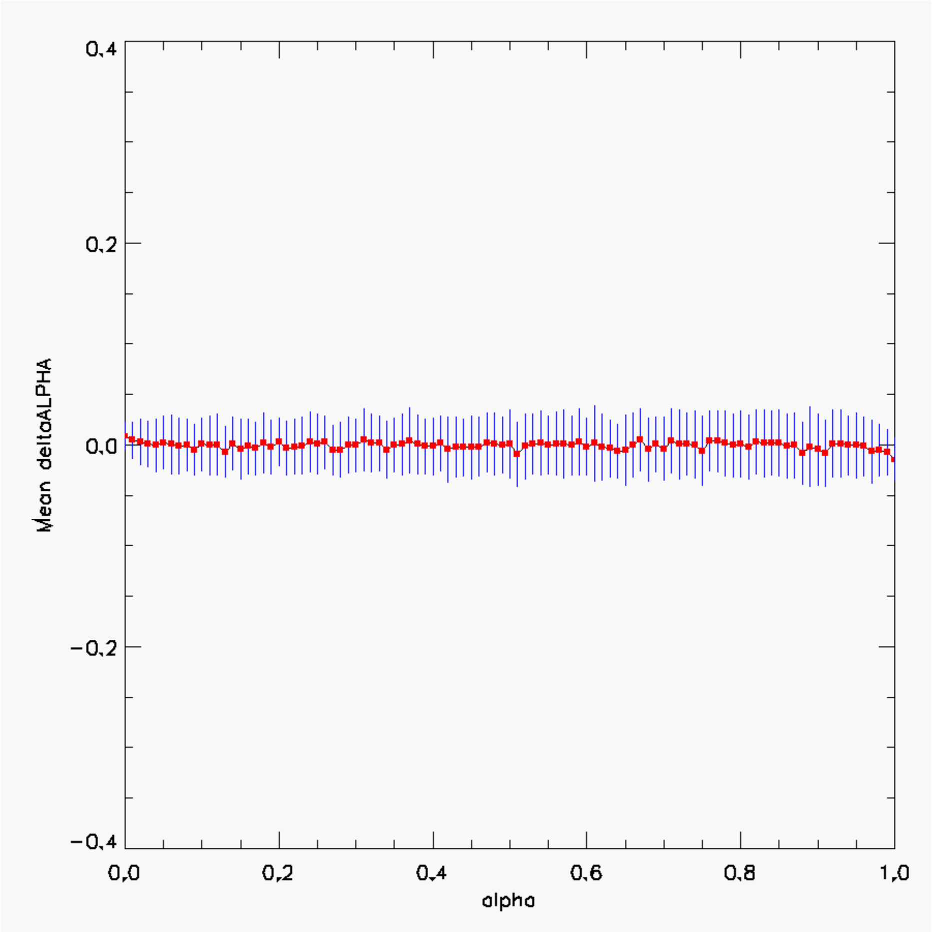
Eq. (12) was applied to randomly selected sets of lightning flashes, each set having a distinct value of α. The retrieval error (deltaALPHA) in α, as a function of α, is given in Fig. 3. Each set contained n = 10,000 flashes, and the errors were generally within 8% of the actual value of α for the set.



Error Plot for Retrieved Mixture
Fig. 3

Further refinements including geographical information, *I* and *O* parameters and perhaps the availability of additional parameters observed by GLM should further improve the results. Fig. 4 is the same experiment, but shows the mean retrieval errors obtained for 100 analyzed sets at a particular α; the vertical lines represent standard deviations about the mean.

Given the small retrieval errors in Figs. 3 and 4, we believe that this approach would be valuable in determining the ratio of cloud to ground flashes in a gridded lightning climatological dataset where any one geographical bin defines a set of lightning flashes with large sample *n*.



Mean Error Plot for Retrieved Mixture
Fig. 4

6. Conclusions

We have examined the feasibility of employing a Bayesian Network (BN) for discriminating lightning flash type (ground or cloud) using the statistics of certain flash optical parameters (e.g., radiance, area, duration, # optical groups, # optical events). It was found that the distribution of any parameter for a ground and cloud flash have sufficient overlap so as to make discrimination of flash type difficult on a flash-by-flash basis. This implies that flash type discrimination for a single flash must involve more parameters than the five we have examined here. However, any of our 5 parameters do have mean values that differ between ground and cloud flashes. Moreover, we have found that the distributions of the mean for any of these five parameters involve much less overlap between the ground and cloud flashes. In the BN framework, this implies that it is conceivable to discriminate on a set of *n* flashes. That is, given a set of flashes collected over a certain time period and composed of so many ground and cloud flashes, the BN could estimate the fraction, α, of ground flashes and the fraction (1-α) of cloud flashes. This paper has shown that the BN evidently can determine α for large *n* to within a reasonable percent (see section 5).

Hence, the primary application of this approach would be for categorizing the ratio of cloud flashes to ground flashes in lightning climatology datasets.

The BN approach is not static. We can begin with estimates of conditional probabilities necessary to start the calculation, but the network can learn from GLM and NLDN data. Methods are available for parameter learning using complete and incomplete data sets (see Appendix). In the case of incomplete data, an approximate method may be used if the exact method becomes computationally expensive.

6. References

Boccippio, D., et al., 2001: Combined satellite and surface-based estimation of the intracloud:cloud-to-ground lighting ratio over the continental United States, *Mon. Weather Rev.*, **129**, 108-122.

Heckerman, D., 1996: A tutorial on learning with Bayesian networks, *Microsoft Research Technical Report* MSR-TR-95-06.

Hogg, R. and E. Tanis, 1997: Probability and Statistical Inference, 5th ed., Prentice-Hall Upper Saddle River, NJ, 82-89.

Hopgood, A., 2001: Intelligent Systems for Engineers and Scientists, CRC Press, Boca Raton, 195, 196.

Jensen, F., 1996: An Introduction to Bayesian Networks, Springer-Verlag, NY.

Korb, K. and A. Nicholson, 2004: Bayesian Artificial Intelligence, Chapman & Hall/CRC, Boca Raton, 320-322.

Koshak, W., 2007: OTD observations of continental US ground and cloud flashes", ICAE, Beijing, China.

Pearl, J., 1988: Probabilistic Reasoning in Intelligent Systems, Morgan Kaufmann, San Mateo, CA, Ch. 2, 4.

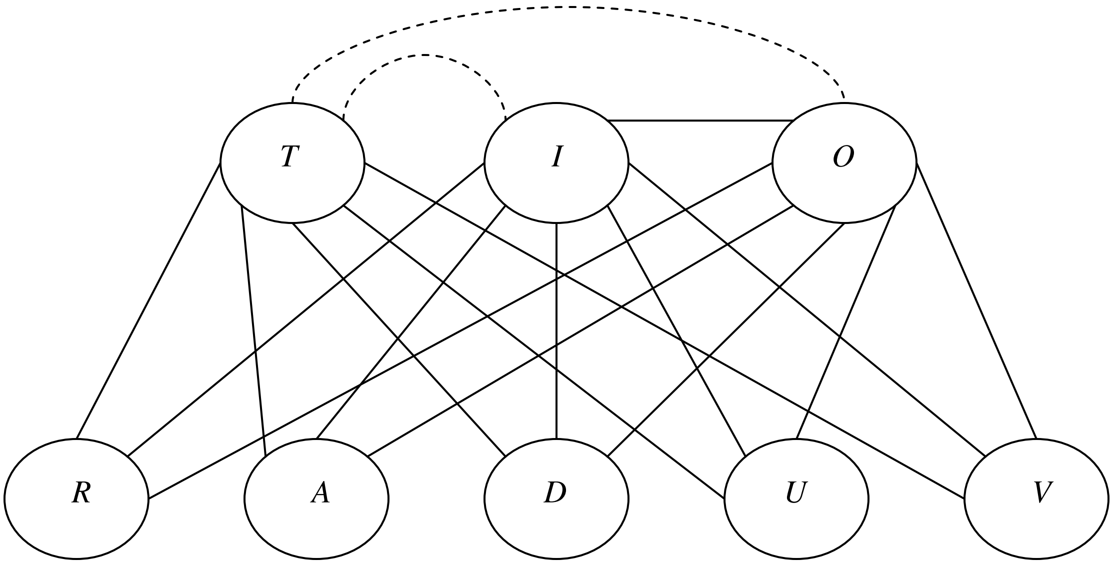
Spiegelhalter, D. and S. Lauritzen, 1990: Sequential updating of conditional probabilities on directed graphical structures, *Networks*, **20**, 579-605.

APPENDIX

A.1 Methods for BN Calculation

A number of computer programs are available for performing BN calculations. A summary of software packages may be found in an appendix of [Korb and Nicholson 2004]. A popular algorithm used by many packages is a message passing scheme using junction trees [Pearl 1988, Jensen 1996].

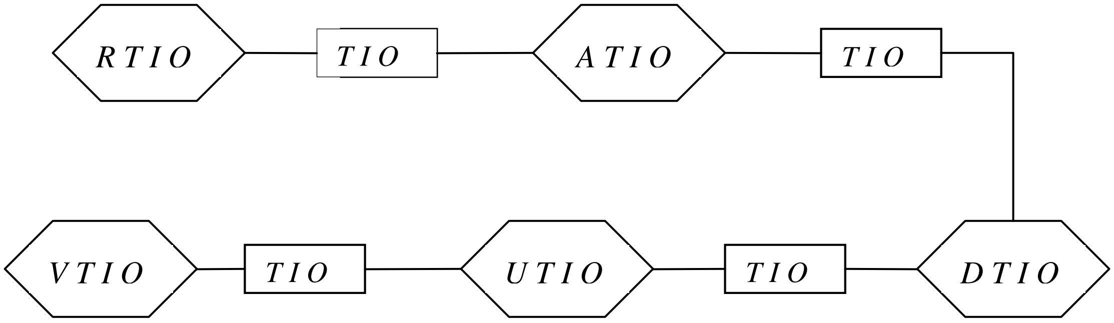
Large BN's can be computationally expensive. One method to improve efficiency is to group subsets of nodes in such a way that the conditional independencies can be exploited. The directions in the DAG are suppressed and connections are made between nodes that share common children. The result is known as a moral graph. See Fig. 5 for a moral graph corresponding to the DAG of Fig. 2.



Moral Graph
Fig. 5

Cycles of length greater than 3 must be removed by adding links. This is called trangulation. The proposed lightning BN's moral graph is already triangulated.

Nodes are selected one at a time, and for each node a clique is formed by considering all of that node's neighbors that are pair wise joined. The node under consideration is then removed together with its links. This process is repeated until every node is a subset of some clique. The cliques are connected with separators that comprise the intersections of the cliques. Messages containing joint probabilities are passed between the cliques through the separators. No message needs more than one path. Therefore, redundant separators may be deleted. The result is known as a junction tree. A junction tree has the property that for each pair of cliques X and Y, all paths between X and Y contain $X \cap Y$. A junction tree for the proposed lightning BN is very simple; see Fig. 6.



Junction Tree
Fig. 6

Associated with each clique, is the joint probability distribution of the variables therein. For example, for the clique *RTIO*, we would have *P*(*T*)*P*(*I*)*P*(*O* | *I*)*P*(*R* | *T*, *I*, *O*). Initially, an appropriately sized table of ones is associated with each separator. In this case, we would use $2 \times n_I \times n_O$ tables, where n_I and n_O are the number of states of *I* and *O* respectively. For each consecutive pair of cliques X and Y with separator S, a message is passed from X to Y via S by replacing the initial table t_S associated with S by a table t_S^* with entries $\sum_{X \setminus S} P(X)$. A table t_Y corresponding to Y, whose entries are P(Y) is replaced with $t_S^* t_Y / t_S$. Once messages are passed in both directions, the junction tree is said to be consistent. Evidence is entered by setting all entries not allowed by the evidence equal to zero. Renormalization and a round of message passing allow for all of the relevant information to be collected in a single clique. The required probability can be obtained by interrogating the table for this clique.

A.2 Verifying Applicability

If there are no data errors, the evidence should be such that the joint probability of all of the findings should exceed the product of the probabilities of independent findings. One way of measuring the amount of conflict is the data is to calculate

$$\ln\left[P(\varepsilon_1)\ldots P(\varepsilon_1)/P(\boldsymbol{\varepsilon})\right]. \tag{14}$$

A positive value of (14) indicates a possible conflict of data. A similar calculation can be preformed for each clique or group of cliques in order to trace flawed evidence.

Let *P* and *P** be 2 joint distributions of the same variables. A distance measure between distributions is defined by

$$D(P,P^*) = \sum_{X \in U}\left[P(X) - P^*(X)\right]^2, \tag{15}$$

where the summation extends over all configurations X in the universe of variables U. A number of other measures of distance between distributions are also in use. See [3].

A.3 Learning

In order to calculate (8), a number of conditional probabilities must be known. These may be estimated or assumed. It is desirable to be able to update these conditional probabilities using experience gained from test cases or actual cases for which the output can be verified. Such updating is known as parameter learning. Using data to update structure is also available in the literature [Heckerman 1996], but it will not be discussed here.

The notation will be compressed by letting $\mathbf{X} = \{T, \mathbf{L}, \mathbf{M}\}$ be a vector of all of the variables in the DAG; $X_1 = T$, $X_2 = L_1$, … . If $Y, Z \subseteq X$, then

$$P(\mathbf{Y} \mid \mathbf{Z}) = \sum\nolimits_{\mathbf{x} \setminus \{\mathbf{y},\mathbf{z}\}} P(\mathbf{X}) \Big/ \sum\nolimits_{\mathbf{X} \setminus \mathbf{Z}} P(\mathbf{X}). \tag{16}$$

Everything that is needed can be obtained once the joint probability distribution is known. From the chain rule (5),

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{pa}_i), \tag{17}$$

where $\mathbf{pa}_i$ is the parent set of X_i. We will adopt the notation and methods in [Heckerman 19968]. Conditioning on the probability parameters and the network structure *S* gives

$$P(\mathbf{X} \mid \boldsymbol{\theta}, S) = \prod_{i=1}^{n} P(X_i \mid \mathbf{pa}_i, \boldsymbol{\theta}_\imath, S). \tag{18}$$

Here, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$. Assume that each X_i can take on r_i states each denoted by x_i^k, $k = 1, \ldots, r_i$.. Each variable in $\mathbf{pa}_i$ can also take on a number of states. Denoting each state of each variable with a superscript makes the notation too cumbersome. Instead, se index each configuration within $\mathbf{pa}_i$ using $\mathbf{pa}_i^j$, $j = 1, \ldots, q_i$. Each factor in

(15) represents a probability table for X_i being in state k and its parent set $\mathbf{pa}_i$ being in configuration j. The individual probabilities are written as

$$P\left(x_i^k \mid \mathbf{pa}_i^j, \boldsymbol{\theta}_{ij}, S\right) = \theta_{ijk} \,;$$
$$\boldsymbol{\theta}_{ij} = \left(\theta_{ij2},\ldots,\theta_{ijr_i}\right),\ \ \theta_{ij1} = 1 - \sum_{k=2}^{r_i} \theta_{ijk}\,,\ \ \boldsymbol{\theta}_i = \left(\boldsymbol{\theta}_{i1},\ldots,\boldsymbol{\theta}_{iq_i}\right). \tag{19}$$

These are the entries in the probability tables that will required. Their values must be given initially, and they will be updated based on the data.

Let D = {$\mathbf{x}_1$, … ,$\mathbf{x}_\mathrm{N}$} be a random sample containing N observations of all n variables. If there are N_{ij1} occurrences of x_i^1 given that its parent set is in configuration j , N_{ij2} occurrences of x_i^2 given that its parent set is in configuration j, etc. in D, then assuming a multinomial distribution,

$$P\left(D \mid \boldsymbol{\theta}_{ij}, S\right) = \theta_{ij1}{}^{N_{ij1}} \theta_{ij2}{}^{N_{ij2}} \ldots \theta_{ijr_i}{}^{N_{ijr_i}}\,. \tag{20}$$

The maximum likelihood configuration of the parameters θ_{ijk} is given by

$$\theta_{ijk} = N_{ijk} \big/ N_{ij}\,,\ \ N_{ij} = \sum_{k=1}^{r_i} N_{ijk}\,. \tag{21}$$

For convenience, a Dirichlet distribution is assumed for the parameters

$$P\left(\boldsymbol{\theta}_{ij} \mid S\right) = \mathrm{Dir}\left(\boldsymbol{\theta}_{ij} \mid \boldsymbol{\alpha}_{ij}\right);\ \ \boldsymbol{\alpha}_{ij} = \left(\alpha_{ij1},\ldots,\alpha_{ijr_i}\right),$$
$$\mathrm{Dir}\left(\boldsymbol{\theta}_{ij} \mid \boldsymbol{\alpha}_{ij}\right) = \frac{\Gamma\left(\sum_{k=1}^{r_i} \alpha_{ijk}\right)}{\prod_{k=1}^{r_i} \Gamma\left(\alpha_{ijk}\right)} \left(1-\theta_{ij2}-\ldots-\theta_{ijr_i}\right)^{\alpha_{ij1}-1} \theta_{ij2}{}^{\alpha_{ij2}-1} \ldots \theta_{ijr_i}{}^{\alpha_{ijr_i}-1}. \tag{22}$$

This is a prior distribution for the parameter set. The Dirichlet hyperparameters α_{ijk} for any complete structure are constrained by

$$\alpha_{ijk} = \alpha P\left(x_i^k, \mathbf{pa}_i^j \mid S\right), \tag{23}$$

where α is an equivalent sample size. The posterior probability depends on the sample D,

$$P\left(\boldsymbol{\theta}_{ij} \mid D, S\right) = \frac{P\left(\boldsymbol{\theta}_{ij} \mid S\right) P\left(D \mid \boldsymbol{\theta}_{ij}, S\right)}{P\left(D \mid S\right)}$$
$$= \left[\frac{\Gamma\left(\sum_{k=1}^{r_i} \alpha_{ijk}\right)}{P(D \mid S) \prod_{k=1}^{r_i} \Gamma\left(\alpha_{ijk}\right)}\right] \left(1-\theta_{ij2}-\ldots-\theta_{ijr_i}\right)^{\alpha_{ij1}+N_{ij1}-1} \theta_{ij2}{}^{\alpha_{ij2}+N_{ij2}-1} \ldots \theta_{ijr_i}{}^{\alpha_{ijr_i}+N_{ijr_i}-1} \tag{24}$$
$$= \mathrm{Dir}(\boldsymbol{\theta}_{ij} \mid \boldsymbol{\alpha}_{ij} + \mathbf{N}_{ij}).$$

The quantity in brackets was recognized as $\Gamma\left[\sum_{k=1}^{r_i}\left(\alpha_{ijk} + N_{ijk}\right)\right] \Big/ \prod_{k=1}^{r_i} \Gamma\left(\alpha_{ijk} + N_{ijk}\right)$ since this is a probability density function, and it must integrate to one.

The probability of obtaining a given observation $\mathbf{x}_{N+1}$ (after the N observations in D) is given by

$$
\begin{aligned}
P(\mathbf{x}_{N+1} \mid D,S) &= \int P(\mathbf{x}_{N+1},\boldsymbol{\theta} \mid D,S)\,d\boldsymbol{\theta} \\
&= \int P(\mathbf{x}_{N+1} \mid \boldsymbol{\theta},D,S)P(\boldsymbol{\theta} \mid D,S)\,d\boldsymbol{\theta} \\
&= \int \prod_{\imath=1}^{n}\left[P\left(x_i^{k_i} \mid \mathbf{pa}_i^{j_i},\boldsymbol{\theta}_{ij_i},D,S\right)\right]\prod_{I=1}^{n}\prod_{J=1}^{q_i}\left[P(\boldsymbol{\theta}_{IJ} \mid D,S)\right]d\boldsymbol{\theta} \\
&= \prod_{\imath=1}^{n}\left\{\int P\left(x_i^{k_i} \mid \mathbf{pa}_i^{j_i},\boldsymbol{\theta}_{ij_i},D,S\right)P\left(\boldsymbol{\theta}_{ij_i} \mid D,S\right)d\boldsymbol{\theta}_{ij_i}\prod_{J\neq j_i}\left[\int P(\boldsymbol{\theta}_{iJ} \mid D,S)\right]d\boldsymbol{\theta}_{iJ}\right\}.
\end{aligned}
\tag{25}
$$

We are taking each variable X_i in $\mathbf{x}_{N+1}$ as appearing in a given state k_i with its parent set in configuration j_i; i.e., each state k and configuration j depends on i. Substituting (19) and using the fact that the integrals in the last product are all ones leaves

$$
P(\mathbf{x}_{N+1} \mid D,S) = \prod_{i=1}^{n}\int \theta_{ij_ik_i}P\left(\boldsymbol{\theta}_{ij_j} \mid D,S\right)d\boldsymbol{\theta}_{ij_j}. \tag{26}
$$

Substituting (19) and performing the integration results in

$$
P(\mathbf{x}_{N+1} \mid D,S) = \prod_{i=1}^{n}\frac{\alpha_{ij_ik_i}+N_{ij_ik_i}}{\alpha_{ij_i}+N_{ij_i}};\quad \alpha_{ij_i} = \sum_{k=1}^{r_i}\alpha_{ij_ik}. \tag{27}
$$

The experience in D has contributed in the expected manner in predicting the outcome of the next sample.

A.4 Incomplete Data

The development in the previous section assumed complete data sets. Some corruption of data should be anticipated. If $\mathbf{Y}$, $\mathbf{Z} \subseteq \mathbf{X}$ are the observed and unobserved variables in a given case ($\mathbf{X}$ denotes the compete set of variables). The posterior distribution of $\boldsymbol{\theta}_{ij}$ may be obtained using [Spiegelhalter and Lauritzen 1990]

$$
P(\boldsymbol{\theta}_{ij} \mid \mathbf{y},S) = \sum_{k=1}^{r_i}P\left(x_i^k,\mathbf{pa}_i^j \mid \mathbf{y},S\right)P\left(\boldsymbol{\theta}_{ij} \mid x_i^k,\mathbf{pa}_i^j,S\right) + \left[1-P\left(\mathbf{pa}_i^j \mid \mathbf{y},S\right)\right]P(\boldsymbol{\theta}_{ij} \mid S). \tag{28}
$$

Unfortunately, each of the probabilities involving $\boldsymbol{\theta}_{ij}$ are Dirichlet distributions for every case. As the number of cases grows, the number of Dirichlet functions proliferate making computation impractical. This formula was included since it is expected that the GLM being new should rarely yield bad or missing data.

If using (28) becomes impractical, a Gaussian approximation may be used. It consists of expanding $\ln[P(\boldsymbol{\theta} \mid D,S)]$ in a Taylor series about the maximum likelihood estimate of $\boldsymbol{\theta}$**.** The first derivative term will be zero, and the second will involve quadratic terms. Exponentiating results in a multivariate Gaussian distribution.