# Lightning Prediction for Australia Using Multivariate Analyses of Large-Scale Atmospheric Variables

BRYSON C. BATES

*CSIRO Oceans and Atmosphere, Wembley, and School of Agriculture and Environment, The University of Western Australia, Crawley, Western Australia, Australia*

ANDREW J. DOWDY

*Bureau of Meteorology, Melbourne, Victoria, Australia*

RICHARD E. CHANDLER

*Department of Statistical Science, University College London, London, United Kingdom*

(Manuscript received 27 July 2017, in final form 22 October 2017)

ABSTRACT

Lightning is a natural hazard that can lead to the ignition of wildfires, disruption and damage to power and telecommunication infrastructures, human and livestock injuries and fatalities, and disruption to airport activities. This paper examines the ability of six statistical and machine-learning classification techniques to distinguish between nonlightning and lightning days at the coarse spatial and temporal scales of current general circulation models and reanalyses. The classification techniques considered were 1) a combination of principal component analysis and logistic regression, 2) classification and regression trees, 3) random forests, 4) linear discriminant analysis, 5) quadratic discriminant analysis, and 6) logistic regression. Lightning-flash counts at six locations across Australia for 2004–13 were used, together with atmospheric variables from the ERA-Interim dataset. Tenfold cross validation was used to evaluate classification performance. It was found that logistic regression was superior to the other classifiers considered and that its prediction skill is much better than using climatological values. The sets of atmospheric variables included in the final logistic-regression models were primarily composed of spatial mean measures of instability and lifting potential, along with atmospheric water content. The memberships of these sets varied among climatic zones.

## 1. Introduction

Appreciable attention has been given to the problems of thunderstorm classification and prediction in recent years. Many techniques have been used, including empirical orthogonal function and canonical correlation analyses (e.g., Muñoz et al. 2016), classification and regression trees (e.g., Burrows et al. 2005), random-forest classification (e.g., Blouin et al. 2016), quadratic discriminant analysis (e.g., Sánchez et al. 1998), logistic regression (e.g., Mazany et al. 2002; Sousa et al. 2013; Romps et al. 2014), and dynamical modeling (e.g., Yair et al. 2010; Lynn et al. 2012; Zepka et al. 2014). Subject areas have included very-short-range forecasting, seasonal prediction, and climatological studies. The systematic evaluation of the performances of several different classification techniques when applied to datasets from a wide range of climatic zones has not received much attention, however.

The principal aim of this study is to investigate the relationships between lightning activity and atmospheric conditions at the coarse spatial and temporal scales of currently available climate models and reanalyses, including a comparison of different modeling techniques that are based on a range of thermodynamic measures. The lightning data used in this study are from several locations in Australia, covering a variety of climate types. The description of study sites and data, multivariate analyses, and cross-validation experiments below parallels that of Bates et al. (2017). The material covered in section 2 is derived from there with minor

*Corresponding author*: Andrew J. Dowdy, andrew.dowdy@bom.gov.au

DOI: 10.1175/JAMC-D-17-0214.1

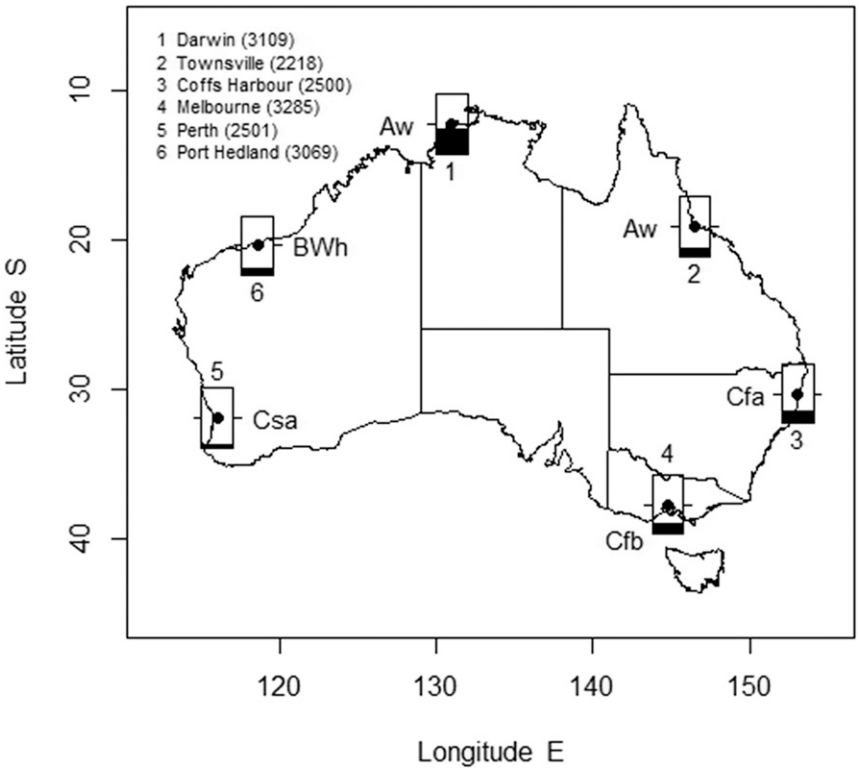


FIG. 1. Locations of CIGRE 500 sensors (black dots), observed proportions of lightning days at the sensors (indicated by relative heights of black bars within rectangles centered on sensor locations), and corresponding Köppen climate-classification zones: Aw = tropical savanna, Cfa = humid subtropical, Cfb = marine west coast, Csa = Mediterranean, and BWh = subtropical desert climate. Numbers in parentheses refer to record lengths (days).

modification and is intended to provide sufficient detail to allow readers to assess the validity and generalizability of the results presented here. Results are presented in section 3, and a summary of key research findings is given in section 4.

## 2. Data and methods

The daily lightning data used in this study were collected from six Comité Internationale des Grandes Réseaux Electriques (International Committee on Large Electric Systems; CIGRE) model "CIGRE 500" lightning-flash counters (Fig. 1; Table 1). The total number of flash counts was considered herein because 1) although the CIGRE 500 sensor was designed specifically to detect cloud-to-ground flashes, it also responded to cloud-to-cloud flashes, with about 68% of the lightning-flash counts recorded being due to cloud-to-ground flashes; 2) estimates of the effective horizontal ranges of the counters for cloud-to-ground flashes and cloud-to-cloud flashes are different (30 and 15 km, respectively); and 3) the ratio of intracloud to cloud-to-ground flashes can vary considerably depending on thunderstorm type and intensity, region of occurrence, and season (Rakov and Uman 2003). The counters were read manually each day between 0800 and 0900 local time. They were selected because of their record length and quality and their different climatic settings. The period of record varies from January of 2004 to at least December of 2010 (Townsville, Queensland, Australia) and at most February of 2013 (Melbourne, Victoria, Australia). A thunderstorm was deemed to have occurred during a 24-h period if the counter registered at least one lightning-flash count (LFC).

A second threshold of two LFCs per 24-h period was also used to assess the degree of the sensitivity to threshold selection, because of the (undocumented) possibility that some counts of one flash may have originated from a source other than lightning. For the Melbourne site, which has a high proportion of lightning days with LFC = 1 (Table 1), it was found that the results obtained from the procedures described below showed slight to some sensitivity in terms of atmospheric-variable selection and levels of prediction skill. There was no impact on the interpretation of the results, however. Therefore, the results obtained using

TABLE 1. Site and data details for Australian CIGRE 500 lightning-flash counters used in the study. Daily records of LFC cover the period from January of 2004 to at least December of 2010 (Townsville) and at most February of 2013 (Melbourne).

| Site No. | Location | Record length (days) | No. of lightning days | Percentage of lightning days with LFC = 1 |
|---|---|---|---|---|
| 1 | Darwin | 3109 | 1350 | 6.22 |
| 2 | Townsville | 2218 | 286 | 14.0 |
| 3 | Coffs Harbour | 2500 | 501 | 25.0 |
| 4 | Melbourne | 3285 | 570 | 34.5 |
| 5 | Perth | 2501 | 148 | 18.9 |
| 6 | Port Hedland | 3069 | 401 | 17.2 |

the threshold of two LFCs in a 24-h period will not be reported here.

Data for 31 atmospheric fields were obtained from the European Centre for Medium-Range Weather Forecasts interim reanalysis (ERA-Interim) archive (Dee et al. 2011): the fields are listed in Table 2. The fields represent a broad variety of physical processes that can be associated with deep convection, including both dynamical and thermodynamical processes. The variables cover various measures of temperature lapse, moisture content, vertical motion, and water phase state at a range of different pressure levels. The spatial and temporal resolution of the dataset is 0.75° and 6 h. For each CIGRE 500 site, atmospheric data were extracted for the 49 reanalysis grid points closest to the sensor's location. The lightning series was synchronized with the ERA-Interim series for 0600 UTC (1600 h eastern Australia time) within the 24-h period represented by the lightning data. This procedure was done because, in general, weather conditions are more favorable for lightning activity to occur during the late-afternoon period than at other times of the day or night (see, e.g., Christian et al. 2003; Dowdy and Mills 2009). Moreover, the additional information provided by data at other time steps was found to be largely redundant because correlations within a 24-h period were invariably high. For example, correlations between individual variables at 0600 and 1200 UTC, spanning the time period during which most deep-convective processes occur in Australia, are greater than 0.85 in every case examined, and most are greater than 0.95.

Quadratic surfaces and low-dimensional summary statistics (LDSS) were used to characterize the main features of the atmospheric fields on each day (appendix A). Six LDSS were considered: the intercept of the quadratic surface (mu), the magnitude of the gradient vector (gd) and its direction (dr), Gaussian curvature (gc), vertical gradient (vg), and adjusted correlation coefficient squared $R^2$ (r2). The adjusted $R^2$, as a goodness-of-fit measure for the quadratic surfaces, was included as a measure of spatial (dis)organization in the atmospheric fields.

For each candidate variable (designated hereinafter by the convention "LDSS code.field name"), comparative box plots were used to contrast its values for lightning and nonlightning days. A data matrix was constructed using variables that showed the greatest contrast. The columns of this matrix were standardized to zero mean and unit variance. This ensures that the variables were placed on a commensurable scale without disturbing the shape of their probability distributions, facilitates interpretation of the results of a discriminant or regression analysis, and helps to concentrate precisely on the conditions that are present during nonlightning and lightning days because it focuses on the relative variations of each variable within its own physical limits. The "colldiag" function from the "perturb" package in the R computing software environment (https://cran.r-project.org/web/packages/perturb/index.html) was used to detect the presence of collinearity in the data matrix. Colldiag is an implementation of the regression collinearity diagnostic procedures found in Belsley (1991). It computes the condition indices of the data matrix and provides the *variance decomposition proportions* associated with each condition index. As a rule of thumb, variables with proportions that are greater than 0.99 were considered to be sources of severe collinearity. Thus, the corresponding columns were removed to form a reduced data matrix.

Six classification techniques were used: a principal component (empirical orthogonal) analysis–logistic-regression approach (PCA&LR), classification and regression trees (CART), random forests (RF), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and logistic regression (LR). Four measures of prediction skill were considered: hit rate (HR), false-alarm ratio (FAR), Brier (1950) score (BS), and (for LR) the area under the receiver-operating-characteristic curve (AUC). For a perfect classification, HR = 1, FAR = 0, BS = 0, and AUC = 1. Values of HR near 0, FAR and BS values near 1, and AUC values near 0.5 indicate poor performance. Further details on the classifiers and the receiver-operating-characteristic curve can be found in appendix B and other sources (see, e.g., Breiman 2001; Venables and Ripley 2002; Hilbe 2009). Tenfold cross validation was used to assess how well the classifiers performed on an independent dataset. All analyses

TABLE 2. Abbreviations, full names, units of measure, and specifications for atmospheric fields. This is the same set of fields that was considered by Bates et al. (2017, their Table 2).

| Abbreviation | Full name | Specification |
|---|---|---|
| Instability and lifting potential | | |
| CAPE | Convective available potential energy ($J\,kg^{-1}$) | As provided in ERA-Interim (max CAPE on the basis of lifting parcels within a near-surface layer) |
| CBH | Cloud-base height (m) | From temperature and dewpoint at a height of 2 m with lifting to condensation level using an idealized constant lapse rate |
| CMF | Convective mass flux ($Pa^2\,s^{-1}\,K^{-1}$) | 500 hPa: calculated as the product of air density, fraction of grid points covered by updrafts within the $7 \times 7$ gridded region, and the vertical velocity averaged across all updrafts |
| CONV1000850 | Mean low-level horizontal wind convergence ($s^{-1}$) | Mean value at 850 and 1000 hPa pressure levels |
| DD | Dewpoint depression (°C) | 500, 700, and 850 hPa |
| DDIV | Density-weighted mean upper-level divergence minus density-weighted mean low-level divergence ($s^{-1}$) | {300, 400} − {850, 1000} hPa |
| EPTL | Mean low-level equivalent potential temperature minus mean midlevel equivalent potential temperature (°C) | Mean value at 1000 and 850 hPa − mean value at 700 and 500 hPa |
| TD850T500 | Cross totals index (°C) | 850 and 500 hPa |
| TGD | Direction of thickness gradient (rad) | {500, 700}, {500, 1000}, and {700, 1000} hPa |
| TGM | Magnitude of thickness gradient ($m^2\,s^{-2}$) | {500, 700}, {500, 1000}, and {700, 1000} hPa |
| THETA_W1000 | Wet-bulb potential temperature (°C) | 1000 hPa |
| THETA_W850500 | Wet-bulb potential temperature diff (°C) | 850 − 500 hPa |
| THK7001000 | Geopotential thickness ($m^2\,s^{-2}$) | 700 − 1000 hPa geopotential heights |
| TL850500 | Temperature lapse (°C) | 850 − 500 hPa |
| TL850700 | Temperature lapse (°C) | 850 − 700 hPa |
| TTI | Total totals index (°C) | 850 and 500 hPa |
| *W* | Vertical velocity ($Pa\,s^{-1}$) | 200, 300, 500, 700, 850, and 1000 hPa |
| Atmospheric water content | | |
| CONVP | Convective precipitation (m) | As provided in ERA-Interim |
| ICE | Total column ice water ($kg\,m^{-2}$) | As provided in ERA-Interim |
| SH | Specific humidity ($kg\,kg^{-1}$) | 500, 700, and 850 hPa |
| TCWV | Total column water vapor ($kg\,m^{-2}$) | As provided in ERA-Interim |
| TOTP | Total precipitation (m) | As provided in ERA-Interim |
| Wind speed | | |
| MVWS | Max vertical wind shear ($m\,s^{-1}$) | From 300 to 850 hPa |
| S06 | Vertical wind shear between 0 and 6 km ($m\,s^{-1}$) | 1000 and 500 hPa |
| *U* | Zonal wind velocity ($m\,s^{-1}$) | 300, 500, 700, 850, and 1000 hPa |
| *V* | Meridional wind velocity ($m\,s^{-1}$) | 300, 500, 700, 850, and 1000 hPa |
| General atmospheric state and variability | | |
| SEASON | Season of year | DJF, MAM, JJA, and SON |
| *T* | Air temperature (°C) | 2 m and 500, 700, and 850 hPa |
| MSLP | Mean sea level pressure (Pa) | As provided in ERA-Interim |
| GPH | Geopotential height ($m^2\,s^{-2}$) | 500 and 700 hPa |
| MING | Min geostrophic vorticity ($s^{-2}$) | Laplacian of geopotential at 500, 700, and 850 hPa |

were carried out in the R computing environment (https://www.r-project.org/).

## 3. Results

The proportions of lightning days for the CIGRE 500 sites are displayed in Fig. 1. The proportions range from 0.06 (Perth, Western Australia, Australia) to 0.43 (Darwin, Northern Territory, Australia). Median adjusted $R^2$ values for the fitted quadratic surfaces varied across atmospheric fields and sites, with 8.1%–16% below 0.5 and 49%–70% above 0.75. Thus, the surfaces gave a reasonable representation of the main features of the fields. For PCA&LR, perusal of comparative box plots revealed that only the first principal component had any predictive power in terms of discriminating between lightning and nonlightning days. Therefore, the remaining principal components were not considered further.

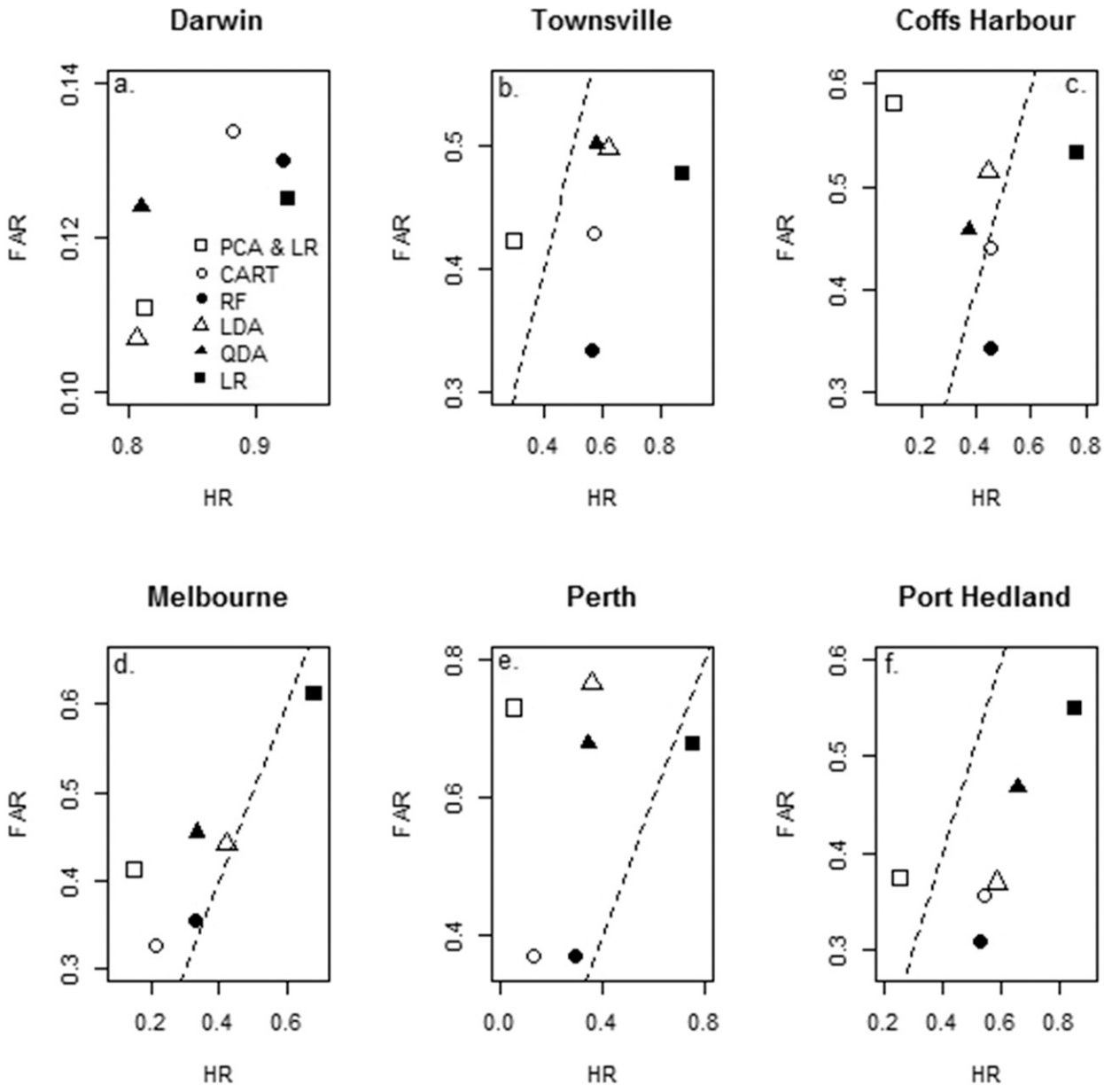


FIG. 2. Cross-validated prediction skill for six classifiers, with line of equality (dashed line).

Summaries of the cross-validation results for lightning days appear in Figs. 2 and 3. Every classifier considered performed well for Darwin: HRs are noticeably greater than FARs and the BSs are low (0.09–0.13). The RF and LR methods produced the highest HRs (0.92 and 0.93, respectively). For the five remaining sites, PCA&LR is the worst-performing classifier, with FAR > HR in every case. The LR method produced the highest HRs across these sites and, for Melbourne and Perth, the only cases in which HR > FAR. Across all sites, AUC values ranged from 0.80 (Melbourne) to 0.96 (Darwin). They indicate a prediction skill that is much better than use of climatological values (known as ''climatology''; AUC = 0.5). Perusal of Fig. 3 reveals that Coffs Harbour (in New South Wales) and Melbourne, both in the vicinity of the extratropical east coast region of Australia, have the highest BSs ($0.15 < BS \leq 0.24$). This is primarily due to low HRs for both lightning and nonlightning days and high FARs for nonlightning days. The relatively low prediction skill in this region could relate to the fact that lightning activity is sometimes associated with other synoptic-scale systems, such as subtropical cyclones

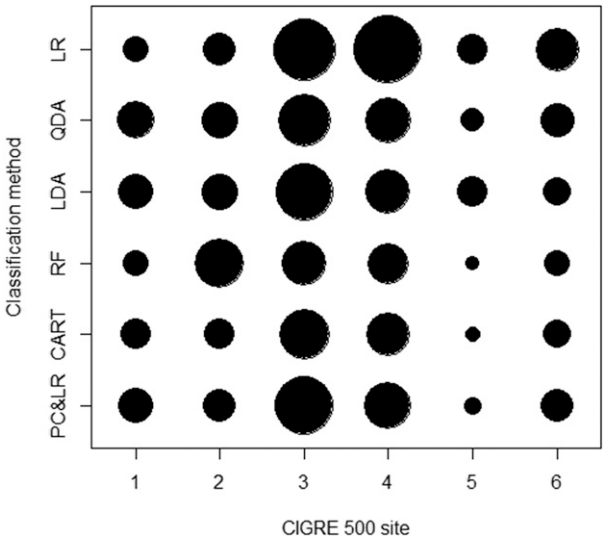


FIG. 3. Bubble chart of Brier scores for six classification methods and six CIGRE 500 sites. The largest bubble diameter corresponds to a Brier score of 0.241, and the smallest bubble corresponds to 0.052. The key to the site numbers appears in Fig. 1.

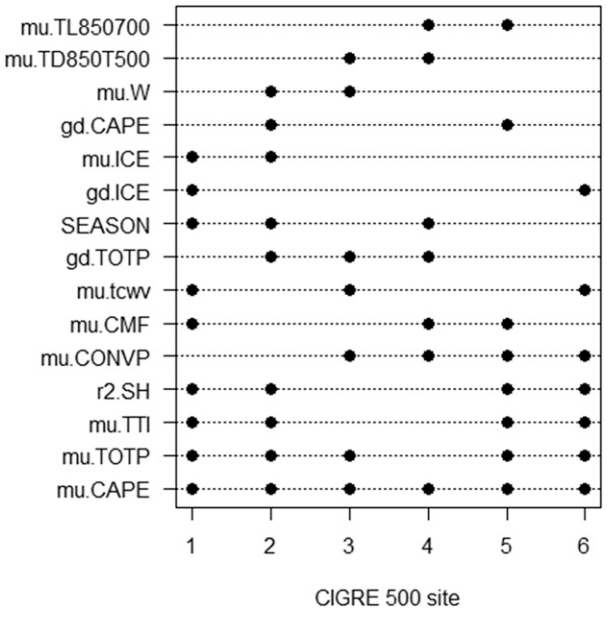


FIG. 4. Dot chart of important atmospheric variables for each final LR model for each of the six CIGRE 500 sites. Variables were declared important if they appeared in at least two models. The key to site numbers appears in Fig. 1.

known as "east coast lows" (Chambers et al. 2014; Dowdy and Kuleshov 2014). Although these synoptic-scale systems may exert some localized control on atmospheric conditions associated with deep convection (such as in relation to low-level convergence and advection of moisture), the cyclones in this region may not be well represented by the specific set of 0600 UTC thunderstorm and convection measures that are considered in this study. Furthermore, an improved understanding of the factors controlling cyclone activity in this region, as well as associated local severe-weather conditions, is an area of active research. Research topics include improved understanding of the region's uniqueness in terms of the hybrid energetics characteristics of these cyclones (Pezza et al. 2014); the regional features of its lightning climatological behavior (Dowdy and Kuleshov 2014); and the interrelationships among cyclones, fronts, thunderstorms/lightning activity, and local atmospheric conditions (Dowdy and Catto 2017). Across the six classifiers, the lowest BSs were obtained for Perth ($0.052 < BS < 0.11$). This result is because of the high HR (0.90) and low FAR (0.017) obtained for nonlightning days and the low number of lightning days for that site.

A dot chart of the 15 variables included in the six final LR models is displayed in Fig. 4, and a key to the atmospheric fields that were used appears in Table 2. The plot and table reveal five key features. First, 10 variables are spatial mean measures of instability and lifting potential and of atmospheric water content. Second, only mu.CAPE appears in all six models, mu.TOTP appears in five, and mu.TTI, r2.SH, and mu.CONVP appear in four. Third, variables representing wind shear are not included in the LR models for these locations. Fourth, the LR models for Darwin and Townsville include mu.CAPE, mu.TOTP, mu.TTI, and r2.SH as variables whereas the models for Perth and Port Hedland (in Western Australia) include those variables as well as mu.CONVP. Fifth, despite the above similarities, the general pattern of scatter in the dot chart suggests that the optimal variable sets for the classification of lightning days may vary among different climatic zones.

The dominance of the spatial mean measures could be related to temporal variations in the timing of thunderstorms with respect to a given location since lightning can, at times, occur during hours other than late afternoon. Although CAPE and TTI are well known measures of storminess, including for Australian conditions (Hanstrum et al. 2002; Niall and Walsh 2005; Allen et al. 2011; Dowdy 2015), the prominence of mu.TOTP and mu.CONVP might reflect the strong relationship between convective parameterizations and dynamic variables such as moisture convergence and vertical velocity *W* at midlevels that was found by Davies et al. (2013). It is also noted that measures that are based on precipitation in combination with CAPE have also been shown to provide a good indication of lightning-flash density in other regions of the world, such as the United

States (Romps et al. 2014). Perusal of the comparative box plots for r2.SH revealed that high values (>0.9) are associated with lightning days at Darwin, Townsville, Perth, and Port Hedland. These days are associated with noticeably high values of mu.SH and low values of vg.SH. Combined, these conditions indicate the presence of high levels of atmospheric moisture content near the surface (as indicated by r2.SH) with little change with height (as indicated by vg.SH). A high concentration of moisture throughout a range of levels in the lower troposphere could help to lead to enhanced moist convection over a considerable depth of the atmosphere. Given the important role of entrainment and detrainment in cumulus convection (De Rooy et al. 2013), these conditions are likely to be conducive to latent heat release (from condensation and/or freezing processes) acting to enhance potential updraft strengths. Although there are considerable uncertainties and complexities around the microphysical processes and combination of physical factors associated with lightning generation, such as the role of aerosols as potential cloud condensation nuclei in processes leading to lightning occurrence (Stolz et al. 2017; Thornton et al. 2017), it is widely accepted from a thermodynamic perspective that strong updraft speed (i.e., kinetic energy) is needed in regions of the cloud where ice is present to help to produce charge separation and the associated high potential differences that are required for atmospheric electrical breakdown (Rakov and Uman 2003).

Two variables that represent vertical wind shear were considered as candidate variables in the analysis (Table 2). Although some previous studies, such as Allen et al. (2011), have used variables for vertical wind shear within large-scale indicators of thunderstorm characteristics in Australia, one plausible explanation for the lack of wind-shear variables in the final LR models for lightning is that it is due to differences in the thunderstorm characteristics under study. For example, Allen et al. (2011) focused on severe-thunderstorm characteristics such as hail, tornados, and extreme winds and rainfall rather than on lightning occurrence.

## 4. Conclusions

The following key results were found for six different locations in Australia:

1) Low-dimensional summary statistics capture useful information about the structure of thunderstorms at coarse spatial and temporal scales. This result is consistent with the finding of Bates et al. (2017) that the use of LDSS adds value to the discrimination of dry and wet thunderstorms.
2) The overall performance of logistic regression was superior to that of the other classifiers considered.
3) The prediction skill of the LR was found to be much better than use of climatology.
4) The variables associated with the final LR models are dominated by spatial mean measures of instability and lifting potential and of atmospheric water content (10 of 15 variables). This dominance might be related to temporal variations in the timing of thunderstorms at a given location.
5) Although the same set of atmospheric variables was used for each CIGRE 500 site, the variables in the final LR models varied across climatic zones. The issue of whether it is possible to use the same variables and same classification method at different sites within a single climatic zone would be a fertile area for future research.

It is envisaged that the combined LDSS–LR approach advocated in this study will find application as finer-scale reanalyses and GCM runs become available. Such work might lead to results that are less dependent on climate-model parameterizations (such as mu.TOTP and mu.CONVP).

*Acknowledgments.* Lightning data were provided by the Observations and Engineering Branch of the Australian Bureau of Meteorology. The National Environmental Science Programme provided partial financial support for author AJD. All data used in this work are available on request from Andrew J. Dowdy (a.dowdy@bom.gov.au). We also thank three anonymous referees and the editor of this journal, Andrew Ellis, for their thoughtful and constructive comments on the original typescript.

## APPENDIX A

### Representation of Atmospheric Variables

The material covered here and in appendix B is taken from Bates et al. (2017) with minor modification and is intended to provide sufficient methodological detail to allow readers to decide whether they need to read further. Most of the information on the daily atmospheric variables that are used herein is available at a single pressure level or is defined as a mean or difference for fixed pressure levels and hence can be considered as a function of two spatial dimensions: $z = f(x, y)$. An exception is convective mass flux (CMF), which, by definition, has a constant value across all 49 reanalysis grid points for a given day and UTC time. Other variables such as air temperature, minimum geostrophic vorticity,

vertical velocity, specific humidity, and zonal and meridional wind are defined for specific atmospheric pressure levels $p$ at each grid point (Table 2). These variables can be considered as a function of three spatial dimensions: $z = f(x, y, p)$. For each day, quadratic surfaces were fitted to the atmospheric fields for 0600 UTC using ordinary least squares. A quadratic surface in two spatial dimensions is defined by

$$z = f(x,y) = c_1 + c_2x + c_3x^2 + c_4y + c_5xy + c_6y^2, \quad \text{(A1)}$$

and the corresponding surface in three spatial dimensions is defined by

$$z = f(x,y,p) = c_1 + c_2x + c_3x^2 + c_4y + c_5xy + c_6y^2 + c_7p + c_8xp + c_9yp + c_{10}p^2. \quad \text{(A2)}$$

Instead of fitting Eqs. (A1) and (A2) directly, the linear and quadratic terms were replaced by orthogonal polynomials to ensure that the intercept and linear and quadratic regression coefficients are independent of each other (i.e., they do not change when higher-order terms are added), and the estimates of the intercept and regression coefficients are placed on the same scale. Also, it allows the decomposition of relationships into general components of magnitude as well as into linear and nonlinear rates of change. The estimates were calculated in a coordinate system that was centered on the CIGRE 500 sensor (i.e., the $7 \times 7$ grid described in section 2). The adjusted $R^2$ was used as a goodness-of-fit measure for the quadratic surfaces and a measure of spatial (dis)organization in the atmospheric fields.

Let $\theta_1, \ldots, \theta_{10}$ denote the orthogonal polynomial regression coefficients. Six LDSS for the above surfaces were used to facilitate physical interpretation: the intercept, which is equivalent to the mean across the domain (mu = $\theta_1$); the magnitude of the gradient vector (gd) and its direction (dr) in the $x$–$y$ plane; Gaussian curvature (gc); vertical gradient (vg = $\theta_7$); and adjusted $R^2$ (r2) since it is a measure of spatial (dis)organization in the atmospheric fields. The magnitude of the gradient vector and its direction in terms of linear rate of change are defined by $\text{gd} = (\theta_2^2 + \theta_4^2)^{1/2}$ and $\text{dr} = \tan^{-1}(\theta_4/\theta_2)$. Given the use of orthogonal polynomial regression, the values of gd and dr are the same as those that would have been obtained had a linear surface been fitted to the data. Gaussian curvature is an intrinsic geometric property of a surface that is independent of the coordinate system that is used to describe it. It is defined by

$$\text{gc} = \det(\mathbf{H}) = \lambda_1\lambda_2, \quad \text{(A3)}$$

where det () denotes the determinant, **H** is the Hessian matrix given by

$$\mathbf{H} = \begin{pmatrix} \dfrac{\partial^2 z}{\partial x^2} & \dfrac{\partial^2 z}{\partial x \partial y} \\ \dfrac{\partial^2 z}{\partial y \partial x} & \dfrac{\partial^2 z}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 2\theta_3 & \theta_5 \\ \theta_5 & 2\theta_6 \end{pmatrix}, \quad \text{(A4)}$$

and $\lambda_1$ and $\lambda_2$ are the eigenvalues of **H** (and also the maximum and minimum principal curvatures).

## APPENDIX B

### Statistical and Machine-Learning Classification Techniques

The first classification technique used in this study involves dimensional reduction using PCA and classification with logistic regression. PCA uses an orthogonal transformation to convert an original set of possibly correlated variables into a new set of mutually uncorrelated variables that are arranged in decreasing order of importance. The first principal component is the linear combination of the original variables that captures as much of the variation in the original dataset as possible. The second component captures the maximum variability that is uncorrelated with the first component, and so on. PCA provides a useful reduction in complexity when a substantial proportion of the total variance in the data is accounted for by a few components. It is not in itself a classification technique.

The LR model can be written as

$$\text{logit}(\pi_i) = \ln[\pi_i/(1-\pi_i)] = \beta_0 + \sum_{j=1}^{p} \beta_j \mathbf{X}_j, \quad \text{(B1)}$$

where $\pi_i$ is the probability of occurrence of class $i$ ($i = 1, 2$), $\pi_i/(1-\pi_i)$ is the odds ratio for class $i$, $p$ is the number of columns in the data matrix **X**, and $\beta_0, \ldots, \beta_p$ are the regression coefficients, which are determined through maximum likelihood estimation. (It is obvious that with only two categories it is only necessary to estimate the coefficients for one of the categories since $\pi_2 = 1 - \pi_1$.) Classification on the basis of the variables is then done by setting a threshold $\tau$, say, and allocating a day to category 1 if $\pi_1 > \tau$. For each site, a receiver-operating-characteristic curve (a plot of HR vs FAR as the threshold $\tau$ is varied across its full range) was used to estimate the threshold by minimizing the distance from the curve to the point representing perfect classification accuracy (HR = 1; FAR = 0). This was done to account for the fact that the sample sizes for nonlightning and

lightning days were noticeably unequal at all sites (Table 1). Experiments using the Youden (1950) index indicated that threshold estimates were not sensitive to the selection technique that was used. With LR, by contrast with LDA and QDA, there is no formal requirement for multivariate normality of the explanatory variables within each category of the response variable, and the use of binary or categorical variables is acceptable. A combination of stepwise selection and analysis of deviance was used to determine the significance of variables in the LR models. All of the variables used in the final LR models are significant at the 0.05 level.

CART uses binary recursive partitioning to divide the data space, splitting it along the coordinate axes of the candidate variables to give increasingly homogenous subsets and hence the maximal separation of the classes until it is infeasible to continue. The measure of node heterogeneity is the deviance (a quality-of-fit statistic). The partitioning leads to a set of decision rules in the form of a binary tree. The tree is ''pruned'' to identify a parsimonious tree with acceptable misclassification rates. Cross validation can be used to determine an appropriate tree size.

The random-forests method is an ensemble learning algorithm that generates a large number of CART from bootstrap samples of the original data. An estimate of the misclassification rate can be obtained by using each tree to predict the data not in the bootstrap sample and averaging the predictions over all trees. Variable importance plots can be produced that reveal how important each variable is in classifying the data and contributing to the homogeneity of the nodes.

LDA is derived from an underlying model in which the distributions of the variables on dry and wet lightning days are both multivariate normal, with possibly different means and a common covariance matrix. LDA is somewhat robust with respect to minor violations of these assumptions. Although serious violations will often result in unreliable estimates of the coefficients, the procedure can still be a good heuristic. The discriminant function is a linear combination of the candidate variables, the coefficients of which are estimated by ordinary least squares so that the ratio of the between-classes variance and the within-classes variance is maximized. This function takes the value zero at the decision boundary. If the value of the discriminant function is negative, the variable vector is assigned to one class; if it is positive, the variable vector is assigned to the other class. Given that the variables are standardized, the coefficients indicate the relative importance of each variable in predicting class assignment.

Quadratic discriminant analysis is a generalization of LDA in which the two classes need not have the same covariance matrix, but the assumption of multivariate normality still applies. The interpretation of the coefficients in terms of the relative importance of each variable is more difficult to assess than for LDA as the discriminant function contains quadratic as well as linear and constant terms.

## REFERENCES


Allen, J. T., D. J. Karoly, and G. A. Mills, 2011: A severe thunderstorm climatology for Australia and associated thunderstorm environments. *Aust. Meteor. Ocean J.*, **61**, 143–158, https://doi.org/10.22499/2.6103.001.

Bates, B. C., A. J. Dowdy, and R. E. Chandler, 2017: Classification of Australian thunderstorms using multivariate analyses of large-scale atmospheric variables. *J. Appl. Meteor. Climatol.*, **56**, 1921–1937, https://doi.org/10.1175/JAMC-D-16-0271.1.

Belsley, D. A., 1991: *Conditioning Diagnostics, Collinearity and Weak Data in Regression.* John Wiley and Sons, 396 pp.

Blouin, K. D., M. D. Flannigan, X. Wang, and B. Kochtubajda, 2016: Ensemble lightning prediction models for the province of Alberta, Canada. *Int. J. Wildland Fire*, **25**, 421–432, https://doi.org/10.1071/WF15111.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Brier, G. W, 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Burrows, W. R., C. Price, and L. J. Wilson, 2005: Warm season lightning probability prediction for Canada and the northern United States. *Wea. Forecasting*, **20**, 971–988, https://doi.org/10.1175/WAF895.1.

Chambers, C. R. S., G. B. Brassington, I. Simmonds, and K. Walsh, 2014: Precipitation changes due to the introduction of eddy-resolved sea surface temperatures into simulations of the ''Pasha Bulker'' Australian east coast low of June 2007. *Meteor. Atmos. Phys.*, **125**, 1–15, https://doi.org/10.1007/s00703-014-0318-4.

Christian, H. J., and Coauthors, 2003: Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. *J. Geophys. Res.*, **108**, 4005, https://doi.org/10.1029/2002JD002347.

Davies, L., C. Jakob, P. May, V. V. Kumar, and S. Xie, 2013: Relationships between the large-scale atmosphere and the small-scale convective state for Darwin, Australia. *J. Geophys. Res. Atmos.*, **118**, 11 534–11 545, https://doi.org/10.1002/jgrd.50645.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

De Rooy, W. C., and Coauthors, 2013: Entrainment and detrainment in cumulus convection: An overview. *Quart. J. Roy. Meteor. Soc.*, **139**, 1–19, https://doi.org/10.1002/qj.1959.

Dowdy, A. J., 2015: Large-scale modelling of environments favourable for dry lightning occurrence. *MODSIM2015, 21st International Congress on Modelling and Simulation*, T. Weber, M. J. McPhee, and R. S. Anderssen, Eds., Modelling and Simulation Society of Australia and New Zealand, 1524–1530.

——, and G. A. Mills, 2009: Atmospheric states associated with the ignition of lightning-attributed fires. Collaboration for Australian Weather and Climate Research Tech. Rep. 019, 34 pp, http://www.cawcr.gov.au/technical-reports/CTR_019.pdf.

——, and Y. Kuleshov, 2014: Climatology of lightning activity in Australia: Spatial and seasonal variability. *Aust. Meteor. Ocean J.*, **6**, 9–14.

——, and J. L. Catto, 2017: Extreme weather caused by concurrent cyclone, front and thunderstorm occurrences. *Sci. Rep.*, **7**, 40359, https://doi.org/10.1038/srep40359.

Hanstrum, B. N., G. A. Mills, A. Watson, J. P. Monteverdi, and C. A. Doswell III, 2002: The cool-season tornadoes of California and southern Australia. *Wea. Forecasting*, **17**, 705–722, https://doi.org/10.1175/1520-0434(2002)017<0705:TCSTOC>2.0.CO;2.

Hilbe, J. M., 2009: *Logistic Regression Models.* Chapman and Hall/CRC, 656 pp.

Lynn, B. H., Y. Yair, C. Price, G. Kelman, and A. J. Clark, 2012: Predicting cloud-to-ground and intracloud lightning in weather forecast models. *Wea. Forecasting*, **27**, 1470–1488, https://doi.org/10.1175/WAF-D-11-00144.1.

Mazany, R. A., S. Businger, S. I. Gutman, and W. Roeder, 2002: A lightning prediction index that utilizes GPS integrated precipitable water vapor. *Wea. Forecasting*, **17**, 1034–1047, https://doi.org/10.1175/1520-0434(2002)017<1034:ALPITU>2.0.CO;2.

Muñoz, Á. G., J. Díaz-Lobatón, X. Chourio, and M. J. Stock, 2016: Seasonal prediction of lightning activity in north western Venezuela: Large-scale versus local drivers. *Atmos. Res.*, **172–173**, 147–162, https://doi.org/10.1016/j.atmosres.2015.12.018.

Niall, S., and K. Walsh, 2005: The impact of climate change on hailstorms in southeastern Australia. *Int. J. Climatol.*, **25**, 1933–1952, https://doi.org/10.1002/joc.1233.

Pezza, A. B., L. A. Garde, A. P. Veiga, and I. Simmonds, 2014: Large scale features and energetics of the hybrid subtropical low 'Duck' over the Tasman Sea. *Climate Dyn.*, **42**, 453–466, https://doi.org/10.1007/s00382-013-1688-x.

Rakov, V. A., and M. A. Uman, 2003: *Lightning: Physics and Effects.* Cambridge University Press, 687 pp.

Romps, D. M., J. T. Seeley, D. Vollaro, and J. Molinari, 2014: Projected increase in lightning strikes in the United States due to global warming. *Science*, **346**, 851–854, https://doi.org/10.1126/science.1259100.

Sánchez, J. L., R. Fraile, M. T. de la Fuente, and J. L. Marcos, 1998: Discriminant analysis applied to the forecasting of thunderstorms. *Meteor. Atmos. Phys.*, **68**, 187–195, https://doi.org/10.1007/BF01030210.

Sousa, J. F., M. Fragoso, S. Mendes, J. Corte-Real, and J. A. Santos, 2013: Statistical–dynamical modeling of the cloud-to-ground lightning activity in Portugal. *Atmos. Res.*, **132–133**, 46–64, https://doi.org/10.1016/j.atmosres.2013.04.010.

Stolz, D. C., S. A. Rutledge, J. R. Pierce, and S. C. van den Heever, 2017: A global lightning parameterization based on statistical relationships among environmental factors, aerosols, and convective clouds in the TRMM climatology. *J. Geophys. Res.*, **122**, 7461–7492, https://doi.org/10.1002/2016JD026220.

Thornton, J. A., K. S. Virts, R. H. Holzworth, and T. P. Mitchell, 2017: Lightning enhancement over major oceanic shipping lanes. *Geophys. Res. Lett.*, **44**, 9102–9111, https://doi.org/10.1002/2017GL074982.

Venables, W. N., and B. D. Ripley, 2002: *Modern Applied Statistics with S.* 4th ed. Springer, 495 pp.

Yair, Y., B. Lynn, C. Price, V. Kotroni, K. Lagouvardos, E. Morin, A. Mugnai, and M. del Carmen Llasat, 2010: Predicting the potential for lightning activity in Mediterranean storms based on the Weather Research and Forecasting (WRF) Model dynamic and microphysical fields. *J. Geophys. Res.*, **115**, D04205, https://doi.org/10.1029/2008JD010868.

Youden, W. J., , 1950: Index for rating diagnostic tests. *Cancer*, **3**, 32–35, https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

Zepka, G. S., O. Pinto Jr., and A. C. V. Saraiva, 2014: Lightning forecasting in southern Brazil using the WRF Model. *Atmos. Res.*, **135–136**, 344–362, https://doi.org/10.1016/j.atmosres.2013.01.008.