The DOI for this manuscript is doi: 10.1175/JAMC-D-16-0271.1

The final published version of this manuscript will replace the preliminary version at the above DOI once it is available.

If you would like to cite this EOR in a separate work, please use the following full citation:

Bates, B., A. Dowdy, and R. Chandler, 2017: Classification of Australian Thunderstorms using Multivariate Analyses of Large-Scale Atmospheric Variables. J. Appl. Meteor. Climatol. doi:10.1175/JAMC-D-16-0271.1, in press.

Classification of Australian Thunderstorms using Multivariate Analyses of Large-Scale Atmospheric Variables

BRYSON C. BATES

CSIRO Oceans and Atmosphere, Wembley, Western Australia, Australia

School of Earth 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, UK

ABSTRACT

Lightning accompanied by inconsequential rainfall (i.e. 'dry' lightning) is the primary natural ignition source for wildfires globally. This paper presents a machine-learning and statistical-classification analysis of 'dry' and 'wet' thunderstorm days in relation to associated atmospheric conditions. The study is based on daily lightning flash count and precipitation data from ground-based sensors and gauges, and a comprehensive set of atmospheric variables based on the ERA-Interim reanalysis for the period from 2004 to 2013 at six locations in Australia. These locations represent a wide range of climatic zones (temperate, subtropical to tropical). Quadratic surface representations and low-dimensional summary statistics were used to characterize the main features of the atmospheric fields. Four prediction skill scores were considered and ten-fold cross validation used to evaluate the

*Corresponding author (Email: andrew.dowdy@bom.gov.au; Tel: + 61 3 9669 4722)

performance of each classifier. The results were compared with those obtained by adopting the approach used in an earlier study for the Pacific Northwest, United States. It was found that: both approaches have prediction skill when tested against independent data, mean atmospheric field quantities proved to be the most influential variables in determining dry lightning activity and no single classifier or set of atmospheric variables proved to be consistently superior to their counterparts for the six sites examined here.

1. Introduction

Although human-caused wildfire ignitions are common in many regions of the world, particularly in densely populated areas, fires ignited by lightning typically burn a larger area than fires ignited by other sources. This is attributable to lightning occurrence in remote locations and in large spatial and temporal clusters which hamper the response efforts of fire management authorities (USDA Forest Service 1992; McRae 1992; Vazquez and Moreno 1998; Wotton et al. 2005; Wotton and Martell 2005; Kasischke et al. 2006; Dowdy and Mills 2012a). Lightning that occurs with relatively little precipitation (i.e., ‘dry’ lightning) has a higher chance of igniting a fire than lightning accompanied by heavier precipitation (‘wet’ lightning) (Rothermel, 1972; Wotton and Martell 2005; Dowdy and Mills 2012a). Therefore, an improved understanding of dry lightning activity and the atmospheric conditions that influence its occurrence is of importance for better preparedness and enhancing the ability to respond to the impacts associated with wildfires ignited by lightning.

There are many physical factors that can influence lightning occurrence as demonstrated in numerous previous climatological, dynamical modeling and seasonal prediction studies including Weisman and Klemp (1982), Goodman et al. (2000), Burrows et al. (2005), Williams et al. (2005), Deierling et al. (2008), Romero et al. (2007), Chronis et al. (2008), Dai et al. (2009), Barthe et al. (2010), Romps et al. (2014), Magi (2015), Dowdy (2016), Muñoz et al. (2016) and references therein. Although aspects of the microphysical processes associated with lightning generation are not well understood in some cases, the role of ice in facilitating charge separation within the cloud appears to be a critical factor in determining whether or not lightning is produced (e.g., as indicated by laboratory experiments (Takahashi and Miyawaki 2002) as well as observations (Lang et al. 2014)). Microphysical processes such as ice formation are not well represented at the spatial and temporal scales of currently available climate models and reanalyses, leading to the use of parameterization schemes for a range of variables associated with convection. For example, the ERA-Interim reanalysis (Dee et al. 2011) uses a convective parameterization based on a bulk mass flux scheme (as originally described by Tiedtke 1989), with parameterizations also used to represent the fallout of precipitation (e.g., Kuo and Raymond 1980) and factors such as virga (streaks of water or ice particles that vaporize before reaching the ground) considered.

In addition to convective parameterization schemes, several studies have demonstrated that statistical indicators of lightning activity can be found at relatively coarse spatial and temporal scales (e.g., similar to the resolution of general circulation models (GCMs) and reanalyses). For example, Romps et al. (2014) combined precipitation and Convective Available Potential Energy (CAPE) based on GCM output for use as an indicator of environments conducive to lightning activity, applying this indicator to examine the influence of global warming on lightning strikes in the United States. A recent study based on reanalyses demonstrated that even at spatial resolutions of 7.5° in latitude and longitude, atmospheric conditions such as lower-tropospheric moisture content, temperature lapse rate and CAPE can be strongly related to lightning activity (Dowdy 2016).

In contrast to the number of studies that have examined atmospheric conditions associated with lightning activity in general, relatively few studies have focused specifically on dry lightning. Notable early studies include Rorig and Ferguson (1999, hereafter designated as

RF99) and Rorig et al. (2007), demonstrating that a linear discriminant rule could separate dry and wet lightning classes. The rule was composed of dewpoint depression at 850 hPa (DD850) and temperature lapse from 850 to 500 hPa (TL850500), with dry lightning defined as lightning accompanied by precipitation of less than one tenth of an inch (about 2.5 mm). Dowdy and Mills (2012b) demonstrated that these two variables were also applicable in southeast Australia, and that the average chance of a sustained fire ignition resulting from the occurrence of lightning in that region is higher than average if the precipitation accompanying the lightning is less than about 2 to 3 mm. Recent studies have examined a somewhat wider range of variables in relation to the occurrence of dry lightning, including studies in North America (Wallmann et al. 2010; Nauslar et al. 2013; Abatzoglou et al. 2016) and Australia (Dowdy 2015), finding that some useful skill can be obtained for predicting the occurrence of dry lightning based on several different methods. However, as dry lightning activity remains relatively unstudied when compared with other aspects of thunderstorm activity and associated convective processes, to date there have been no climatological studies of the spatial and temporal variability of dry lightning activity, or the influence of large-scale atmospheric drivers of dry lightning variability.

The approach presented in this paper (hereafter designated as BDC) represents a more general approach to the two-category classification problem of dry and wet lightning days than that of RF99. The paper has four objectives with a view to building on previous studies of dry lightning occurrence. The first is to consider a wider range of atmospheric conditions associated with dry lightning activity and precipitation occurrence than has been the case to date. The second is to build on the suggestion put forward by Blouin et al. (2016) that a comparison of classification methods (classifiers) may provide useful guidance for future research. The third is to consider lightning, precipitation and atmospheric data from a wide range of climatic zones. The fourth objective is to identify a subset of influential atmospheric variables across climatic zones and different classifiers. The method of RF99 is used as a benchmark for assessments of prediction accuracy and the applicability of the new approach proposed in this paper. In this way, the paper provides a useful addition to the toolkit for addressing questions related to lightning activity. The paper is divided into five sections. Section 2 provides a description of the study sites, data and classifiers used. Results are presented in Section 3. A summary and conclusions are given in Section 4. The quadratic surface representations and low-dimensional summary statistics (LDSS) used to characterize the main features of the atmospheric fields considered in this study are described in the Appendix.

2. Data and methods

a. Description of study sites and data

The description of the daily lightning flash count datasets used herein parallels that of Bates et al. (2015), and the text in the next two paragraphs is derived from there with minor modifications. The data were collected from ground-based CIGRE 500 (Comité Internationale des Grands Réseaux Electriques, 500 Hz peak transmission filter circuit) sensors located at six weather stations operated by the Australian Bureau of Meteorology (Figure 1 and Table 1). The sensors were selected because of their record length and quality, and their locations in a variety of climatic settings including temperate, subtropical and tropical sites. The records cover the period from January 2004 to at least December 2010 (Townsville) and at most February 2013 (Melbourne).

< Insert Figure 1 and Table 1 about here >

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. We considered the total number of lightning flash counts since the CIGRE 500 sensor did not distinguish between intracloud and cloud-to ground flash counts; and the ratio of intracloud to cloud-to-ground flashes can vary significantly depending on thunderstorm type and intensity, region of occurrence and season (Rakov and Uman, 2003). Estimates of the effective horizontal ranges of the sensor are 30 km for cloud-to-ground flashes and 15 km for cloud-to-cloud flashes (Kuleshov and Jayaratne, 2004). As with other studies of this nature, these effective ranges should be taken into consideration when interpreting results for specific purposes such as fire ignition from cloud-to-ground lightning flashes. The electromechanical counters attached to the CIGRE 500 sensors were read manually each day between 0800 and 0900 h local time. Further details can be found in Jayaratne and Kuleshov (2006), Kuleshov et al. (2009) and Bates et al. (2015).

For a given weather station, thunderstorms were deemed to have occurred during a 24-h period if at least one lightning flash count was registered by the CIGRE 500 sensor. They were categorized as either ‘dry’ or ‘wet’ according to the concurrent daily precipitation reading recorded by the storage gauge at the station. A thunderstorm was classified as dry if the precipitation reading was less than 2.5 mm or wet otherwise. In Australia, daily precipitation is nominally measured each day at 0900 h local time. Station data were obtained from the SILO patch-point data set (Australian Bureau of Meteorology). There is a large disparity in spatial scales between the detection range of the sensor and the diameter of a precipitation gauge (15-30 km versus 203 mm). Thus it is possible for precipitation amounts greater than 2.5 mm to occur within the sensor’s detection limit but away from the station gauge. However, the use of gridded station data has its own set of limitations in that the interpolation involved is a form of smoothing that reduces precipitation variability. Thus, the process of gridding can considerably increase (decrease) the frequency of low (high) precipitation amounts (Ensor and Robeson, 2008), and this might have implications for the classification of dry thunderstorms. The reduction in variability is dependent on the distance from a grid point to the nearest gauge. A further concern is the relatively low density of precipitation gauge networks in Australia. For example, the numbers of gauges within a 30 km radius of the Darwin, Townsville, Coffs Harbour and Port Hedland sites are 12, 8, 18 and 3, respectively. This low network density is likely to lead to excessive smoothing in some instances and affect the distribution of daily precipitation amounts. Given the above, days with station precipitation values flagged as interpolated were discarded.

With future applications in mind, the study was designed to be conducted at spatial and temporal resolutions similar to that of current general circulation models and reanalyses. Atmospheric information was obtained from the ERA-Interim reanalysis archive (Dee et al. 2011). The spatial and temporal resolution of the dataset used is 0.75 degrees (in both latitude and longitude) and 6 hours, respectively. For each CIGRE 500 site, atmospheric data were extracted for the 49 reanalysis grid points closest to the sensor’s location. The aim of the grid was to capture the presence of a thunderstorm over or in the proximity of a sensor. The lightning and precipitation series were synchronized with the ERA-Interim series for 0600 UTC (1600 h Eastern Australia Time) within the 24-hour period represented by the lightning and precipitation data. This is because the diurnal variation in temperature lapse rate over land, due to solar radiation, produces conditions that are more favorable for lightning activity to occur during the late-afternoon period in general than at other times of the day or night (Christian et al. 2003; Dowdy and Mills 2009; Allen et al. 2011). Thus, the synchronization ensures that the atmospheric variables for each daily lightning flash count correspond to the time at which the lightning is most likely to have occurred. Since the use of a single time point can be viewed as reductive, the possibility that atmospheric variables at other times of day may also be relevant was considered. However, the additional information was found to be largely redundant because correlations within a 24-hour period are invariably high (e.g. correlations between individual variables at 0600 UTC and 1200 UTC, spanning the time period during which most deep convective processes occur in Australia, are typically greater than 0.95 and greater than 0.85 in every case examined). The atmospheric variables considered herein are listed in Table 2.

< Insert Table 2 about here >

The set of atmospheric variables examined here represents a wider range than has typically been examined in previous studies, particularly those studies focused on climate-scale analyses rather than finer-resolution numerical weather prediction or radar observational studies. This is because there have been very few studies that have specifically examined dry lightning activity and the atmospheric conditions that influence its occurrence. Consequently, the literature on dry and general lightning activity was combed and physical understanding used to reduce the number of variables as far as possible. The variables listed in Table 2 represent a broad variety of physical processes that can be associated with deep convection, including both dynamical and thermodynamical processes. The variables comprise various measures of temperature lapse, moisture content, vertical motion and water phase state, including at a range of different pressure levels (to allow potential variations in height between dry and wet thunderstorm characteristics to be distinguished).

b. Variable selection

To identify the dominant large-scale controls on lightning activity from among the variables listed in Table 2 is a challenging statistical problem: there are dependencies among the variables leading to collinearity and, moreover, the processes controlling lightning activity are complex so that the variables must be considered concurrently rather than in isolation. Regression-based approaches, notably those based on generalized linear models, are ideally suited to this kind of problem (e.g. Yan et al. 2002, Chandler 2005). However, an additional complication in the present application is that the explanatory variables are spatial fields over a 7×7 grid, rather than individual values. In principle, this can be handled using modern statistical techniques such as functional regression (e.g. Morris, 2015). However, in their current state of development such methods are most effective when the number of candidate variables is relatively small. The current state of knowledge is insufficient to identify a small number of candidate variables from the list in Table 2 with high confidence. The strategy adopted here is therefore to use a combination of approaches that are designed to isolate the most influential variables from many candidates.

To handle the spatial nature of the atmospheric variables listed in Table 2, the daily fields for each variable were reduced to a set of five LDSS designed to capture the main synoptic features. This was done by fitting quadratic surfaces to each daily field (see Appendix) and using the fitted surfaces to derive physically-interpretable daily summaries (overall means, vertical and horizontal gradients, and curvature). Note that the intention is not to provide highly accurate descriptions of the fields, but rather to provide indices that broadly describe the synoptic structure. The use of LDSS reduces the dimensionality of the problem from 49 grid point values per atmospheric variable per day to 5. Other dimension reduction techniques are available, notably principal component (empirical orthogonal) analysis which was explored as an alternative to the LDSS considered here. It was found that five or more components were necessary to explain 70 to 80% of the variance for each data set. Only the first component had any predictive power in terms of discriminating between dry and wet lightning. Although the loadings for this component would often indicate a contrast between two sets of variables, a defensible interpretation of the contrast proved elusive. Moreover, its predictive skill was lower than that obtained with LDSS.

At this point, a LDSS of an atmospheric variable will be referred to as a potential candidate variable. As an initial screening procedure, for each potential candidate variable, comparative boxplots of each LDSS were used to contrast its values for dry and wet lightning cases. Two variable selection criteria were considered. First, potential candidate variables where the 75^{th} (25^{th}) percentile for one lightning type was below (above) the 25^{th} (75^{th}) percentile for the other were deemed informative in terms of discriminating between dry and wet lightning days. These variables were reserved for further analysis. Second, depending on the number of such candidate variables found, they were supplemented by including additional candidate variables where the median in one lightning type was above the 75^{th} percentile or below the 25^{th} percentile of the other (see, e.g., Figure 2). The resulting candidate variables formed the columns of an atmospheric data matrix. This approach could be criticized as ad hoc: it is natural to ask whether alternative techniques, such as automatic variable selection procedures, would be preferable. The main reason for the approach taken here is that manual inspection of boxplots can provide checks on the data, as well as preliminary insights that may aid subsequent interpretation and that cannot be obtained from an automated analysis. In any case, the aim is merely to carry out a very preliminary screening of the data so as to focus subsequently on quantities that may have some predictive power in discriminating between dry and wet lightning.

< Insert Figure 2 about here >

Many of the candidate variables are measured on very different scales and thus are not commensurable in terms of magnitude or variability. This means that some variables could dominate or influence the results of the classification analysis because of their measurement units alone (Everitt and Hothorn, 2011). Thus, the columns of the data matrix were standardized to zero mean and unit variance prior to further analysis. This process places candidate variables on the same relative scale without disturbing the shape of the distribution of the data. It facilitates interpretation of the results of a discriminant or regression analysis, and helps to concentrate precisely on the conditions that are present during thunderstorms 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 environment (Hendrickx 2012; R Core Team 2015) 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 et al. (1980). 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 greater than 0.99 were considered sources of severe collinearity. Thus the corresponding columns were removed to form a reduced data matrix. A second proportion threshold of 0.8 was used to assess the degree of the sensitivity to threshold selection. It was found that the results obtained from the procedures described below showed only a slight sensitivity. Therefore, the results obtained using the proportion threshold of 0.8 will not be reported here.

c. Multivariate statistical analysis

Two machine-learning and three statistical methods were used for classification: classification and regression trees (CART); random forests (RF); linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and logistic regression (LR). Detailed descriptions of CART, RF and LR can be found in Faraway (2016), and LDA and QDA in Everitt and Dunn (2001). The R packages used in this work were: DiscriMiner (Sanchez 2013); MASS (Venables and Ripley, 2002); randomForest (Liaw and Wiener 2002); and tree

(Ripley, 2014). 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. RF is an ensemble learning algorithm which 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. The randomForest package can be used to produce variable importance plots which 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 positive it 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. QDA 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. The LR model can be written as

$$\mathrm{logit}(\pi_i) = \ln[\pi_i/(1-\pi_i)] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{1}$$

where π_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 and $\beta_0, \ldots, \beta_p$ are the regression coefficients which are determined via maximum likelihood estimation. (Obviously, 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 τ say, and allocating a day to category 1 if $\pi_1 > \tau$. For each site, a receiver operating characteristic (ROC) curve was used to select the threshold τ by minimizing the distance from the curve to the point representing perfect classification accuracy: this was done to account for the fact that the sample sizes for dry and wet lightning days were noticeably unequal for several sites (Table 1). Experiments using Youden's (1950) Index indicated that threshold estimates were not sensitive to the selection technique 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. Further details on the above classifiers can be found in Breiman (2001), Venables and Ripley (2002) and Hilbe (2009).

Although the approach of RF99 used only LDA to discriminate between dry and wet lightning, the four other classifiers considered herein (CART, RF, QDA and LR) were applied to the means of the DD850 and TL850500 fields (designated mu.DD850 and mu.TL850500, see Appendix) to ascertain whether a higher classification performance could be achieved. This extended approach is hereafter designated as E-RF99. The analysis was conducted in parallel with an identical study of a much larger set of variables to determine the extent to which it is possible to improve on the RF99 variable pair. Four measures of prediction skill were considered: the hit rate for dry lightning (HR), the false alarm ratio for dry lightning (FAR), the Brier (1950) score (BS) and for LR the area under the ROC (AUC). HRs, FARs, BSs, ROC curves and AUCs were determined using the verification package in R (NCAR 2015). For a perfect classification, HR=1, FAR=0, BS=0 and AUC=1. HR values near 0, FAR and BS values near 1, and AUC values near 0.5 indicate poor classification performance. For the convenience of the reader, in what follows, a list of the acronyms and abbreviations used in this paper and their meaning is given in Table 3.

< Insert Table 3 about here >

d. Cross validation experiments

Initial assessments of the prediction skill of the five classifiers (CART, RF, LDA, QDA and LR) were based on the data matrices for the six CIGRE 500 sites. As this can lead to optimistic bias in the estimated skill scores, ten-fold cross validation experiments were used to assess how well the results generalized to an independent dataset. Here the lightning, precipitation and candidate variable data were partitioned into ten subsamples of equal size. From these subsamples, a single subsample was retained for testing the model, and the remaining nine subsamples used for training (model fitting). The process is then repeated ten times with each of the subsamples used exactly once for validation. The R packages used in the cross validation experiments were cvTools (Alfons 2012) and verification (NCAR 2015).

3. Results

a. Preliminary analyses

Lightning activity at the six sites occurs primarily during the warmer months of the year (November to April). However, the most severe fire weather conditions in Australia occur at different times of the year, generally ranging from summer in the south to winter (i.e., the dry season) in the north. There are some regional variations to this, particularly along the eastern seaboard (including Coffs Harbour) where the peak fire weather conditions occur somewhat earlier (around October) than at other similar latitudes in Australia (Luke and McArthur, 1978). Further details on the lightning climatology of Australia may be found in Kuleshov et al. (2009), Dowdy and Kuleshov (2014), Bates et al. (2015) and references therein, and will not be repeated here.

The proportions of dry and wet lightning days for the six CIGRE 500 sites are reported in Table 1 and illustrated in Figure 1. Darwin is one of the most lightning prone areas in Australia. The number of lightning days for Darwin is markedly higher than those for the remaining sites, even for the case of Townsville which is in the same climatic zone. Perth has the lowest number of lightning days by a wide margin. Port Hedland has the highest proportion of dry lightning days, reflecting its desert environment. For the tropical and subtropical sites, the proportion of dry lightning days exceeds that of wet lightning days. This is somewhat surprising, and may in part be explained by the use of a single precipitation gauge to characterize rainfall over the detection range of the CIGRE 500 sensor (Section 2). To a lesser extent, it might reflect the effects of the precipitation threshold of 2.5 mm on lightning day classification. For example, the proportions of dry and wet lightning days at Darwin are essentially equal if the precipitation threshold is reset to 2.0 mm.

Median adjusted R^2 values for the fitted quadratic surfaces varied from variable to variable with 5.2 to 16% below 0.5 across the six CIGRE 500 sites and 47 to 68% above 0.75. The highest values were obtained for GPH500 and GPH700 (> 0.97), and the lowest for W and MING, (0.09 to 0.43). This pattern was consistent across all sites. Thus, overall, the quadratic surfaces described in the Appendix gave a reasonable representation of the main features of the atmospheric fields considered herein.

b. Classification analyses

Scatterplots of the skill scores obtained from the five classifiers are shown in Figure 3. The HR and FAR are for dry lightning and the radii of the circles represent the magnitude of the BS. For each CIGRE 500 site, the convex hull of the five data points obtained using only mu.DD850 and mu.TL850500 as candidate variables is displayed to facilitate their delineation. (The convex hull of a set of points is the smallest convex set enclosing the points.) The plots reveal six key features. First, for any site, approach (E-RF99 or BDC) and classifier, the HR exceeds the FAR (note the differences in the axis scales). Thus, both approaches and all five classifiers have some skill in discriminating dry lightning from wet lightning. Apart from Port Hedland, it is also evident that the RF99 approach (denoted by filled squares enclosed by green circles) provides lower HRs. Second, for Darwin and Townsville the approach of BDC often provides higher HRs than those for E-RF99 but at expense of higher FARs for some classifiers. Third, for Coffs Harbour, Melbourne and Perth, the approach of BDC produced simultaneously higher hit rates and lower FARs than those for E-RF99. Fourth, in the case of Port Hedland, the HRs and FARs obtained for a given classifier and the two approaches considered herein (E-RF99 and BDC) are similar despite the differences in candidate variable sets: the BDC candidate variable set included terms such as mu.TOTP, mu.CONVP and mu.TCW (see Appendix for details regarding their derivation). Fifth, except for Darwin, the application of LR to the BDC candidate variable set produced low FARs. Sixth, the approach of BDC produces similar or lower BSs than those for E-RF99.

< Insert Figure 3 about here >

c. Influential variables

The relative frequency histogram of influential variables across the six CIGRE 500 sites and four classifiers with easily interpreted decision rules or boundaries (CART, RF, LDA and LR) is shown in Figure 4. Overall, 16 out the 28 variables are means, nine are magnitudes of gradient vectors, two are vertical gradients and one is SEASON (Table 2). The seven most frequent variables are associated with atmospheric water content (mu.TOTP, mu.CONVP, gd.TOTP and mu.TCWV) and instability and lifting potential (mu.CBH, mu.DD700 and vg.T). Thus, five of the seven most frequent variables are means. In terms of the raw atmospheric variables listed in Table 2, not one of the variables used by RF99 (DD850 and TL850500) is present in this subset. Additionally, DD850 does not appear to be as influential as DD700, with DD700 and DD850 having relative frequencies of 0.0879 and 0.0220. As shown in Fig. 2c, high values of DD700 are typically associated with dry lightning rather than wet lightning, with a physical interpretation of this being that relatively dry air results in an increased likelihood of precipitation evaporating before reaching the ground (i.e., virga). The absence of CAPE and the near absence of W in Figure 4 suggest that these variables are not informative in terms of discriminating between dry and wet lightning conditions. This is unlikely to be the case for lightning activity studies involving discrimination between lightning and non-lightning days.

The dominance of the mean terms in the set of influential variables could be related to temporal variations in the timing of thunderstorms with respect to a given location, noting that although our analyses is based on afternoon values of the atmospheric variables (as this is when lightning most frequently occurs in these regions), lightning can also occur at other times of the day and night. The apparent influence of mu.TOTP and mu.CONVP must give rise to concern that information about precipitation has been used twice: once as daily precipitation readings at ground-based storage gauges were used to classify lightning days as either dry or wet; and twice as mu.TOTP and mu.CONVP values at 0600 UTC were derived from modeled precipitation and used as explanatory variables. However, scatter plots and quantile-quantile plots of mu.TOTP and mu.CONVP against the precipitation readings (not shown) revealed little evidence of relationships for all six CIGRE 500 sites. Except for Melbourne, robust estimates of the correlation coefficients ranged from 0.1 to 0.3. For Melbourne, the estimates were about 0.4. This lack of a simple relationship, and the positions of mu.TOTP and mu.CONVP in the histogram depicted in Figure 4, suggest that the construction of these variables captures useful additional information about atmospheric conditions that cannot be obtained from the other potential candidate variables considered. Some evidence for this conjecture is provided in Davies et al. (2013). For one of the tropical sites considered herein (Darwin), they created two concurrent long-term data sets that described the large-scale atmosphere and the characteristics of small-scale convection. They found that estimates of convective precipitation have a strong relationship with dynamical variables such as moisture convergence and vertical velocity at mid-levels. Wind rather than moisture convergence was used in the current study (Table 2), and vertical velocities in reanalyses can suffer from large inaccuracies (Abalos et al., 2015). The latter may have also contributed to the position of mu.W and gd.W in Figure 4.

< Insert Figure 4 about here >

Figure 5 displays the relative frequency histograms of the most-frequent atmospheric variables on a site-by-site basis. Here the maximum frequency for any variable is limited to four (the number of classifiers with interpretable decision rules or boundaries). Furthermore, the minimum count for any variable can be zero as not every one of the variables was found to be influential for every site. Colored bars indicate the seven most-frequent variables depicted in Figure 4. For the sake of clarity, white bars indicate additional variables that have a frequency of at least two. The variables depicted in Figure 5 are primarily associated with atmospheric water content and instability and lifting potential. Comparison of Figures 5a-d and 5e-f indicates a marked difference in the shapes of the histograms for sites located in western Australia (Perth and Port Hedland) and those in central and eastern Australia (Darwin, Townsville, Coffs Harbour and Melbourne). In the case of Perth (Figure 5e), five of the seven most-influential variables across all sites and classifiers depicted in Figure 4 have zero frequencies and the frequencies of the remaining two (mu.CBH and gd.TOTP) are low. It is the only site not to include both mu.TOTP and mu.CONVP amongst its set of influential variables. The most common variables across the four classifiers for Perth are indicated by white bars. Three of these four variables (mu.TGM7001000, gd.GPH500 and gd.GPH700) are not included in the variable sets for the other sites (cf. Figure 4). These variables are potential indicators of convective systems associated with fronts. The fourth variable (mu.TL850500) is selected for Coffs Harbour by LR only. For Port Hedland (Figure 5f), three of the seven most frequent variables in Figure 4 have zero frequencies. It is the only site to not include mu.CBM amongst its set of influential variables. The remaining four variables (mu.TOTP, mu.CONVP, gd.TOTP and m.TCWV) characterize atmospheric water content. There are four additional variables (gd.TOTP, gd.CONVP, mu.MING and mu.T2) that are not depicted in Figure 5f since they have a frequency of one. Port Hedland is also different to the other sites in that it has a notably higher HRs and lower FARs (Figure 3). This is because the ratio of dry to wet lightning proportions for Port Hedland is 4.3 which is much higher than that for the other sites where it is between 1.1 and 1.8 (Table 1). With the exceptions of Townsville and Coffs Harbour, the frequencies of vg.T are zero for the four remaining sites. Sharp temperature gradients are a potential indicator of troughs, and convergence along troughs can lead to showers and thunderstorms. The so-called inland (or easterly) trough is located on the inland side of the Great Dividing Range in Australia, forming a boundary between the moist air near the coast and dry air inland. It typically extends through central Queensland and into central New South Wales and is active during the months from September to May. Furthermore, the frequency of mu.ICE is greater than zero for Melbourne alone. Ice water content and lightning activity are highly correlated (Xu et al., 2010 and references therein), and this variable may provide information about the low (high) lightning flash rates associated with dry (wet) lightning. These results, and those illustrated in Figure 4, suggest that the optimal variable sets for lightning classification problems may vary between different climatic zones.

< Insert Figure 5 about here >

d. Cross validation experiments

Scatterplots of the mean skill scores obtained from the cross validation experiments are shown in Figure 6. Again, the HR and FAR are for dry lightning and the radii of the circles represent the magnitude of the BS. The radii of the circles have been placed on the same scale as those shown in Figure 3. The plots in Figure 6 reveal five key features. First, in all cases the mean HR exceeds the mean FAR. This indicates that both approaches (E-RF99 and BDC) and the classifiers considered herein have prediction skill when tested with independent data. While this is also true for the approach of RF99, the mean HRs are relatively low compared to those of either the E-RF99 or BDC approach. Second, the plots confirm the earlier finding that the approach of BDC generally provides either higher hit rates, or simultaneously higher hit rates and lower FARs, than that of E-RF99. Third, the mean FARs obtained using QDA are not always robust. This is particularly evident for Townsville, Melbourne, Perth and Port Hedland. This reflects the method's sensitivity to outliers. Fourth, when tested with independent data, applying LR to the BDC variable set often produced the lowest or competitive mean FARs. Fifth, the approach of BDC often produces competitive or lower BSs when tested with independent data than that of E-RF99.

< Insert Figure 6 about here >

A scatterplot of AUC values obtained from cross-validation of the LR models is shown in Figure 7. For all sites and both approaches (E-RF99 and BDC), the AUC values are greater than 0.5 indicating that prediction skill is better than climatology. However, the AUCs for the BDC approach are greater than those for E-RF99. The lowest AUC values were obtained for Darwin and the highest for Port Hedland (E-RF99 approach) and Perth (BDC approach).

< Insert Figure 7 about here >

4. Summary and conclusions

Daily lightning flash count and precipitation data from ground-based sensors and gauges, atmospheric information from the ERA-Interim reanalysis and five classification techniques (classifiers) were used to distinguish between 'dry' and 'wet' thunderstorm days for the period from 2004 to 2013 at six locations in Australia. The locations of the lightning flash (CIGRE 500) sensors represent a range of climatic settings (including temperate, subtropical and tropical regions). The earlier approach of Rorig and Ferguson (1999, RF99), which used two atmospheric variables and one classifier (linear discriminant analysis) for one region in the United States (the Pacific Northwest), was used as a benchmark to test whether the inclusion of additional atmospheric information and a wider range of classifiers resulted in a notable improvement in prediction accuracy for the climatic settings considered herein.

With future applications in mind, the study was designed to be conducted at the spatial resolution of current GCMs and reanalyses. Quadratic surfaces and determination of low-dimensional summary statistics (LDSS) were used to capture the main features of the atmospheric fields. Five classifiers were considered: classification and regression trees (CART); random forests (RF); linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and logistic regression (LR). Four prediction skill scores were considered, with a focus on dry lightning since it is the primary cause of wildfire ignition. Ten-fold cross validation was used to estimate the prediction accuracy of the classifiers. The study findings can be summarized as follows:

1) The use of LDSS captured useful and interpretable information in terms of the large-scale spatial structure of thunderstorms. While it can be argued that the LDSS are somewhat crude, our results suggest that there is value in their application to the problem of thunderstorm classification.
2) The approach outlined in this paper (BDC) and an extended version of that of Rorig and Ferguson (1999, herein designated as E-RF99) have prediction skill when tested against independent data for a wide range of climatic zones.
3) Overall, while five LDSS were used to better capture the main features of the atmospheric fields used, the mean field proved to be the most useful. The seven most-frequent

variables across the six sites and five classifiers considered are associated with atmospheric water content (mu.TOTP, mu.CONVP, gd.TOTP and mu.TCWV) and instability and lifting potential (mu.CBH, mu.DD700, and vg.T). The preceding lists of variables contain five spatial means, two gradient terms and variables derived from convective parameterizations. The results presented herein suggest that the latter may provide unique information that is not contained in ground-based precipitation data.

4) Despite the finding above, the set of influential atmospheric variables varied from site-to-site and between classifiers. This result needs to be tested using data from dense monitoring networks in different countries and a wide variety of climatic zones. The question of whether it is legitimate to use the same atmospheric variables and statistical classification techniques at different locations within the same climatic zone will be the subject of future research.

5) No single classifier proved to be consistently superior to its counterparts across the six sites considered. However, LR often produced lower FARs while the predictive accuracy of QDA was compromised by the presence of outliers in the variables.

6) Although the BDC variable selection approach requires more effort than that of E-RF99, with the exception of the Port Hedland site it produced either higher hit rates or simultaneously higher hit rates and lower false alarm ratios for dry lightning than that of E-RF99. It also tended to produce lower Brier (1950) scores and higher AUCs for LR models.

Although a number of previous studies have examined lightning and thunderstorm activity at the spatial and temporal scales of current reanalyses and GCMs, very few of these studies have considered ‘dry’ and ‘wet’ systems separately. The results presented here are intended to lead to an improved ability to classify deep convective systems in terms of their likelihood of being ‘dry’ or ‘wet’, as well as enhanced capability to understand the observed climatological characteristics of these systems. It is envisaged that the approach of this study will find application in future studies involving finer-scale reanalyses and GCM runs as they become available. Such work might lead to classification decision rules and boundaries that are less dependent on model parameterizations. Given the importance of dry thunderstorms for the ignition of wildfires by lightning, as well as wet thunderstorms in relation to a range of associated hazards (including extreme rainfall), a greater understanding of dry and wet thunderstorms could have significant benefits for improved resilience to the impacts of lightning and thunderstorms throughout the world.

Acknowledgments. The CIGRE 500 lightning flash counter registration data were provided by the Observations and Engineering Branch, Australian Bureau of Meteorology. Financial support for this work was provided by the Australian Climate Change Science Program, administered by the Department of Environment. We are indebted to Lorraine Bates for her assistance with data checking and coding. All data used in this work are available on request from Andrew J. Dowdy (andrew.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 manuscript.

APPENDIX

Representation of atmospheric variables

Most of the daily atmospheric variable information is available at 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 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: $z = f\left(x,\ y,\ p\right)$. 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

$$z = f(x, y) = c_1 + c_2 x + c_3 x^2 + c_4 y + c_5 xy + c_6 y^2 \qquad \text{(A.1)}$$

and the corresponding surface in three spatial dimensions by

$$\begin{aligned} z = f(x, y, p) &= c_1 + c_2 x + c_3 x^2 + c_4 y + c_5 xy + c_6 y^2 + c_7 p + \\ &\quad c_8 xp + c_9 yp + c_{10} p^2 \qquad \text{(A.2)} \end{aligned}$$

Instead of fitting (A.1) and (A.2) directly, the linear and quadratic terms were replaced by orthogonal polynomials in order 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); the estimates of the intercept and regression coefficients are placed on the same scale; and 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 centered on the CIGRE 500 sensor (i.e., a 7 × 7 grid described in Section 2a). The adjusted R^2 was used as a goodness-of-fit measure for the quadratic surfaces.

Let $\theta_1, \ldots, \theta_{10}$ denote the orthogonal polynomial regression coefficients. Five low-dimensional summary statistics (LDSS) for the above surfaces were used to facilitate physical interpretation: the intercept which is equivalent to the mean across the domain $(\mathrm{mu} = \theta_1)$; the magnitude of the gradient vector (gd) and its direction (dr) in the $x-y$ plane; Gaussian curvature (gc); and vertical gradient $(\mathrm{vg}=c_7)$. The magnitude of the gradient vector and its direction in terms of linear rate of change are defined by $\mathrm{gd}=\sqrt{\theta_2^2+\theta_4^2}$ and $\mathrm{dr}=\tan^{-1}\left(\theta_4/\theta_2\right)$. 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 which is independent of the coordinate system used to describe it. It is defined by

$$\mathrm{gc}=\det(\mathbf{H})=\lambda_1\lambda_2 \tag{A.3}$$

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

$$\mathbf{H}=\begin{bmatrix}\dfrac{\partial^2 z}{\partial x^2} & \dfrac{\partial^2 z}{\partial x\partial y}\\ \dfrac{\partial^2 z}{\partial y\partial x} & \dfrac{\partial^2 z}{\partial y^2}\end{bmatrix}=\begin{bmatrix}2\theta_3 & \theta_5\\ \theta_5 & 2\theta_6\end{bmatrix} \tag{A.4}$$

and λ_1 and λ_2 are the eigenvalues of **H** (also the maximum and minimum principal curvatures).

REFERENCES

Abalos, M., B. Legras, F. Ploeger, and W. J. Randel, 2015: Evaluating the advective Brewer-Dobson circulation in three reanalyses for the period 1979-2012. *J. Geophys. Res. Atmos.*, **120**, 7534-7554, doi:10.1002/2015JD023182.

Abatzoglou, J. T., C. A. Kolden, J. K. Balch, and B. A. Bradley, 2016: Controls on interannual variability in lightning-caused fire activity in the western US. *Environ. Res. Lett.*, **11**, 1-11, doi:10.1088/1748-9326/11/4/045005.

Alfons, A., 2012: cvTools: Cross-validation tools for regression models. R package version 0.3.2. http://CRAN.R-project.org/package=cvTools.

Allen, J. T., D. J. Karoly, and G. A. Mills, 2011: A severe thunderstorm climatology for Australia and associated thunderstorm environments. . *Aust. Meteorol. Oceanogr. J.*, **61**, 143-158.

Barthe, C., W. Deierling, and M. C. Barth, 2010: Estimation of total lightning from various storm parameters: A cloud-resolving study. *J. Geophys. Res*., **115**, D24202, doi:10.1029/2010JD014405.

Bates, B. C., R. E. Chandler, and A. J. Dowdy, 2015: Estimating trends and seasonality in Australian monthly lightning flash counts. *J. Geophys. Res. Atmos*., **120**, 3973–3983, doi:10.1002/2014JD023011.

Belsley, D., E. Kuh, and R. Welsch, 1980: *Regression Diagnostics Identifying Influential Data and Sources of Collinearity*. Wiley, 292 pp.

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

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

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78,** 1–3, doi: http://dx.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.

Chandler, R. E., 2005: On the use of generalized linear models for interpreting climate variability. *Environmetrics,* **16**, 699-715, doi:10.1002/env.731.

Christian, H. J., R. J. Blakeslee, D. J. Boccippio, W. L. Boeck, D. E. Buechler, K. T. Driscoll, S. J. Goodman, J. M. Hall, W. J. Koshak, D. M. Mach, and M. F. Stewart, 2003: Global frequency and distribution of lightning as observed from space by the Optical Transient Detector. *Journal of Geophysical Research: Atmospheres*, 108(D1).

Chronis T. G., S. J. Goodman, D. Cecil, D. Buechler, F. J. Robertson, J. Pittman, and R. J. Blakeslee, 2008: Global lightning activity from the ENSO perspective. *Geophys. Res. Lett*., **35**, L19804, doi: (2008).10.1029/2008GL034321.

Dai, J., Y. Wang, L. Chen, L. Tao, J. Gu, J. Wang, X. Xu, H. Lin, and Y. Gu, 2009: A comparison of lightning activity and convective indices over some monsoon-prone areas of China. *Atmos. Res.*, **91**, 438-452, doi:10.1016/j.atmosres.2008.08.002.

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, doi:10.1002/jgrd.50645.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc*., **137**, 553-597, doi:10.1002/qj.828.

Deierling, W., W. A. Petersen, J. Latham, S. Ellis, and H. J. Christian, 2008: The relationship between lightning activity and ice fluxes in thunderstorms. *J. Geophys. Res*., **113**, D15210, doi:10.1029/2007JD009700.

Dowdy, A. J., 2015: Large-scale modelling of environments favourable for dry lightning occurrence. *MODSIM2015, 21st International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand*, Gold Coast, Queensland, Australia, T. Weber, M. J. McPhee, and R. S. Anderssen, Eds, 1524-1530, ISBN: 978-0-9872143-5-5. [Available online at www.mssanz.org.au/modsim2015/G4/dowdy.pdf.]

Dowdy, A. J., 2016: Seasonal forecasting of lightning and thunderstorm activity in tropical and temperate regions of the world. *Sci. Rep.*, **6**, doi:10.1038/srep20874.

Dowdy, A. J., and Y. Kuleshov, 2014: Lightning climatology of Australia: temporal and spatial variability. *Aust. Meteorol. Oceanogr. J.*, **64**, 103-108.

Dowdy, A. J., and G. A. Mills, 2009: Atmospheric States Associated with the Igntion of Lightning-Attributed Fires. CAWCR Tech. Rep. No. 019, 34 pp.

Dowdy, A. J., and G. A. Mills, 2012a: Characteristics of lightning-attributed fires in south-east Australia. *Int. J. Wildland Fire*, **21**, 521–524, doi.org/10.1071/WF10145.

Dowdy, A. J., and G. A. Mills, 2012b: Atmospheric and fuel moisture characteristics associated with lightning-attributed fires. *J. Appl. Meteor. Climatol.*, **51**, 2025-2037, doi:10.1175/JAMC-D-11-0219.1.

Ensor, L. A., and Robeson, S. M., 2008: Statistical characteristics of daily precipitation: Comparisons of gridded and point datasets. *J. Appl. Meteor. Climatol.*, **47**, 2468-2476, doi:10.1175/2008JAMC1757.1.

Everitt, B .S., and G. Dunn, 2001: *Applied Multivariate Data Analysis.* Second edition. Arnold, 342 pp.

Everitt, B., and T. Hothorn, 2011: *An Introduction to Applied Multivariate Analysis with R.* Springer, 273 pp.

Faraway, J. J., 2016: *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* Second edition. Chapman and Hall / CRC Press, 399 pp.

Goodman S. J., D. E. Buechler, K. Knupp, D. Driscoll, and E. E. McCaul Jr., 2000: The 1997–98 El Niño event and related wintertime lightning variations in the southeastern United States. *Geophys. Res. Lett*. **27**, 541–544, doi:10.1029/1999GL010808.

Hendrickx, J., 2012: perturb: Tools for evaluating collinearity. R package version 2.05. http://CRAN.R-project.org/package=perturb.

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

Jayaratne, E. R., and Y. Kuleshov, 2006: Geographical and seasonal characteristics of the relationship between lightning ground flash density and rainfall within the continent of Australia. *Atmos. Res*., **79**, 1–14, doi:10.1016/j.atmosres.2005.03.004.

Kasischke, E. R., T. S. Rupp, and D. L. Verbyla, 2006: Fire trends in the Alaskan Boreal Forest. *Alaska's Changing Boreal Forest*, F. S. Chapin, M. Oswood, K. van Cleve, L. Viereck, and D. Verbyla, Eds., Oxford University Press, 285-301.

Kuleshov, Y., and E. R. Jayaratne, 2004: Estimates of lightning ground flash density in Australia and its relationship to thunder-days. *Aust. Meteorol. Mag.*, **53**, 189–196.

Kuleshov, Y., D. Mackerras, and M. Darveniza, 2009: Spatial distribution and frequency of thunderstorms and lightning in Australia. *Lightning: Principles, Instruments and Applications*, H. D. Betz, U. Schumann, and P. Laroche, Eds., Springer, 187–207, doi:10.1007/978-1-4020-9079-0_8.

Kuo, H. L., and W. H. Raymond, 1980: A quasi-one-dimensional cumulus cloud model and parametrization of cumulus heating and mixing effects. *Mon. Wea. Rev.*, **108**, 991–1009, doi: http://dx.doi.org/10.1175/1520-0493(1980)108<0991:AQODCC>2.0.CO;2.

Lang, T. J., S. A. Rutledge, B. Dolan, P. Krehbiel, W. Rison, and D. T. Lindsey, 2014: Lightning in wildfire smoke plumes observed in Colorado during summer 2012. *Mon. Wea. Rev.*, **142**, 489-507, doi:http://dx.doi.org/10.1175/MWR-D-13-00184.1.

Liaw, A., and M. Wiener, 2002: Classification and Regression by randomForest. *R News*, **2**, 18-22.

Luke, R. H., and A. G. McArthur, 1978: *Bushfires in Australia*. Australian Government Publishing Service, 368 pp.

Magi, B. I., 2015: Global lightning parameterization from CMIP5 climate model output. *J. Atmos. Oceanic Technol.*, **32**, 434-452, doi:10.1175/JTECH-D-13-00261.1.

McRae, R. H. D., 1992: Prediction of areas prone to lightning ignition. *Intl. J. Wildland Fire*, **2**, 123-130, doi:10.1071/WF9920123.

Morris, J. S., 2015: Functional regression. *Annu. Rev. Stat. Appl.*, **2**, 321-359, doi:10.1146/annurev-statistics-010814-020413.

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, doi:10.1016/j.atmosres.2015.12.018

Nauslar, N. J., M. L. Kaplan, J. Wallmann, and T. J. Brown, 2013: A forecast procedure for dry thunderstorms. *J. Oper. Meteor*., **1**, 200-214, doi:http://dx.doi.org/10.15191/nwajom.2013.0117.

NCAR - Research Applications Laboratory, 2015: verification: Weather Forecast Verification Utilities. R package version 1.42. http://CRAN.R-project.org/package=verification.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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

Ripley, B., 2014: tree: Classification and regression trees. R package version 1.0-35. http://CRAN.R-project.org/package=tree.

Romero, R., M. Gayà, and C. A. Doswell, 2007: European climatology of severe convective storm environmental parameters: A test for significant tornado events. *Atmos. Res*., **83**, 389-404, doi:10.1016/j.atmosres.2005.06.011.

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, doi:10.1126/science.1259100.

Rorig, M. L., and S. A. Ferguson, 1999: Characteristics of lightning and wildland fire ignition in the Pacific Northwest. *J. Appl. Meteor.*, **38**, 1565-1575, doi:http://dx.doi.org/10.1175/1520-0450(1999)038<1565:COLAWF>2.0.CO;2.

Rorig, M. L., S. J. McKay, S. A. Ferguson, and P. Werth, 2007: Model-generated predictions of dry thunderstorm potential. *J. Appl. Meteor. Climatol.*, **46**, 605-614, doi:http://dx.doi.org/10.1175/JAM2482.1.

Rothermel, R. C., 1972: A mathematical model for predicting fire spreads in wildland fuels. U.S. Forest Service Res. Paper INT-115, 40 pp.

Sanchez, G., 2013: DiscriMiner: Tools of the Trade for Discriminant Analysis. R package version 0.1-29. http://CRAN.R-project.org/package=DiscriMiner.

Takahashi, T., and K. Miyawaki, 2002: Reexamination of riming electrification in a wind tunnel. *J. Atmos. Sci.*, **59**, 1018-1025, doi: http://dx.doi.org/10.1175/1520-0469(2002)059<1018:ROREIA>2.0.CO;2.

Tiedtke, M., 1989: A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Mon. Wea. Rev.*, **117**, 1779–1800, doi:http://dx.doi.org/10.1175/1520-0493(1989)117<1779:ACMFSF>2.0.CO;2.

USDA Forest Service, 1992: 1984-1990 Wildfire Statistics. State and Private Forestry, Fire and Aviation Management Staff, USDA Forest Service (Washington, DC), 240 pp.

Vazquez A., and J. M. Moreno, 1998: Patterns of lightning- and people-caused fires in peninsula Spain. *Intl. J. Wildland Fire***, 8**, 103-115, doi:10.1071/WF9980103.

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

Wallmann, J., R. Milne, C. Smallcomb, and M. Mehle, 2010: Using the 21 June 2008 California lightning outbreak to improve dry lightning forecast procedures. *Wea. Forecasting*, **25**, 1447-1462, doi:http://dx.doi.org/10.1175/2010WAF2222393.1.

Weisman, M. L., and J. B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Mon. Wea. Rev*., **110**, 504-520, doi:http://dx.doi.org/10.1175/1520-0493(1982)110<0504:TDONSC>2.0.CO;2.

Williams, E., V. Mushtak, D. Rosenfeld, S. Goodman, and D. Boccippio, 2005: Thermodynamic conditions favorable to superlative thunderstorm updraft, mixed phase microphysics and lightning flash rate. *Atmos. Res*., **76**, 288-306, doi:10.1016/j.atmosres.2004.11.009.

Wotton B. M., and D. L. Martell, 2005: A lightning fire occurrence model for Ontario. *Can. J. For. Res*., **35**, 1389-1401, doi:10.1139/x05-071.

Wotton, B. M., B. J. Stocks, and D. L. Martell, 2005: An index for tracking sheltered forest floor moisture within the Canadian Forest Fire Weather Index System. *Intl. J. Wildland Fire*, **14**, 169-182, http://dx.doi.org/10.1071/WF04038.

Xu, W., E. J. Zipser, C. Liu, and H. Jiang, 2010: On the relationship between lightning frequency and thundercloud parameters of regional precipitation systems. *J. Geophys. Res., Atmos*., **115**, D12203, doi:10.1029/2009JD013385.

Yan, Z., S. Bate, R. E. Chandler, V. Isham, and H. S. Wheater, 2002: An analysis of daily maximum windspeed in northwestern Europe using generalized linear models. *J. Climate*, **15,** 2073-2088, doi:http://dx.doi.org/10.1175/1520-0442(2002)015<2073:AAODMW>2.0.CO;2.

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

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

Site No.	Location	Altitude (m)	Köppen classification	No. of lightning days	Proportion dry lightning days	Proportion wet lightning days
1	Darwin	30	Tropical savanna climate (Aw)	1350	0.53	0.47
2	Townsville	4	Tropical savanna climate (Aw)	286	0.53	0.47
3	Coffs Harbour	5	Humid subtropical climate (Cfa)	501	0.58	0.42
4	Melbourne	113	Marine west coast (Cfb)	570	0.64	0.36
5	Perth	15	Mediterranean (Csa)	148	0.55	0.45
6	Port Hedland	6	Subtropical desert (BWh)	401	0.81	0.19

TABLE 2. Abbreviations, full names, units of measure and specifications for atmospheric variables.

Abbreviation	Full name	Specification
	Instability and lifting potential	
CAPE	Convective available potential energy (J kg^{-1})	As provided in ERA-Interim reanalysis (maximum CAPE based on lifting parcels within a near-surface layer)
CBH	Cloud base height (m)	Based on 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 7x7 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 mid-level 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 difference (°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 reanalysis
ICE	Total column ice water (kg m^{-2})	As provided in ERA-Interim reanalysis
SH	Specific humidity (kg kg^{-1})	500, 700 and 850 hPa
TCWV	Total column water vapor (kg m^{-2})	As provided in ERA-Interim reanalysis
TOTP	Total precipitation (m)	As provided in ERA-Interim reanalysis
	Wind speed	
MVWS	Maximum vertical wind shear (m s^{-1})	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 meters, 500, 700 and 850 hPa
MSLP	Mean sea level pressure (Pa)	As provided in ERA-Interim reanalysis
GPH	Geopotential height (m^2 s^{-2})	500 and 700 hPa
MING	Minimum geostrophic vorticity (s^{-2})	Laplacian of geopotential at 500, 700 and 850 hPa

TABLE 3. Frequently used acronyms and abbreviations

Acronym or abbreviation	Full name
AUC	Area under receiver operating characteristic curve
BDC	Approach of Bates, Dowdy and Chandler (this paper)
BS	Brier (1950) score
CART	Classification and regression trees
E-RF99	Extended approach of Rorig and Ferguson (1999)
FAR	False alarm ratio for dry lightning
GCM	General circulation model
HR	Hit rate for dry lightning
LDA	Linear discriminant analysis
LDSS	Low-dimensional summary statistics
LR	Logistic regression
QDA	Quadratic discriminant analysis
RF	Random forests
RF99	Approach of Rorig and Ferguson (1999)
ROC	Receiver operating characteristic

LIST OF FIGURES

FIG. 7. Scatter plot of mean area under receiver operating characteristic curve (AUC) values obtained from cross-validation of logistic regression (LR) models. Key to numerals is given in Table 1.

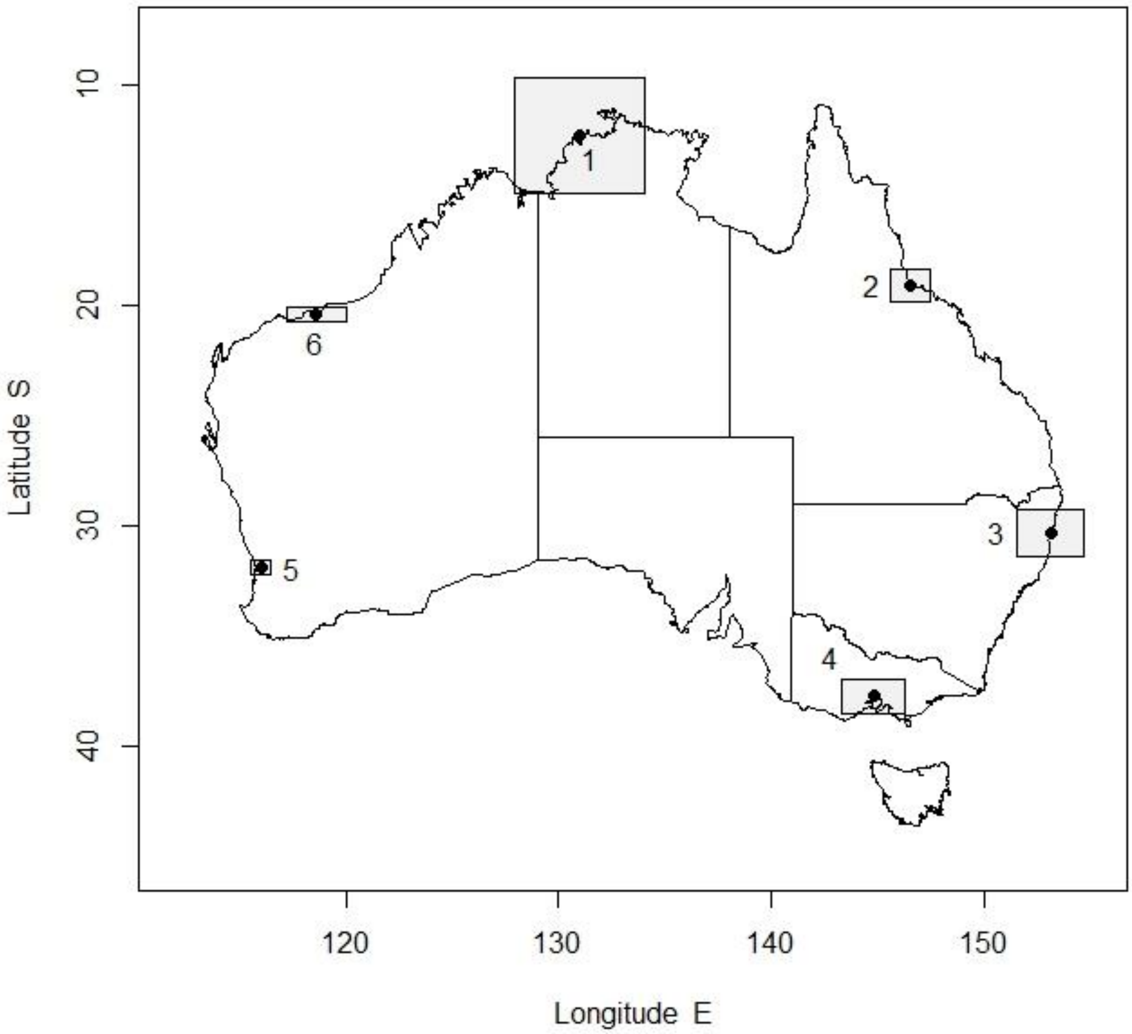

FIG. 1. Locations of CIGRE 500 sensors (filled circles) and relative proportions of dry lightning and wet lightning days in daily lightning flash count series. Key to numerals is given in Table 1. Widths of gray rectangles indicate proportions of dry lightning days and heights proportions of wet lightning days.

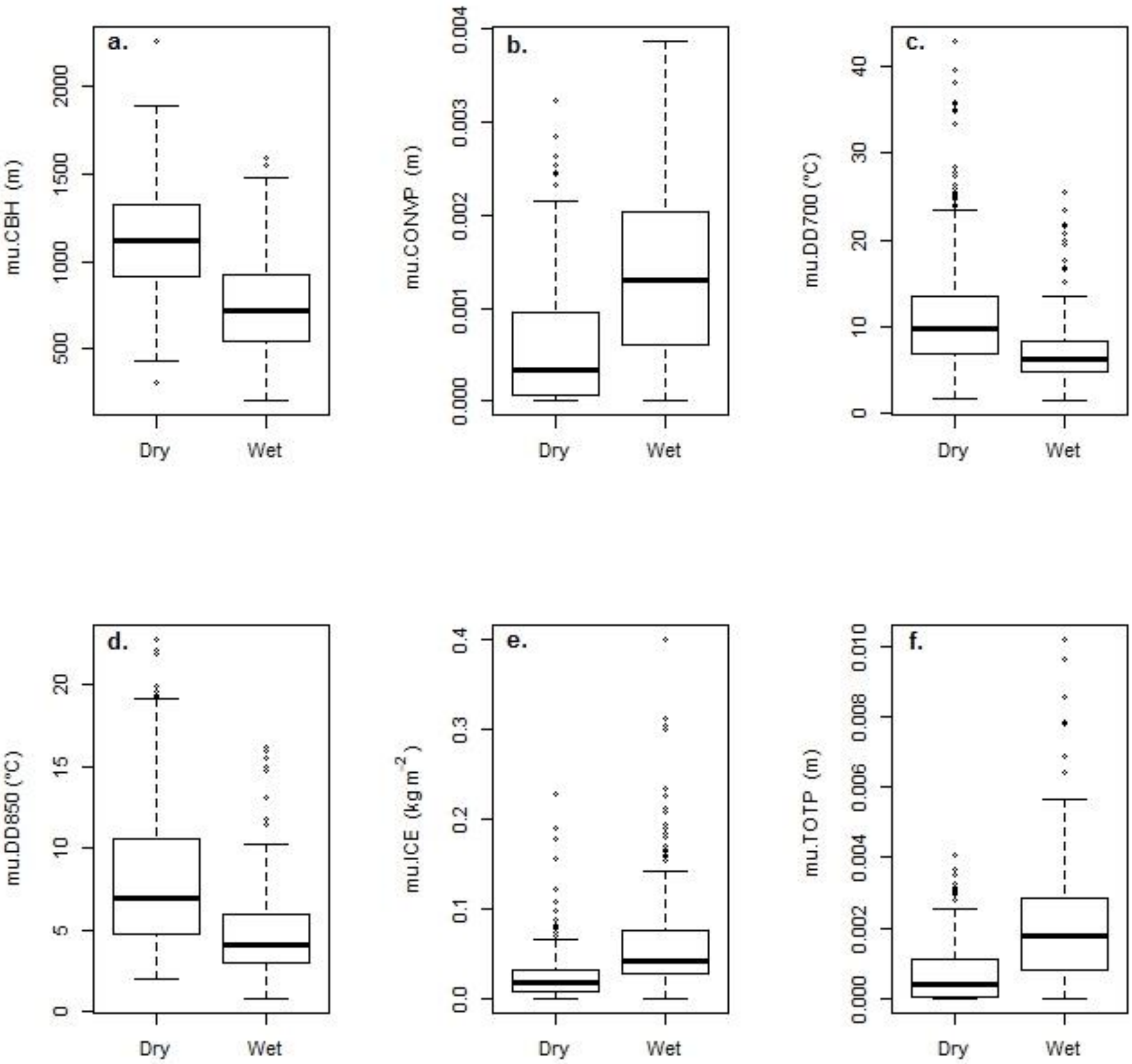

FIG. 2. Examples of comparative boxplots of potential candidate variables for the Coffs Harbour CIGRE 500 site: (a) mu.CBH, (b) mu.CONVP, (c) mu.DD700, (d) mu.DD850, (e) mu.ICE, and (f) mu.TOTP.

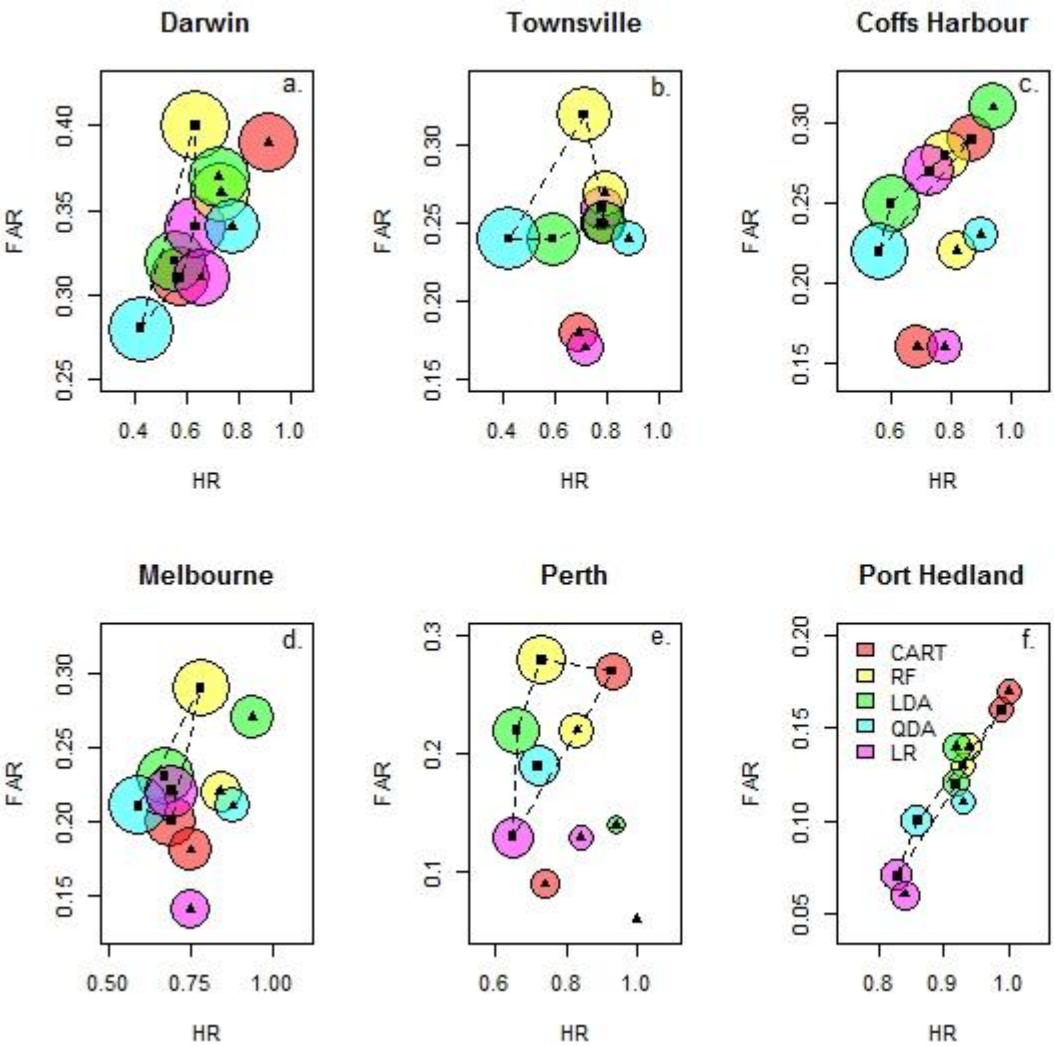

FIG. 3. Skill scores obtained using the methods of E-RF99 (filled squares) and BDC (filled triangles) for five classifiers and six CIGRE 500 sites. Radii of the circles are proportional to the Brier score. Dashed lines represent the convex hull of the false alarm ratio (FAR) and hit rate (HR) values for dry lightning obtained using the methods of E-RF99.

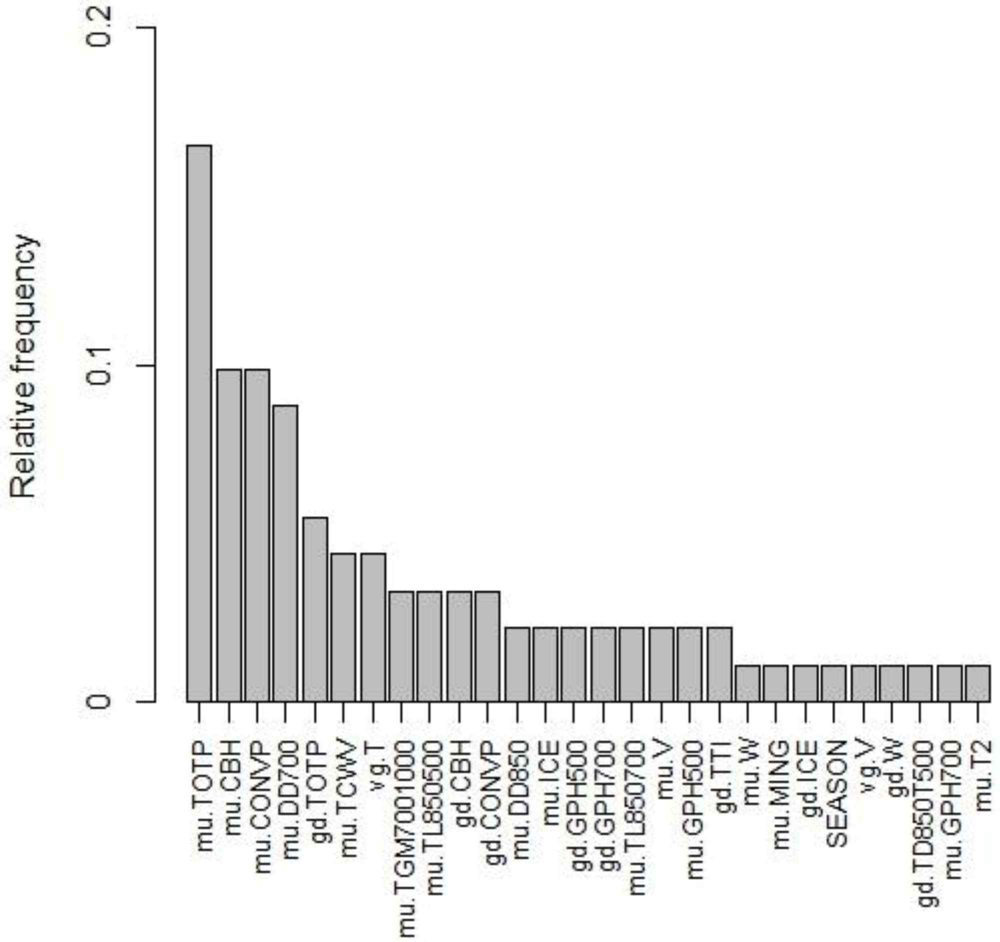

FIG. 4. Relative frequency histogram of selected variables across six CIGRE 500 sites and four classifiers: classification and regression trees (CART), random forests (RF), linear discriminant analysis (LDA) and logistic (LR).

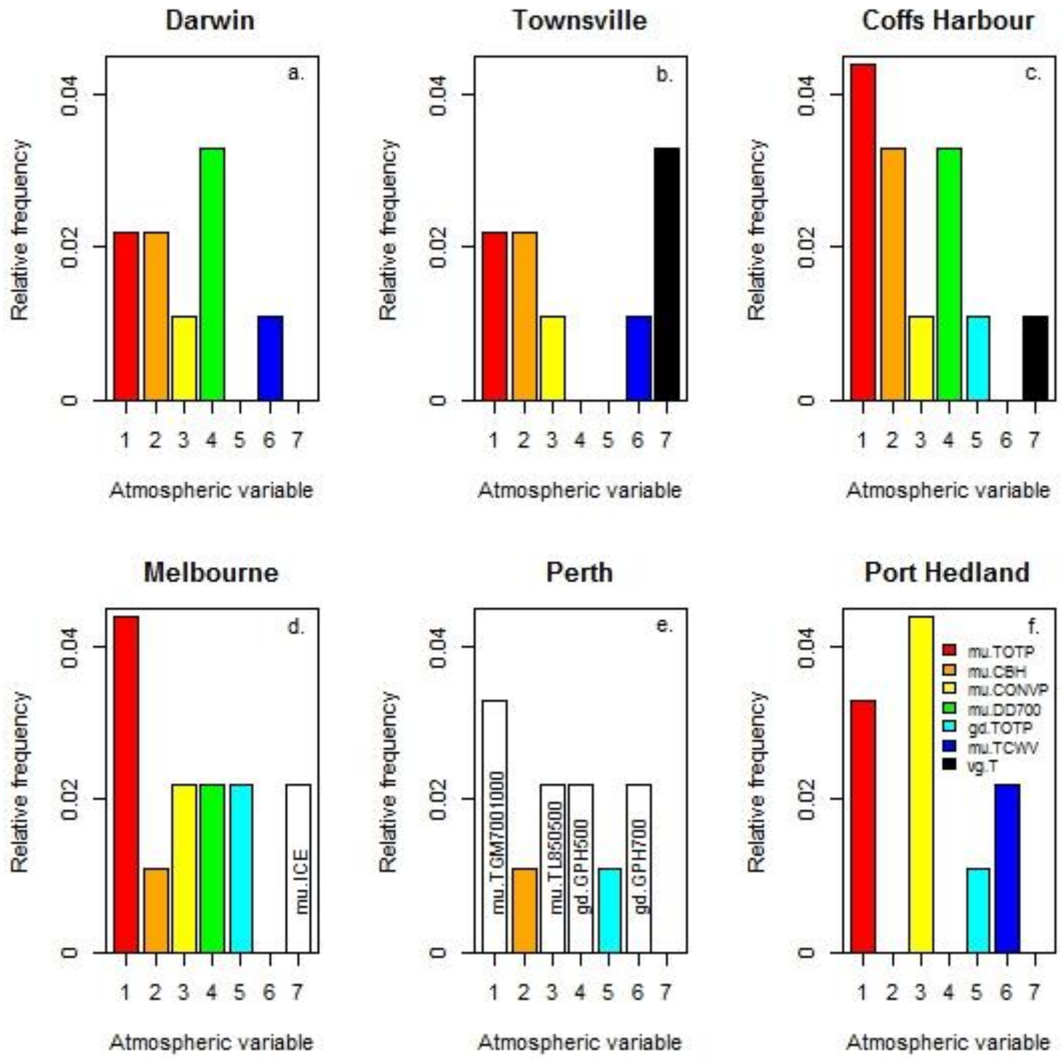

FIG. 5. Relative frequency histograms of influential variables for discriminating dry lightning from wet lightning days for each CIGRE 500 site across four classifiers: classification and regression trees (CART), random forests (RF), linear discriminant analysis (LDA) and logistic (LR).

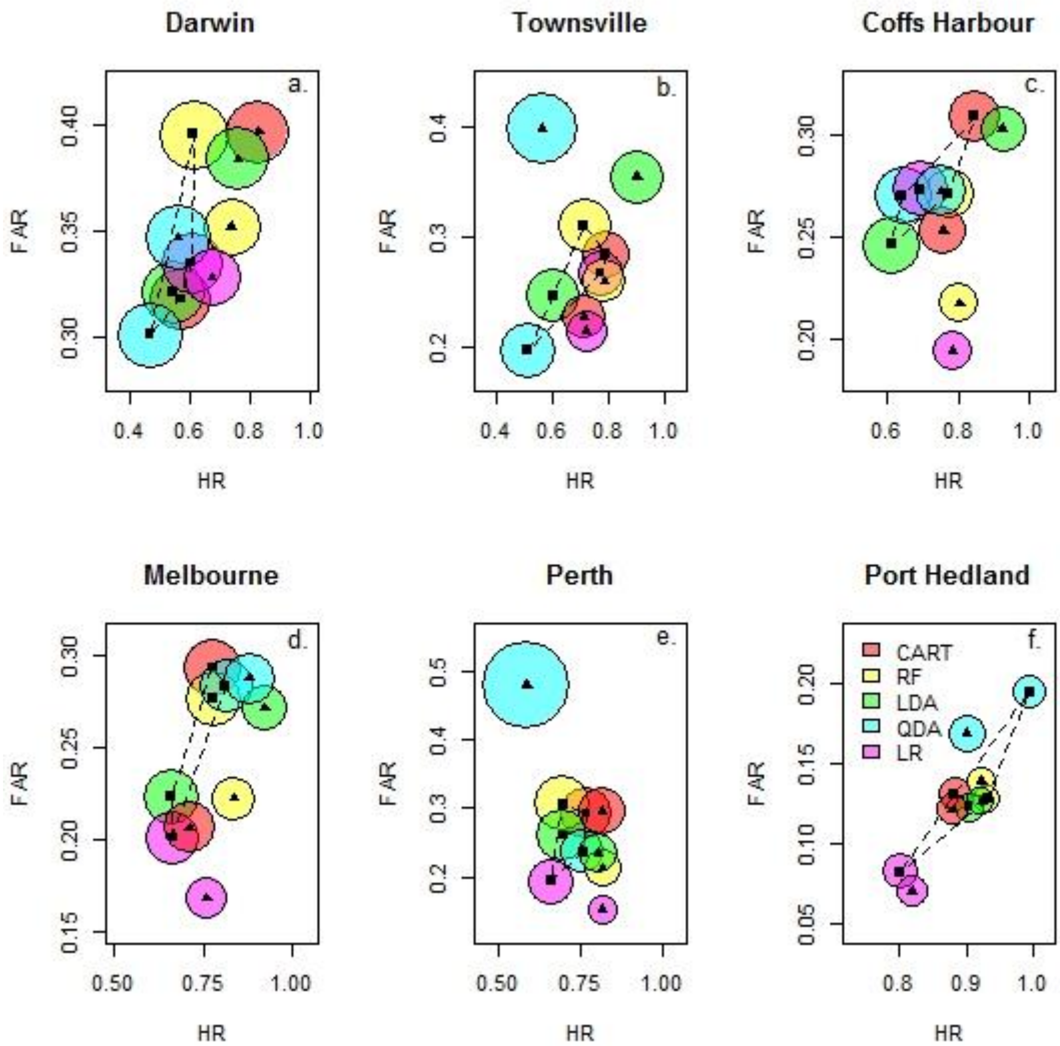

FIG. 6. Mean skill scores obtained from cross validation experiments using the methods of E-RF99 (filled squares) and BDC (filled triangles) for five classifiers and six CIGRE 500 sites. Radii of the circles are proportional to the Brier score. Dashed lines represent the convex hull of the mean false alarm ratio (FAR) and hit rate (HR) values for dry lightning obtained using the methods of E-RF99.

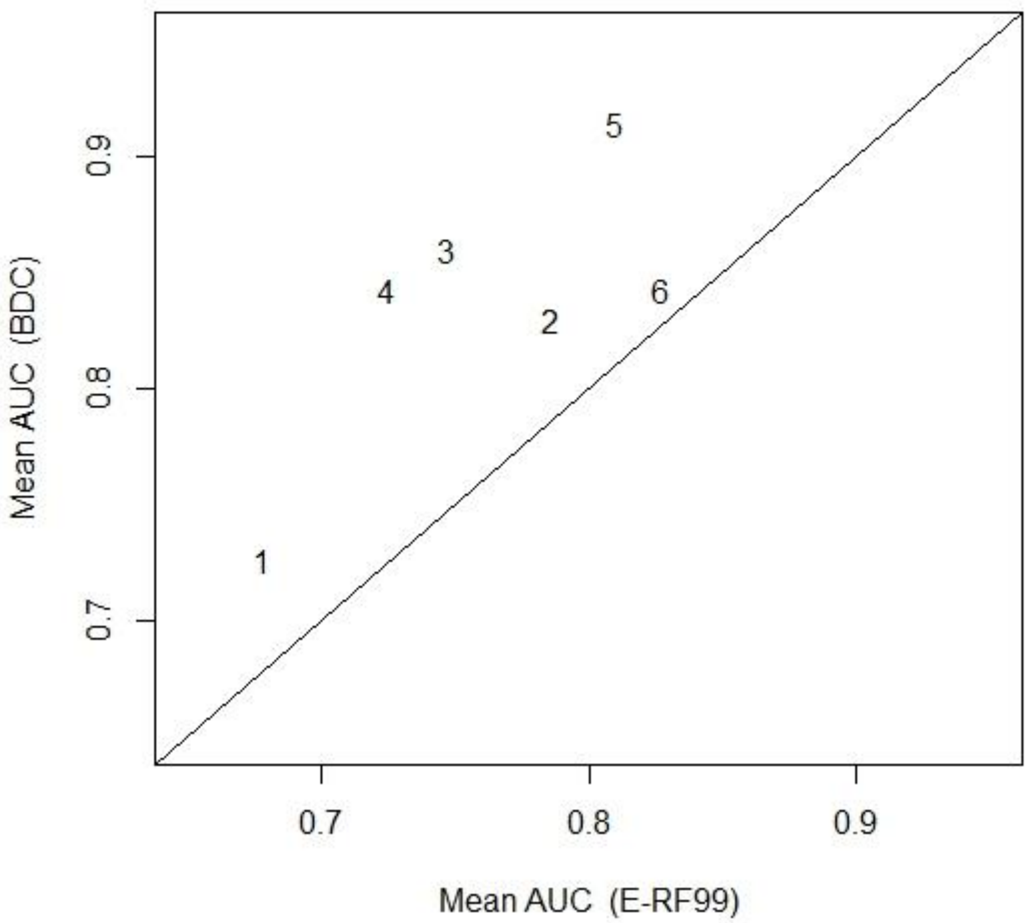

FIG. 7. Scatter plot of mean area under receiver operating characteristic curve (AUC) values obtained from cross-validation of logistic regression (LR) models. Key to numerals is given in Table 1.