Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

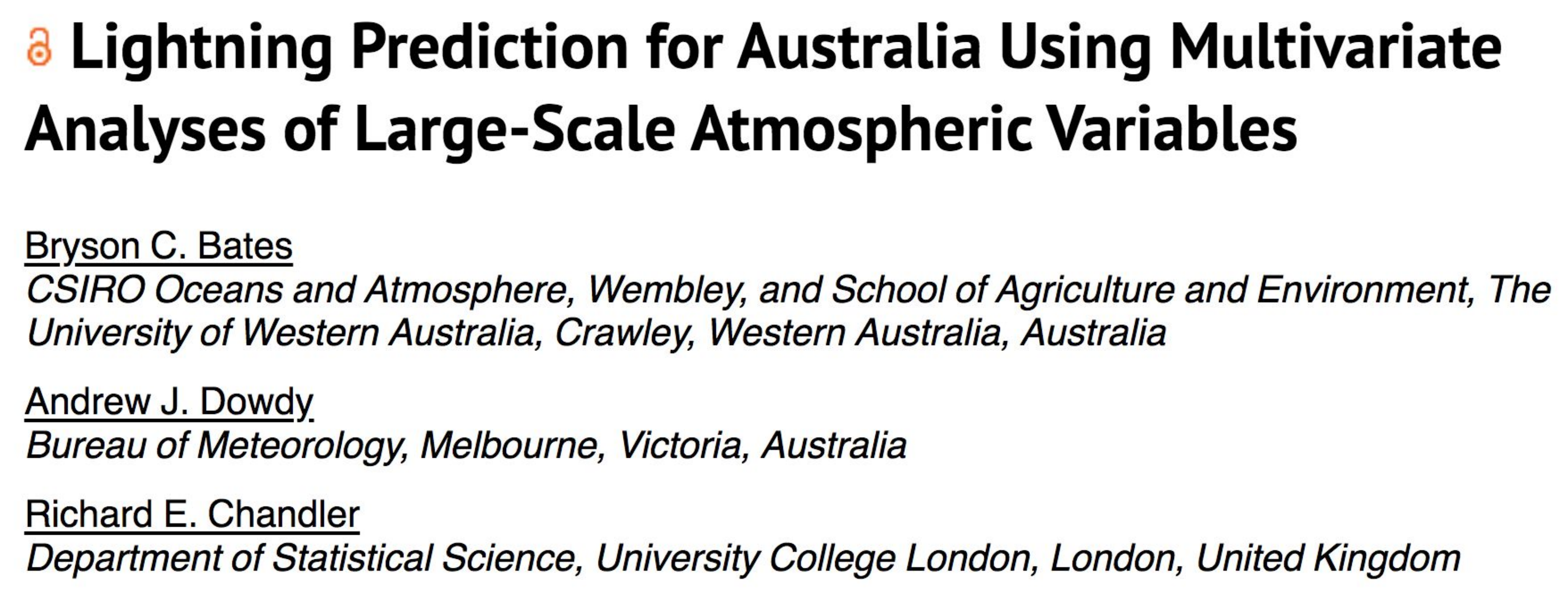
EXCELENCIA
SEVERO
OCHOA

AI4ES Journal Club #3

Núria Pérez-Zanón
Computational Earth Science group

Barcelona
21/04/2020

Article of the day



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

Link: https://journals.ametsoc.org/doi/full/10.1175/JAMC-D-17-0214.1?mobileUi=0

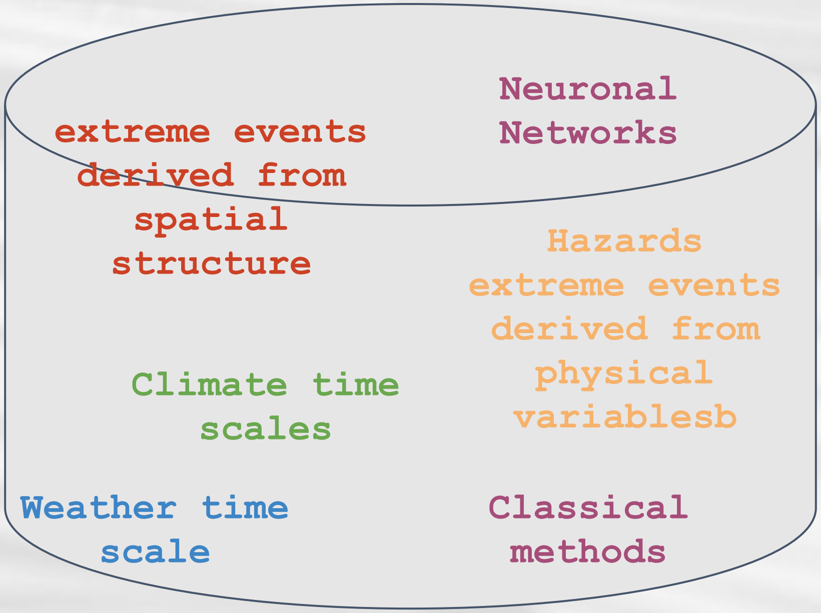
Context

Why this article?

I would like to study natural hazards with negative impacts on society whose origin is an atmospheric event, e.g. lightning storms

These events are different from cyclons and storm tracks since they don’t create a spatial pattern in an atmospheric field if not they are recorded in observations (they are not a physical variables output of climate models).

I would like to apply modern techniques, such as Neuronal Networks, to forsee the occurrence of these events in climate time scales (from sub-seasonal to decadal) but also I need to understand how to tackle the problem to construct the correct question (one I can find an answer or a sort of answer).



Neuronal Networks

extreme events derived from spatial structure

Hazards extreme events derived from physical variablesb

Climate time scales

Weather time scale

Classical methods

Context

Aim of the article:

- The principal aim of this study is to investigate the relationships between lightning activity and atmospheric conditions.

How?

- Applying classification methods to **distinguish between nonlightning and lightning days**

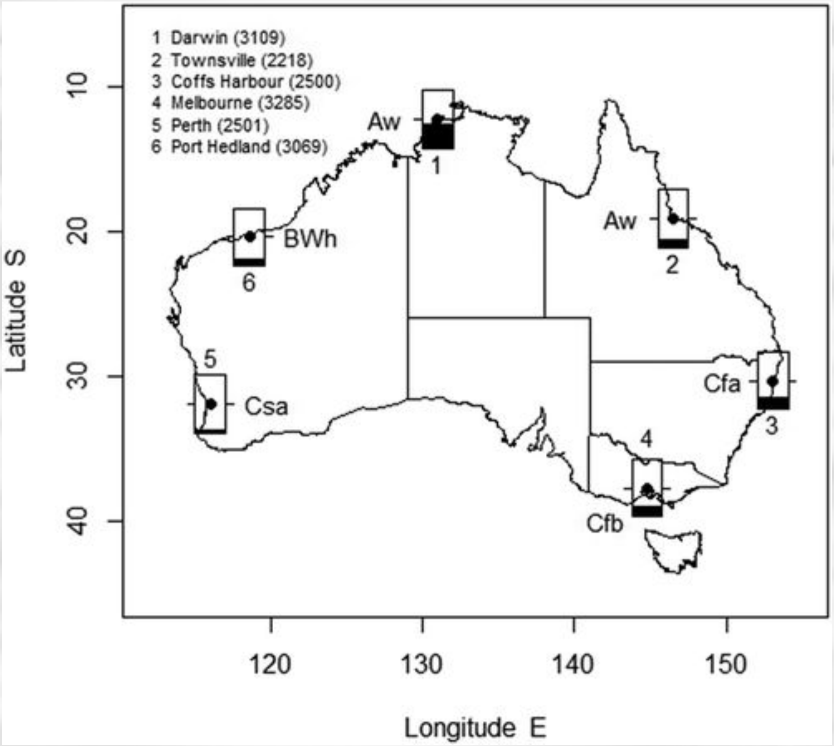
Why?

- Because there is a lack on the researches conducted: 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.

Data

Ligjtning-flash counts at six locations

- the locations belongs to different **climatic zones**
- the **total number of flashes** are used even being the sensor able to distinguish between cloud-to-ground and cloud-to-cloud flashes
- daily manual observation at 0800-0900 local time the



Definition of the event

- A lightning day is a day in which 1 flash has been recorded.
- A lightning day is a day in which 2 flashes have been recorded (not shown - same results).

Data

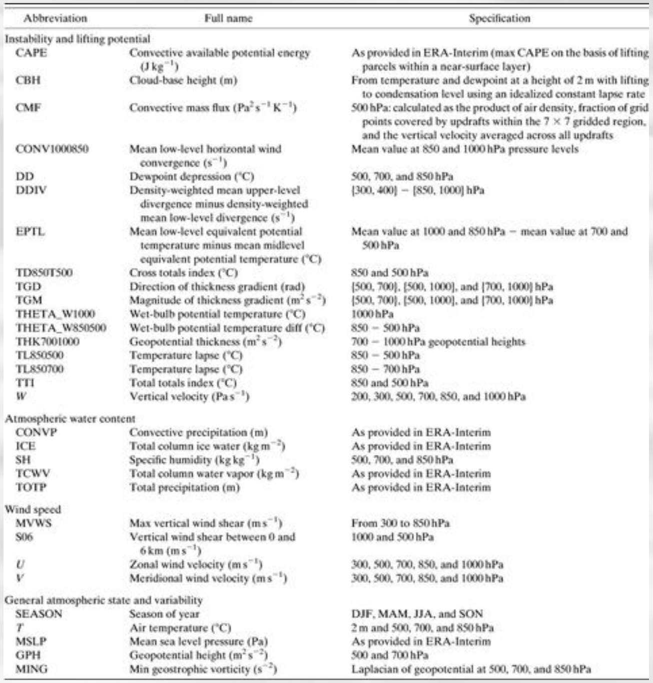
ERA-Interim variables

- 31 atmospheric fields associated with deep convection (dynamical and thermodynamical processes)

- 49 reanalysis grid points closest to the sensor’s

- synchronization with lightning observations the 0600 UTC field is chosen:
  - higher lightning activity in the afternoon
  - 6hourly ERA-Interim resolution
  - other time steps are redundant (high correlations)

- Quadratic suraces and low-dimensional summary statistics (LDS):
  - 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 R2 (r2).



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 × 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

Matrix

- Selection of the LDS that shows greater contrast
- Standardization of these statistics
- collinearity detection by variance decomposition proportions → remove the statistics with colliniarity

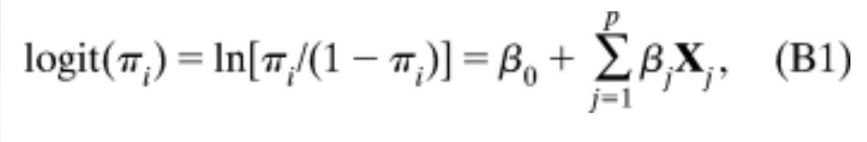
Methods

Classification methods:

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

PCA is used to reduce the dimensions of the matrix

LR model depends on the probability of occurrence of the event and beta parameters are regression coefficients



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

Random Forest is an ensemble learning algorithm creating bootstrap samples of the original data.

LDA is a method used to to find a linear combination of features that characterizes or separates two or more classes of objects or events.

Quadratic discriminant analysis is a generalization of LDA in which two classes need not have the same covariance matrix.

Methods

Classification methods:

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

Measures of prediction skills:

- hit rate (HR),	1	0
- false-alarm ratio (FAR),	0	1
- Brier (1950) score (BS), and	0	1
- (for LR) the area under the receiver-operating-characteristic curve (AUC)	0	0.5

Tenfold cross validation was used to assess how well the classifiers performed on an independent dataset.

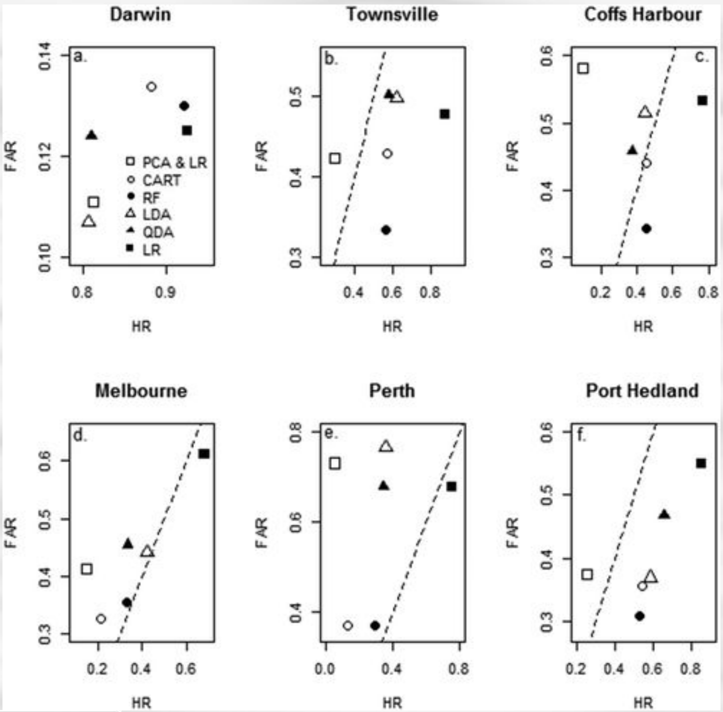
Results

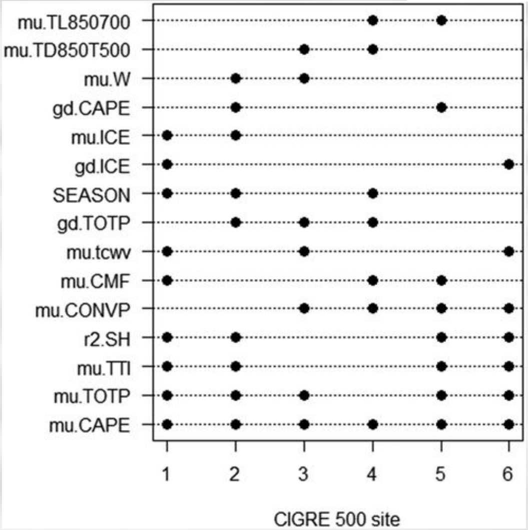
Figure 2 shows that **Logistic Regression** (LR) is the best method in all cases.

Figure 4 shows the significant variables in the LR. The once appearing in most of the locations

- mu.CAPE (convective energy)
- mu.TOTP (precipitation)
- mu.TTI (inestability)
- r2.SH (specific humidity)

Fig. 2. Cross-validated prediction skill.





Results

There are a few very interesting remarks about the variables included in the LR:

Measure of storminess in Australia:

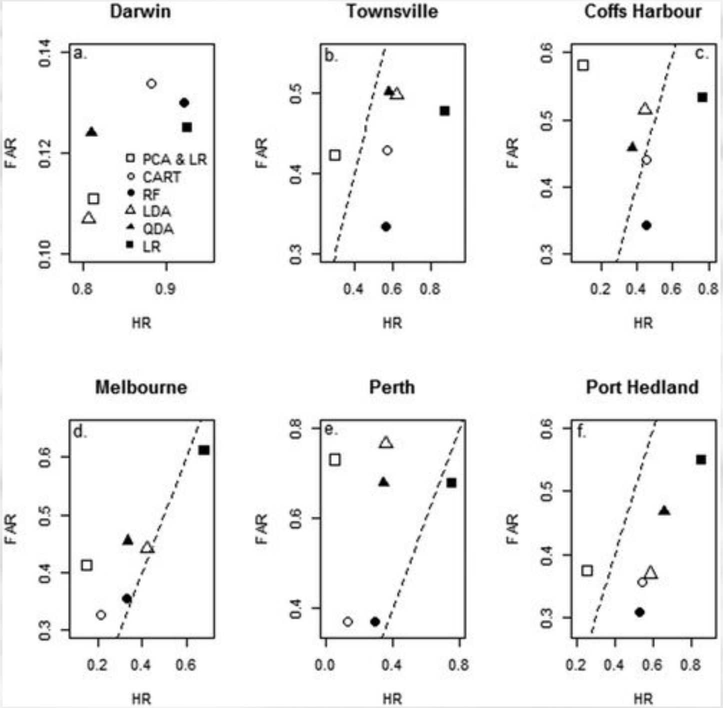
mu.CAPE (convective energy)
mu.TOTP (precipitation)

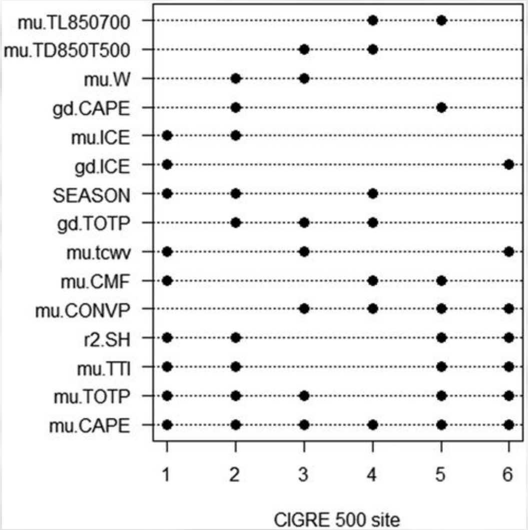
Parametrization of variables:

mu.TTI (inestability)

r2.SH (specific humidity)

- Wind variables were not part of the final LR model probably because of the sort of event.





Conclusions

1. Low-dimensional summary statistics (LDS) capture useful information about the structure of thunderstorms at coarse spatial and temporal scales.

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. Predominan variables are spatial mean measures of instability and lifting potential and of atmospheric water content (10 of 15 variables).

5. The variables in the final LR models varied across climatic zones.

Critique and questions

- **It has dangerous impact when 1 flash is recoreded in a day** … which is the probability of 1 flash producing damage? What about to consider lightning storm with much higher number of flashes?

- **only atmospheric fields. Could be useful ocean fields such as sea surface temperature?** I guess it will be useful depending on the time-scale

- why 49 grid points?
- Could this study be perform keeping the spatial fields?

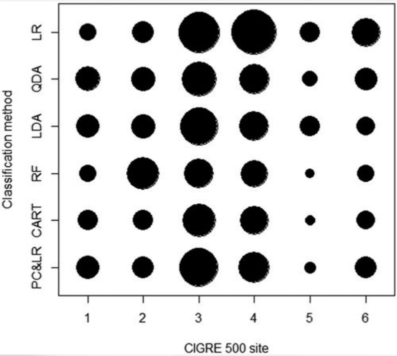
- How good is Tenfold compare to having training and validation sets? Why there is a single dot in Figure 2?

- Personally, I think figure 3 has a few information.

- **Is it possible to use Neuronal Networks** into this article to distinguish between lightning and non-lightning days?

- **Are all the variables** obtained from ERAInterim **available in a forecast model ou** not, a forecast model based on the LR could not be operational.

- Should parameterized variables in atmospheric models be avoided?



Comments and questions

Thanks

Minutes

- Options to try other event definitions (Núria)
- Options to try other time step on ERA-Interim (daily mean) or a sequence (Suso)
- Use a different number of gridpoint or a single field for Australia region (several)
- Tenfold cross validation is better to a single cross validation set. (Carlos)
- Tenfold is different than having training and validation set for training a Neuronal Network. (Carlos)
- Figure 2 shows the mean of all Tenfold samples but the spread could be also interesting information to be shown. (Carlos)
- Neuronal Network approach would be useful to distinguish between lightning and non-lightning days and avoid to reduce the sample size (Carlos)
- For operational objectives, the variables should be common in forecast model output and observational gridded datasets. (Lluis and Núria)
- Variables that are parameterized in the models simulations must be avoided in the ML model. (Núria)
- Figure 2 shows strange values for most of the models, that may indicate something wrong in the data or the models. (Hervé)
- Discussion about collinearity should be more clearly considered in the early stages (Hervé, Amalia)