24th International Congress on Modelling and Simulation, Sydney, NSW, Australia, 5 to 10 December 2021
mssanz.org.au/modsim2021

Enhancement of lightning strike forecasts using a machine learning approach

Gow, Daniel[a], Julien Lerat[b], Rob Warren[c] and Ivor Blockley[c]

[a] Australian Bureau of Statistics, Canberra
[b] CSIRO, Land & Water, Canberra
[c] Bureau of Meteorology, Melbourne
Email: julien.lerat@csiro.au

Abstract: This study represents a preliminary attempt to explore how a machine learning approach could be used to generate operational forecasting products in the Bureau of Meteorology. The objective of this paper is to compare feed-forward neural networks with the current operational system to generate forecasts on the likelihood of lightning strikes. The results presented here suggest that a network model based on off-the-shelf machine learning tools can rival or even surpass the performance of the current system when evaluated during a 90-day period and using a single performance metric (Brier skill score). The performance improvement was particularly noticeable when predicting lightning over the sea, where the current system consistently over-forecasts lightning. These results, although remaining a pilot study with no intention of operational deployment, are encouraging and show the potential for machine learning to support lightning strike forecast. The fact that off-the-shelf machine learning tools were used in this work indicate that the approach is cost effective with limited customised development needed, and potentially rapid deployment to operations.

Views expressed in this paper are those of the first author while on secondment to the Bureau of Meteorology and do not necessarily represent those of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the authors.

Keywords: *Lightning strike forecast, deep learning, machine learning*

1. INTRODUCTION

Thunderstorms represent a significant hazard in Australia. By definition, all thunderstorms produce lightning, which can trigger bushfires, damage electrical infrastructure, disrupt airport operations, and cause injury or death (Rabbani et al., 2014). Accurately predicting the potential for thunderstorms is therefore one of the most critical activities undertaken by forecasters at the Bureau of Meteorology (BoM). Since 2016, thunderstorm forecasting at the BoM has relied on forecasts from the Calibrated Thunder (CT) system (Warren et al., 2021). CT combines numerical weather prediction (NWP) model forecast data with recent lightning observations to produce calibrated probabilistic forecasts of lightning out to a maximum of 8 days. Specifically, the system predicts the probability of lightning within 10 km of a point during successive 3- and 24-hour periods over a domain covering Australia and a 500 km coastal margin. In the latest version of CT (operational since March 2021), NWP data are provided by the global ensemble configuration of the Australian Community Climate and Earth System Simulator (ACCESS-GE) and total lightning observations come from the Weatherzone Total Lightning Network (WZTLN).

This paper presents a preliminary investigation on the potential of a machine learning model to enhance the performance of the current CT system. Machine learning is now becoming an important tool to complement traditional physics-based and statistical models used in weather forecasting in general (Düben et al., 2021; LeCun et al., 2015; Xie et al., 2021). Recently, a particular type of machine learning introduced by Chen and Guestrin (2016) named "gradient boosting" has been explored to improve short-range lightning predictions by Mostajabi et al. (2019), with forecasts up to 30 minutes at 10 minute intervals. They used four predictors (surface pressure, temperature, relative humidity and wind speed) sourced from the ERA5 reanalysis product (Hersbach et al., 2020) over a period of 12 years. The machine learning model was shown to have skill up to 30 minutes of lead time compared to a persistence benchmark. This work is promising but remains limited to a very short lead time, which is insufficient for operational lightning products developed by the Bureau. In addition, the predictors were derived from a reanalysis product, which cannot be used for real-time predictions beyond a short nowcasting horizon. Similar limitations apply to the work of Zhou et al. (2020) who generated lightning strike nowcasts using a deep neural network and predictors generated from Himawari-8 satellite products, Doppler radar and lightning observation data. The network model they used is a variant of the SegNet encoder-decoder (Simonyan & Zisserman, 2014), which allowed the authors to successfully generate lightning nowcasting products. This work is an example of transfer learning where an existing network architecture is adapted for a specific use case. This is an attractive proposition for an organisation like BoM where it is important to balance the cost of system development with maintenance cost, which is reduced significantly with a transfer learning approach. Consequently, the aims of this paper are threefold:

- Develop a lightning strike forecast model using off-the-shelf machine learning tools within the well-established Tensorflow and Keras frameworks to allow potential access to transfer learning in future developments.
- Use predictor data derived from an operational NWP model and extend the lead times up to several hours ahead to ensure that the forecasts generated by the deep learning model remain relevant in an operational context for the BoM.
- Compare model performance against the existing CT system to assess the potential for machine learning to support future BoM operational products. As this is a pilot study, the verification metrics used to compare performance will be limited to the Brier Score (see section 2). However, this approach cannot replace a detailed analysis of performance that will require more metrics and evaluation over longer periods.

The paper describes the data, network architecture and verification metrics in section 2. Results are presented and discussed in section 3 followed by conclusions in section 4.

2. METHOD

2.1. Predictors and predictands data

The development and testing of network models presented in this paper relies on historical re-forecast data, referred to as a "hindcast", generated for the verification of the current CT system. Hindcasts were performed for the 90-day period from 1 December 2020 to 28 February 2021 (i.e., Southern Hemisphere summer). Climatologically, this represents a period of increased lightning activity across Australia, particularly in the tropical north of the country and along the east coast. Consistent with the operational CT system, 3-hour predictions of lightning probability were produced four times a day (starting at 00, 06, 12, and 18 UTC) with a maximum lead time of 216 h.

For this proof-of-concept study we focus on a single lead time covering the period 3–6 h from the time the forecast was issued (referred to as the "base time"). This allows us to consider diurnal variations in lightning probability without the need to address the systematic decrease in forecast skill with lead time.

The two main predictors used in our network models are the same as those used by CT, both computed from ACCESS-GE. The first is the probability between 0 and 1 of CPTP exceeding a threshold of 1, where CPTP is the cloud physics thunder parameter (Bright et al., 2005). CPTP measures instability within the layer of the atmosphere in which charge separation occurs and is scaled such that values above 1 indicate favourable conditions for lightning. The probability of CPTP > 1 was computed by counting the number of ACCESS-GE ensemble members exceeding the threshold of 1 and dividing by the total number of ensemble members. The second predictor is the probability of precipitation > 0.25 mm during each 3-hour period. This provides an indication of where rain-bearing cloud systems are likely to occur. Combined, the two predictors indicate where (precipitating) thunderstorms are possible. Both predictors are provided on a regular latitude–longitude grid spanning 107.175–160.125°E and 5.15–47.95°S with grid spacing of 0.15° in longitude and 0.1° in latitude. This gives a total of 354 x 429 grid points; however, points are masked beyond a 500 km margin around the Australian coastline, leaving a total of 95,180 valid data points for each time.

In addition to the probabilities of CPTP > 1 and precipitation > 0.25 mm, six other predictors were considered: the grid cell longitude and latitude (to account for location of the grid cell within the modelling domain), the sine and cosine of the hour of the day of the forecast base time (to account for the time of the day the forecast is issued), a binary flag indicating whether the grid cell is over land (1) or sea (0), the cosine of the solar zenith angle, and the climatological lightning frequency. The climatological lightning frequency was estimated for every day of the year and 3-hour period (00–03 UTC, 03–06 UTC, etc.) based on historical WZTLN lightning observations for the 6-year period from 1 December 2014 to 30 November 2020.

The predictand data are binary and indicate if any lightning strikes occurred within a radius of 10 km around every grid point of the modelling domain for each 3-hour interval during the 90 days hindcast period. These are the observations used to train, validate, and verify the network models.

2.2. Network architecture

A range of feed-forward networks with increasing complexity were built using the Keras and Tensorflow frameworks. All network configurations tested share the following characteristics:

- As indicated in the previous section, predictors (and hence lightning prediction) were limited to variables issued with a +3 to +6 hours lead time.
- All networks are “feed-forward”. In other words, the data flows from the input layers towards the output layers only, with no recurrence or feed-back loop between output and internal layers.
- All layers used to build the networks rely on the Rectified Linear Unit activation function (Ramachandran et al., 2017), except the last one which uses a logistic function to provide a prediction of lightning probability in the [0, 1] interval.
- Input data are processed sequentially for a single cell and a single forecast base time. Note that this is distinct from convolutional networks implemented by Zhou et al. (2020) which process all grid cells from the modelling domain for a single time step simultaneously. The setup adopted in this paper has the advantage of simplifying the formatting of input data (which become a large tabular dataset) and reducing the number of network parameters, because a single set of weights is applied to all cells and forecast base times during the testing period.
- The input layers are always normalized to eliminate the influence of the scale of predictor variables on the network training algorithm.

Three items were explored to compare network architectures:

- The number of hidden layers was varied from zero, where the network becomes identical to a logistic regression model (i.e., linear mapping between inputs and output layer), to four.
- The number of nodes per hidden layer ranged from 4 to 256 using increments in power of 2.
- The use or not of a drop-out layer (Srivastava et al., 2014), which reduces over-fitting of deep-learning models by randomly setting some of the weights to 0.

Note that additional configurations were tested by reducing the set of predictors but did not lead to increased performance, so are not reported here. For the sake of brevity, only a subset of all architectures tested are reported here, as detailed in Table 1.

Table 1. Selected network architectures

Configuration	Architecture
Model 000	Current CT system (benchmark)
Model 024	Two layers of 32 nodes each.
Model 025	Three layers with 128, 64 and 32 nodes.
Model 028	Four layers of 32, 64, 128, 256 nodes
Model 029	Four layers of 32, 64, 128, 256 nodes with a final dropout layer of 40%.

2.3. Network training and testing

All networks were trained using the Adam (Kingma & Ba, 2014) algorithm and a binary cross entropy loss function defined as

$$L = \sum_i [o_i \log(f_i) + (1 - o_i) \log(1 - f_i)] \quad \text{Eq. 1}$$

where i is the sample number, $o_i \in \{0, 1\}$ is the binary observed lightning occurrence, and $f_i \in [0, 1]$ is the predicted lightning probability. Note that no class weight or re-sampling was implemented to account for the severe imbalance between the frequency of events and non-events. Training was performed for 100 epochs, with early stopping based on validation to prevent overfitting (see below).

All networks were trained and verified using a rolling window of 29 days. Using the classical ML terminology we distinguished a period for training (i.e. parameter calibration), validation (i.e. monitoring of the loss function over independent data to prevent overfitting) and testing (assessment of performance outside of training and validation period). The three periods were defined as follows: we used 21 days for training, followed by 7 days for validation and one day of testing. This was applied for each day from 29 December 2020 to 28 February 2021. Finally, all test forecasts were stitched together to form a continuous test dataset of 62 days in length. Compared to the classical split between training, validation and test periods (for e.g. in Zhou et al., 2020), the rolling window implemented here offers two main advantages. First, it reduces the amount of data to be processed during network training, thereby reducing runtime. Second it allows the network to account for seasonality, with parameters gradually varying during the season.

2.4. Forecast verification

The skill of our lightning forecasts is assessed using the Brier skill score (BSS). The Brier score is a measure of forecast accuracy, computed as

$$BS = \frac{1}{N} \sum_i (f_i - o_i)^2 \quad \text{Eq. 2}$$

where N is the number of forecast–observation pairs. BS varies from zero to one, with $\mathrm{BS} = 0$ for a perfectly accurate forecast. BSS measures the skill of a forecast with respect to some reference forecast and is computed as

$$BSS = 1 - \frac{BS}{BS_{\mathrm{ref}}} \quad \text{Eq. 3}$$

where $\mathrm{BS}_{\mathrm{ref}}$ is the Brier score for the reference forecast. Positive BSS indicates a skilful forecast (i.e., better accuracy than the reference forecast) and negative BSS indicates a forecast with no skill (i.e., worse accuracy than the reference forecast). We use the climatological lightning frequency at each grid point, for each 3-hour period and day of the year, as our reference forecast (herein referred to as "climatology"). As indicated in the introduction, the use of a single performance metric restricts the conclusions that can be drawn from the results. However, the intent of this pilot study is to explore the potential for ML forecasting system, rather than to present a full analysis of their performance, which would require additional metrics such as the attributes diagram or the area under the receiver operator characteristic curve.

Verification is performed separately for each forecast base time (00 UTC, 06 UTC, 12 UTC and 18 UTC) and for the land and sea portions of the domain. This stratification of the results is important as the predictability

of lightning shows significant spatial and temporal variations. Predictability is known to be higher over land during the day compared to at night and over the sea.

3. RESULTS AND DISCUSSION

Figure 1 presents lightning observations on 1 January 2021 at 00 UTC (top left) and the corresponding lightning forecast from three models: the current CT system (top right) and two network models (bottom left and right). The forecasts appear comparable with similar strength and deficiencies: all three models broadly capture observed lightning activity over central New South Wales and the western part of Western Australia but overestimate the likelihood of lightning over northern Queensland and the northern part of the Northern Territory. Forecast probabilities generally do not exceed 0.4, suggesting low sharpness; however, Model 025 (bottom left plot) exhibits a higher upper limit of forecast probabilities compared to the other two models.

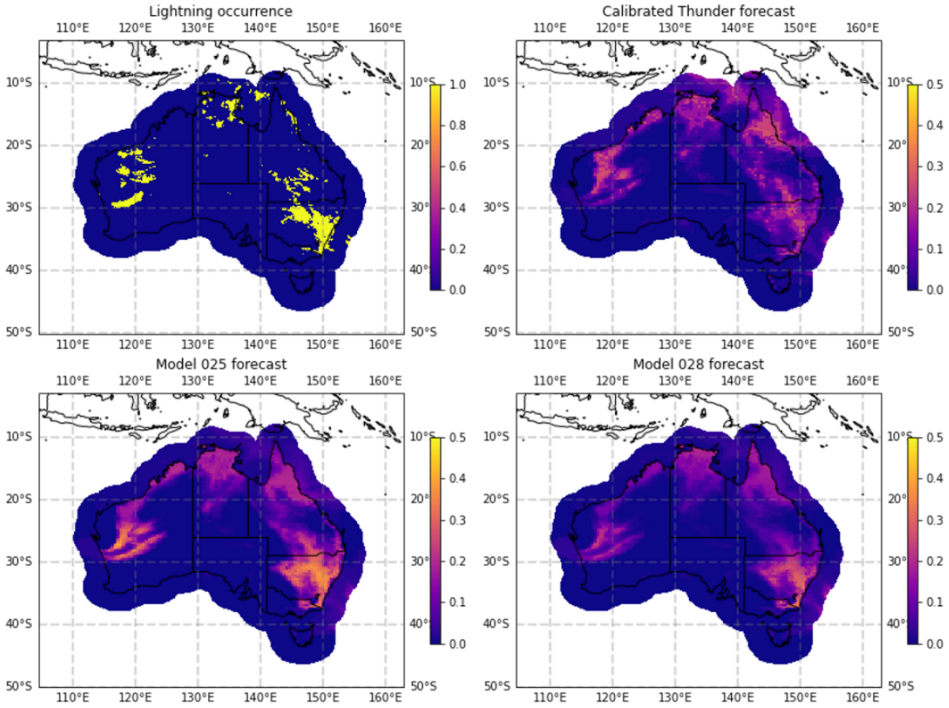

Figure 1. Example of lightning forecast on the 1st January 2021 at 00 UTC, 3 hour lead time.

Figure 2 shows the Brier skill score for the five models listed in Table 1 and the four base times, computed over the whole domain (left column), over land points only (middle column), and over sea points only (right column). All models generally outperform the lightning climatology for the same 3-hour period and spatial neighbourhood (BSS > 0), but skill remains modest. The best models appear in Figure 2b where BSS approaches 0.1, indicating a 10% reduction of the Brier score compared to climatology.

The network models listed in Table 1 outperform the current CT system for all domains and base times. This is particularly pronounced for the sea domain, where the current system performs poorly with negative BSS values compared to near-zero values for all network models. This result is significant for the future of the CT system because it indicates that off-the-shelf machine learning models can rival or even surpass forecasting tools operated by the BoM for lightning. It is worth noting that the performance of the different networks remains very similar, which suggests that network architecture does not need to be highly customised to deliver satisfactory performance. This is a positive outcome because it suggests that network models could be developed quickly without requiring much tuning, hence reducing the cost of initial development.

The results presented in this paper remain preliminary because they use a limited hindcast dataset of 90 days, and a single performance metric only. However, we believe that the satisfactory performance of network models shown in this paper corroborates the rise of machine learning in supporting operational weather forecasting as outlined by Düben et al. (2021). Additional questions remain to be explored related to the performance of the ML model compared to the current system. More specifically, it is not clear if the ML

model reduces forecast bias or improves correlation between predictors and predictand. Both forecast attributes contribute to the Brier score, which does not distinguish them. Additional metrics would be required to answer address this point.

Figure 2 also shows that the performance of lightning forecasts varies with the time of day the forecast is issued as revealed by the decrease in BSS from the top to bottom rows. Model skill, including that of the current CT system, is highest for forecasts issued at 00 and 06 UTC, which, for a lead time of 6h, corresponds to predicting lightning in the afternoon. This could be explained by the higher number of lightning strikes occurring in the afternoon, which provides more information for the lightning prediction model to calibrate their parameters.

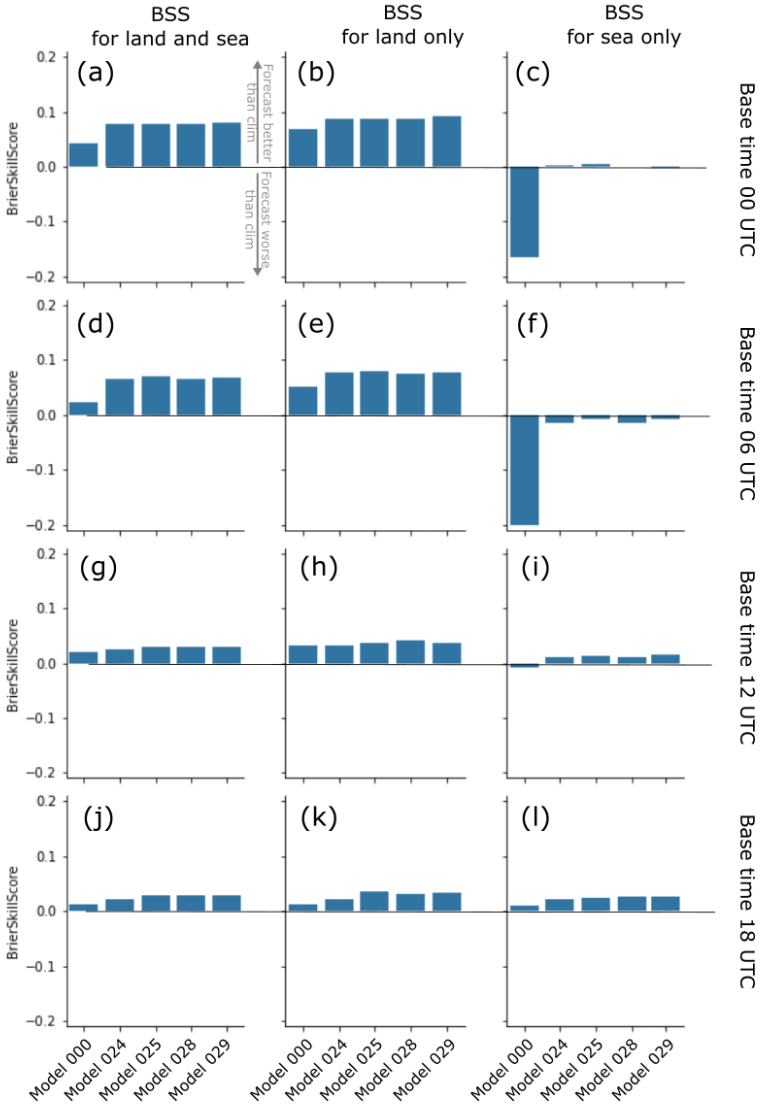

Figure 2. Brier skill score comparing the current CT system (Model 000) with the four network models.

4. CONCLUSION

This study represents a preliminary exploration of how a machine learning approach could be used to generate operational forecasting products in the Bureau of Meteorology. The objective was to compare feed-forward neural networks with the current operational system to generate forecasts on the likelihood of lightning strikes based on the Brier skill score verification metric.

The results presented here suggest that a network model based on off-the-shelf machine learning tools can rival or even surpass the performance of the current system when evaluated over a 90-day period and using a single performance metric (Brier skill score). The performance improvement was particularly noticeable when

predicting lightning over the sea, where the current operational system shows a consistent overforecasting bias. These preliminary results are encouraging and show the potential for machine learning to support lightning strike forecasts. By using off-the-shelf machine learning tools, this approach is cost effective, with limited customised development needed, offering the potential for rapid deployment to operations.

Two areas for future work are noted. First, the networks used in this study remain simple compared to modern deep learning architectures, such as those using recurrent layers (e.g., Hochreiter & Schmidhuber, 1997) or an encoder-decoder structure (e.g., Zhou et al., 2020). It is likely that the encouraging results presented here could be improved significantly with these modern architectures. Second, the physical predictors considered here (i.e., probability of CPTP > 1 and precipitation > 0.25 mm) could be enhanced by using the full set of ensemble values generated by the NWP model. These additional data provide information on the intensity of predicted thunderstorms, which is currently lost by applying thresholds to the predictors.

REFERENCES

Bright, D. R., Wandishin, M. S., Jewell, R. E., & Weiss, S. J. (2005). A physically based parameter for lightning prediction and its calibration in ensemble forecasts. *Preprints, Conf. on Meteor. Appl. of Lightning Data, Amer. Meteor. Soc., San Diego, CA*, *3496*, 30.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

Düben, P., Modigliani, U., Geer, A., Siemen, S., Pappenberger, F., Bauer, P., Brown, A., Palkovic, M., Raoult, B., & Wedi, N. (2021). Machine learning at ECMWF: A roadmap for the next 10 years. *ECMWF Technical Memoranda*, *878*.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., & Schepers, D. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Mostajabi, A., Finney, D. L., Rubinstein, M., & Rachidi, F. (2019). Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *Npj Climate and Atmospheric Science*, *2*(1), 1–15.

Rabbani, M., Oo, A. M. T., & Stojcevski, A. (2014). Analysis of lightning current characteristics to investigate lightning strike damages to energy pipeline. *2014 International Conference on Lightning Protection (ICLP)*, 528–532.

Ramachandran, P., Zoph, B., & Le, Q. V. (2017). Searching for activation functions. *ArXiv Preprint ArXiv:1710.05941*.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929–1958.

Warren, R., Richter, H., Caroll, M., & Blockley, I. (2021). *Upgrade of Calibrated Thunder : The GE3 Upgrade* (Issue Operations Bulletin Number 131).

Xie, H., Wu, L., Xie, W., Lin, Q., Liu, M., & Lin, Y. (2021). Improving ECMWF short-term intensive rainfall forecasts using generative adversarial nets and deep belief networks. *Atmospheric Research*, *249*, 105281.

Zhou, K., Zheng, Y., Dong, W., & Wang, T. (2020). A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *Journal of Atmospheric and Oceanic Technology*, *37*(5), 927–942.