JGR Atmospheres

RESEARCH ARTICLE

10.1029/2021JD035388

Key Points:

- Explainable artificial intelligence methods capture relevant large-scale quantities to predict convective activity
- Mid-level ω shows stronger relation to total convective area than to its degree of organization
- High degrees of convective organization in Darwin can occur in both very dry and moist environments

Correspondence to:

M. H. Retsch,
matthias.retsch@monash.edu

Citation:

Retsch, M. H., Jakob, C., & Singh, M. S. (2022). Identifying relations between deep convection and the large-scale atmosphere using explainable artificial intelligence. *Journal of Geophysical Research: Atmospheres*, *127*, e2021JD035388. https://doi.org/10.1029/2021JD035388

Received 10 JUN 2021
Accepted 7 JAN 2022

Identifying Relations Between Deep Convection and the Large-Scale Atmosphere Using Explainable Artificial Intelligence

M. H. Retsch[1,2], C. Jakob[1,2], and M. S. Singh[1,2]

[1]School of Earth, Atmosphere and Environment, Monash University, Melbourne, VIC, Australia, [2]Australian Research Council Centre of Excellence for Climate Extremes, Monash University, Melbourne, VIC, Australia

Abstract

Deep moist convection is responsible for a large fraction of rainfall in the tropics, but the interaction between deep convection and the large-scale atmosphere remains poorly understood. Here, we apply machine learning techniques to examine relationships between the large-scale state of the atmosphere and two measures of its convective state derived from radar observations in northern Australia: the total area occupied by deep convection and the degree of deep convective organization. Specifically, we use a neural net to predict convective area and convective organization as a function the large-scale state, defined as the thermodynamic and dynamic properties of the atmosphere averaged over the radar domain. Building on research into explainable artificial intelligence, we apply so-called "attribution methods" to quantify the most important large-scale quantities determining these predictions. We find that the large-scale vertical velocity is the most important contributor to the prediction of both convective measures, but for convective area, its absolute and relative influence are increased. Thermodynamic quantities like atmospheric moisture also contribute to the prediction of convective area, but they are found to be unimportant for convective organization. Instead, the horizontal wind field appears to be more relevant for the prediction of convective organization. The results highlight unique aspects of the large-scale state that are associated with organized convection.

Plain Language Summary

Deep moist convection, which can be typically encountered in a thunderstorm, is responsible for a large fraction of tropical rainfall. However, it is not clear how its development and presence interacts with the surrounding atmosphere on scales of kilometers. Thus we examine the relationship between the large-scale atmosphere and two measures of deep convection by using a neural net, that is, a machine learning approach. The neural net enables us to capture possible nonlinear relations when predicting two measures of convection with the large-scale atmospheric quantities as input. More interestingly, however, we also apply methods quantifying which large-scale quantities are most important for these predictions, and thus have possibly a strong relation to deep convection. We take radar measurements providing us with two measures of deep convective activity: (a) The total area occupied by deep convection, and (b) how strongly deep convection is spatially clustered. The large-scale atmosphere consists of the thermodynamic and dynamic properties over this radar domain. We find that ascending winds are contributing the most to the predictions of both convective measures, but more strongly to the convective area than to its spatial clustering. For strong clustering, the horizontal winds are also important, whereas atmospheric moisture is not.

1. Introduction

Precipitation from convective systems accounts for the majority of precipitated water in the tropics and therefore is a vital part of the hydrological cycle. Clouds formed by convective processes strongly affect the radiative heat balance of the atmosphere, which through energy constraints, feeds back directly on the hydrological cycle. This makes convection a key process in connecting the energy and water cycles on Earth. While the importance of convection is well recognized, computational constraints require convection to be parametrized in most global weather and climate models (e.g., Arakawa, 2004). Hence, an important question regarding convection is, how the relatively small convective cloud scales, and the heating produced in them, relate to larger atmospheric scales. An example of such a relationship is the well-known relation of convective precipitation to atmospheric moisture content (e.g., Bretherton et al., 2004; Holloway & Neelin, 2009; Schiro et al., 2018). At short timescales, observational studies have also revealed relationships between convective precipitation and various measures of (in) stability (e.g., Ahmed & Neelin, 2018; Davies et al., 2013; Louf et al., 2019). At longer timescales, such instability might be expected to be removed by convection itself (Arakawa & Schubert, 1974; Emanuel et al., 1994; Xu & Emanuel, 1989). It is important to note that many of the studies above consider an area-average of precipitation as a measure of convective activity, where the size of the chosen area is close to typical grid-boxes of general circulation models (GCMs).

A characteristic of convective activity that has received much recent attention in both model and observational studies is the organization of deep convection (e.g., Bretherton et al., 2005; Houze, 1977; Muller & Held, 2012; Tobin et al., 2012; Zipser, 1977). In models, the degree of convective organization is often assessed by the variance of moist static energy (MSE; e.g., Wing & Emanuel, 2014) or precipitable water (PW; e.g., Bretherton et al., 2005; Holloway & Woolnough, 2016) and the locations where MSE or PW are high are usually the locations where deep convection clusters and thus becomes aggregated. Observational studies of convective organization (see Holloway et al. (2017) for a review) often rely on satellite-borne observations, and several studies have proposed simple indices to characterize convective organization applicable to both observations and models. For example, Tobin et al. (2012) defined the Simple Convective Aggregation Index (SCAI), which uses radiation measurements from satellites to assess the degree of aggregation in classes of convective precipitation rate. They found that increased convective aggregation, as assessed by SCAI, is associated with a decrease in tropospheric relative humidity. Following the development of SCAI, other simple organization measures have been proposed (Kadoya & Masunaga, 2018; Tompkins & Semie, 2017; White et al., 2018), including the Radar Organisation Metric (ROME), recently introduced by Retsch et al. (2020) for use with radar observations, which will be used extensively in this study.

Convective organization has long been known to be an important characteristic of deep convection, as it is responsible in setting the location of intensive rain and accompanying severe weather (e.g., Leary & Houze, 1979; Tan et al., 2015; Weickmann & Khalsa, 1990) and might even play a role for Earth's equilibrium climate sensitivity through its effect on radiative fluxes (e.g., Becker & Wing, 2020; Bony et al., 2016). Despite much recent work, relatively little is known about the relationship of convective organization with the large-scale state of the atmosphere, a central theme of this study.

The main goal of this study is to identify and quantify the dominant relations between large-scale variables and deep convective activity, based on radar observations near Darwin, Australia. Here, large-scale variables refer to thermodynamic and dynamic properties of the atmosphere averaged over the radar domain, roughly corresponding to the resolved scale of a GCM. We further wish to investigate if explainable artificial intelligence (XAI) techniques, and in particular "attribution methods" (e.g., Mamalakis et al., 2021), applied to a neural network (NN) can reveal the main influences of the large-scale atmosphere on convection. We first test three attribution methods by seeking to reproduce the results of a recent study by Louf et al. (2019), using the same set of radar observations. They identified a strong connection of the large-scale vertical velocity and relative humidity to the overall convective activity, as measured by the total area experiencing convection, referred to as total convective area (TCA) here. We then apply these same three attribution methods to identify the main relationships between the large-scale state and convective organization specifically.

We use ROME to assess convective organization. This measure of organization has been specifically defined and validated for use with radar data sets and has shown skill in identifying relationships of convective organization with different states of the Australian monsoon (Retsch et al., 2020). ROME quantifies the degree of organization of a radar scene based on the arrangement of contiguous regions of convective activity. It is based on the principle that convective organization requires interaction between multiple individual convective updrafts. Such interaction is promoted by the existence of large contiguous regions of convective activity (that may contain multiple updrafts), or multiple convective regions in close proximity.

We prefer the use of NNs over other approaches to establish relations between time series, such as multiple linear regression (MLR, e.g., Igel & van den Heever, 2015), because we want potential nonlinear relationships to be captured, which is easily accomplished by applying NNs. NNs can be thought of as "predictive tools" and in our specific example, we can apply an NN to predict either TCA or ROME as a function of the large-scale atmospheric state. However, the main aim of using the NN approach in this study is to "look inside" the NN to quantitatively infer, which large-scale variables contribute the most to the predictions of the convective activity and organization. The general use of NNs to advance our understanding of observed relationships is an emerging area in atmospheric and climate science (e.g., McGovern et al., 2019; Toms et al., 2020) and several techniques, known as attribution methods, exist to identify the most influential input variables of an NN (e.g., Bach et al., 2015; Simonyan et al., 2014; Zhou et al., 2016). Here, we apply Integrated Gradients (IG; Sundararajan et al., 2017) and

Layer-wise Relevance Propagation (LRP; Bach et al., 2015), which both assign each input of an NN with a value representing the input's significance for a single prediction. In addition to applying a traditional LRP method we also develop a "percentage-backtracking" implementation of LRP.

To achieve our goals, the paper is structured as follows. We first present the data sets used in Section 2, followed by an introduction to our methodology involving the neural net and attribution methods in Section 3. Section 4 then presents the relations of the large-scale state to the area and degree of organization of convection. Finally, Sections 5 and 6 discuss and summarize our results.

2. The Data Sets

To infer relations between deep convection and the large-scale atmospheric state, we use radar observations from Darwin in Australia to represent the convective state combined with a large-scale analysis data set that blends ECMWF analyses with local observations (Davies et al., 2013). Convection in the Darwin area occurs mainly in the wet season, and we therefore only examine data in the months of October–March. We use 12 wet seasons of concurrent radar and large-scale observations for the years 2001–2007 and 2009–2015. Radar observations are available every 10 min and large-scale information is available every 6 hr. All together this provides 1,345 days of concurrently observed convective and large-scale information at the radar site, including 191,952 individual radar scans.

The radar used to observe deep convection is "C-POL", which is a C-band polarimetric radar located close to Darwin in northern Australia (Keenan et al., 1989). It is located at 12.249°S and 131.044°W and covers a radius of 138.75 km, enabling it to capture sufficiently large convective systems for our purpose of characterizing overall convective activity and organization.

2.1. Total Convective Area and the Radar Organisation Metric

We quantify the convective state of the atmosphere in two ways. We use the total area that is experiencing convection, or TCA, as a measure of the overall convective activity within the radar domain, and ROME (Retsch et al., 2020) to quantify convective organisation. It has long been known that the TCA relates very strongly to area average convective rainfall (e.g., Byers & Union, 1948; Davies et al., 2013; Doneaud et al., 1984). Indeed, Louf et al. (2019) found a correlation of 0.98 between TCA and the average convective rainfall over the radar domain used here. ROME quantifies convective organization based on the arrangement of contiguous convective regions within the radar domain and is described further below.

Determining both the TCA and ROME requires a classification of precipitating pixels in the radar scene to identify those most likely associated with convection. Here, we apply the algorithm of Steiner et al. (1995) to reflectivity fields interpolated to a Cartesian grid of 2.5 by 2.5 km^2 at 2.5 km height. The Steiner algorithm is commonly used in studies of radar-derived rainfall estimates (e.g., Hitchcock et al., 2021; Louf et al., 2019). It defines convective pixels based on regions of high absolute reflectivity or regions with high reflectivity relative to their immediate surroundings. Full details of the algorithm may be found in Steiner et al. (1995). Once convective pixels have been identified in a scene, the TCA may then be calculated simply by multiplying the total number of convective pixels with the area of one pixel.

To quantify convective organization using the ROME metric, we further define convective objects as regions of contiguous convective pixels, where a pixel is contiguous with another if their edges or corners touch (8-connectivity). Unlike other recently derived organization metrics that are based on the spatial proximity of convective objects (Tobin et al., 2012; Tompkins & Semie, 2017; White et al., 2018), ROME is defined to take both the proximity and size of convective objects into account. This definition is based on the assumption that individual updrafts that are in close proximity are more likely to interact with their environment in a way that enables the further growth of the convective area (e.g., Fovell & Tan, 1998; Redelsperger & Lafore, 1988). As the Steiner algorithm "does not attempt to identify echoes produced by individual updrafts, but rather attempts to delineate general regions of active convection" (Houze, 2014), large convective objects are likely to have more than one convective updraft within them, such that their size is important for assessing organisation. Retsch et al. (2020) have shown that ROME, when applied to radar observations, captures convective organization more accurately

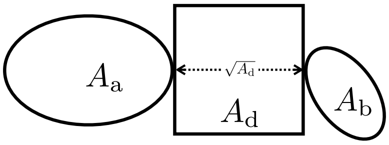

Figure 1. Schematic of two convective objects *a* and *b* with their respective areas A_a and A_b as well as the distance area A_d between them.

than other metrics that consider only the arrangement of convective objects and not their size. The details of ROME are provided in Retsch et al. (2020), but we will briefly summarize the main idea here.

ROME is constructed by defining "connections“ between any two convective objects *a* and *b* and assigning a scalar value c_i to each connection representing its strength. For a given radar scene, ROME is evaluated by determining all unique connections between every pair of convective objects in the scene and then taking the arithmetic mean of all scalar c_i values (Equation 1). Here, the scalar c_i describes a single connection. Figure 1 shows the main terms used to calculate c_i. For the connection between two objects *a* and *b*, the scalar c_i is defined as the area of a pair's larger object A_a [km^2] plus the area of the smaller object A_b [km^2], weighted by the distance between the two. The weighting is calculated as the ratio of A_b to the area between the two objects, A_d, where A_d is the square of the shortest distance *d* between the objects. Therefore, the smaller or further away *b* is from *a*, the less will be added to area A_a. If object *a* is the only convective object in a radar scene, then nothing will be added to its area. For *n* objects with $K = \frac{n\cdot(n-1)}{2}$ unique connections, ROME is then defined as

$$\mathrm{ROME} := \begin{cases} \frac{1}{K}\sum_{i=1}^{K} c_i, & \text{for } n > 1 \\ A_\mathrm{a}, & \text{for } n = 1 \end{cases}, \qquad c := A_\mathrm{a} + \min\left(1, \frac{A_\mathrm{b}}{A_\mathrm{d}}\right)\cdot A_\mathrm{b}. \tag{1}$$

ROME attains a higher value for higher degrees of organisation, and because it is not only sensitive to the distance between convective objects but also their area, high ROME indicates a convective state with large clustered clouds.

2.2. The Large-Scale Atmospheric State

The data set for the large-scale atmospheric state is obtained by applying a variational-analysis technique proposed by Zhang and Lin (1997) to the ECMWF operational analysis. The technique adjusts the ECMWF analysis such that it balances independently observed vertically integrated budgets of mass, energy, and moisture, obtained from satellite and radar observations around the Darwin region. The budgets are given by independent observations of the radiative fluxes at the top of the atmosphere and surface as well as surface precipitation, evaporation and pressure. The variational-analysis of Zhang and Lin (1997) was originally designed to adjust observations of a radiosonde array, but as observations from radiosonde arrays are not available outside of dedicated field campaigns, we treat grid points from the ECMWF analysis as "pseudo-radiosondes." Combining these with satellite and radar observations has been shown to provide a more accurate estimate of the large-scale state of the tropical atmosphere than using the ECMWF analysis directly (Davies et al., 2013; Xie et al., 2004). A comparison of our large-scale input data set with the TWP-ICE field study (May et al., 2008) confirms this to be true for all of the key variables used in this study (not shown). Because the variational-analysis involves an adjustment to balance the energy and water budgets, it also produces estimates for the tendency and advection of water vapor and dry static energy in addition to the usual thermodynamic and dynamic fields. Since the tendency and advection of thermodynamic quantities are potentially useful in relating the large-scale state to the convective state, we allow some of these variables to act as inputs to the NN.

The large-scale state estimate is available every 6 hr, at 3:30 a.m., 9:30 a.m., 3:30 p.m., and 9:30 p.m. local Darwin time. Variables contained in the large-scale data set are either one- or two-dimensional, that is either scalar time series or profile time series with 39 pressure levels, from 1015 hPa to 40 hPa in 25 hPa intervals. As the large-scale data set is available for the years 2001–2015, except for a two year gap between 2007 and 2009 when there was no radar data available, we have a total of 12 wet season covered by both the large-scale and the radar data set. A recent study by Louf et al. (2019) successfully applied both data sets to qualitatively infer relations between the large-scale state and convective properties, providing confidence to our desire to link them more quantitatively using an XAI approach. More details on the construction of the large-scale data set can be found in Davies et al. (2013).

To each time step in the large-scale data set, there is a corresponding TCA or ROME, as observed by the radar. Because the large-scale data has a time resolution of 6 hr, we average TCA over 6 hr, from three hours before to three hours after the time the large-scale data is available. As TCA is strongly related to area-mean convective rainfall, its average value is representative for a 6-hourly rainfall average. While averaging over 6 hr is an obvious choice for TCA, it is less clear what a 6-hourly mean of the organization index ROME represents. For example, if strongly organized convection only existed for one of the 6 hours, the average of ROME might be similar to a situation where convection was present throughout but not organized at all. We therefore consider the maximum ROME inside the ±3 hrs around the large-scale time as a more representative value of ROME over the 6 hr time interval. To avoid selecting outliers of ROME, we average its values over the 20 minutes before and after the time of the maximum and thereby create a time-series which matches the times of the large-scale state data. We refer to this time series as the time series of 6hr-maximum ROME. Applying the downsampling yields a time series for TCA and ROME with 5,089 samples each. The 6hr-average TCA has a mean and standard deviation of (501 ± 647) km^2 and the 6hr-maximum ROME of (157 ± 149) km^2.

3. Methodology

The goal of this study is to use XAI methods to better understand how convection and the large-scale state interact, and in particular, which variables may relate to convective organization. Our approach to achieve this is to first choose a quantity for which the relationship to the large-scale state is relatively well-known. Based on the recent work of Louf et al. (2019), we choose TCA as this quantity. This part of the study is meant to identify strengths and weaknesses of the approach. We then apply the same techniques to convective organization represented by ROME.

The connection between the datasets is made by a mapping from the large-scale data set as input to TCA and ROME as output using a "multilayer perceptron" (MLP), which is a class of neural networks. The mapping may be interpreted as an attempt to predict the value of TCA or ROME, given the large-scale atmospheric state. We refer to this as the "prediction" step. A necessary condition to find meaningful relationships is that the predictions are credible, and the model architecture and predictive skill is presented in the next section. Having made the predictions, we apply three XAI methods to identify those input variables that are most relevant to the predicted value. These kind of methods are known as "attribution methods" and in this study, we apply Integrated Gradients (Sundararajan et al., 2017) and two implementations of Layer-wise Relevance Propagation (LRP; Bach et al., 2015; Montavon et al., 2018). The attribution methods are presented in Section 3.2. For both the predictions and attributions, we compare the results from the NN to those from applying multiple linear regression as a baseline.

3.1. The Predictions

Here, we present the method to predict TCA and ROME using a neural network. We choose to employ a neural net for this task, because it can not only be used for predictions, but also for studying the influence of different input variables. Specifically, we are using a regression MLP. The MLP is an archetypical neural net in which each node in a layer is connected to every node in the previous layer. Because both MLPs used here, one to predict TCA and one to predict ROME, solve a regression task, they each have a single output node yielding TCA_{NN}, the predicted value of TCA, and ROME_{NN}, the predicted value of ROME, respectively. For a more detailed description of how a prediction is calculated by an MLP, see Appendix A1.

As mentioned above, the large-scale data set serves as the input to the NN. The large-scale data set includes both vertical profiles (e.g., relative humidity) as well as scalar variables (e.g., convective available potential energy) derived from the large-scale analysis fields. To enable comparability between variables in the large-scale data set, each variable is normalized and nondimensionalized by subtracting its mean and dividing by its standard deviation. Further, to avoid high correlations between different levels within the vertical profile inputs, we apply an empirical orthogonal function (EOF) analysis to each variable that has a vertical profile. The resulting principal component (PC) time series of these EOFs are uncorrelated by construction and are taken as the input to the NNs. For each variable, we retain as many EOF patterns as are needed to explain 90% of the variance in the original data. We also choose to discard variables available in the large-scale data set that have a correlation higher than ±0.6 to other variables. While, given enough data, MLPs allow for the extraction of useful relationships even with

Table 1
Variables of the Large-Scale Data Set Used as Input to the Neural Net to Predict 6hr-Maximum ROME and 6hr-Average TCA

Profile variables			Scalars	
Variable	Abbreviation	# EOFs	Variable	Abbreviation
Vertical velocity [hPa/hour]	ω	4	Precipitable water	PW
Zonal wind [m/s]	u	4	Longwave radiation at top of atmosphere	OLR
Meridional wind [m/s]	v	5	Surface latent heat flux	LH
Vertical wind shear [s^{-1}]	dUdz	11	Surface sensible heat flux	SH
Dry static energy [K]	S	5	Convective Available Potential Energy	CAPE
Advection of s [K/hour]	adv(s)	8	Convective Inhibition	CIN
Tendency of s [K/hour]	dsdt	8		
Relative humidity [%]	RH	4		
Advection of mixing ratio [g/kg/hour]	adv(r)	6		
Tendency of mixing ratio [g/kg/hour]	drdt	7		

Note. For the profile variables the number of used EOFs is given as well.

highly correlated inputs, such correlation of the input data makes the attribution of the results to individual inputs more difficult. Our desire for clear explainability of the MLP, when using the attribution methods presented below, motivates our use of EOF patterns and the discarding of highly correlated inputs. We discard mostly surface and radiative flux variables (see Table A1). The remaining variables result in a vector $\vec{x}_t$, containing the whole input for one prediction of TCA or ROME at time t. Each element $x_{i,t}$ represents one input variable. The complete list of input variables is given in Table 1.

3.1.1. Model Architecture

When setting up an NN for a specific task, there are many parameters to select, so-called "hyper-parameters", which define the size and inner workings of the NN and can affect its predictive skill. Thus, the first step is to split the available input and output data into three parts, namely the training, validation, and test set, and use the first two parts to find the appropriate combination of hyper-parameters for the task (Beucler et al., 2021). Once the hyper-parameters are chosen, and the NN is trained, the predictive skill of the NN is assessed using the test set. Here, the data is broken into chunks of 10 consecutive samples; samples 1–4 and 7–9 are placed into the training set (70% in total), sample 5 and 6 into the validation set (20% in total) and sample 10 into the test set (10% in total). This breakdown ensures each set contains samples at all times of day and across the whole data set, avoiding biases introduced by uneven diurnal sampling and trends, respectively.

To determine the hyper-parameters, we consider a variety of MLP architectures and apply the KerasTuner package (O'Malley et al., 2019) to choose the most appropriate architecture for our task. Specifically, our search space includes MLPs with 1–5 layers each with 30–1000 nodes, and we vary the learning rate of stochastic gradient descent, a parameter determining how the node weights are optimized, between 0.1 and 10^{-5}. In all cases, we use the Rectified Linear Unit (ReLU) as the activation function, as it works well with the LRP method. Given the regression task and our interest in high values of ROME, we want the loss function to emphasize errors at high ROME, and thus, we take the loss function to be the mean squared error. Both MLPs used here, one to predict TCA and one to predict ROME, are set up in the same way. The optimum structure in both cases with the least validation loss and overtraining is found to be three layers with 300 nodes and a learning rate of 10^{-4}. We use these hyper-parameters and save the weights and biases of each epoch to find the epoch with the global minimum validation loss. With this procedure, the trained net for TCA is obtained after 61 epochs, and for ROME after 39 epochs.

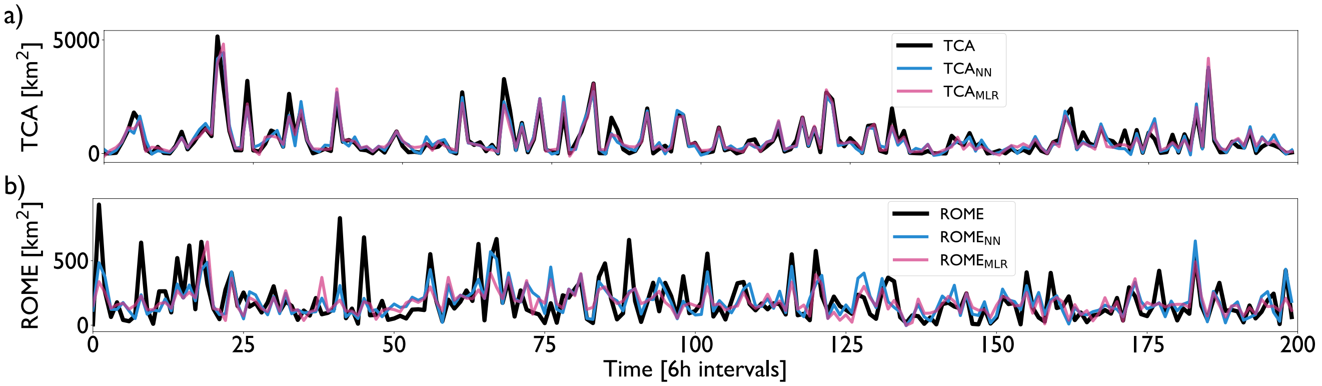

Figure 2. Observed (a) 6hr-average total convective area (TCA) and (b) 6hr-maximum Radar Organisation Metric (ROME) and their predictions by the neural net (TCA_{NN} and $ROME_{NN}$) and multiple linear regression (TCA_{MLR} and $ROME_{MLR}$). Shown are 200 time steps joined together from the test data set, which is the data not used to train or validate the neural net.

3.1.2. Comparison to Multiple Linear Regression

Having trained both NNs, that is, with the weights and biases of the NNs fixed, we now predict both ROME and TCA for the test set and we compare these NN predictions to those from a simple MLR using ordinary least squares regression with the same input variables.

For the total convective area, TCA_{NN} and its predictions by MLR (TCA_{MLR}) are both well aligned with the observed TCA and have a correlation of 0.91 and 0.90, respectively (Figure 2a). When considering the total data set, the correlation of TCA_{NN} to observed TCA is 0.95. The mean and standard deviation of TCA_{NN} is (487 ± 613) km^2 and for TCA_{MLR}, it is (501 ± 586) km^2. Both approaches, NN and MLR, match the target of (501 ± 647) km^2 closely, and with only small differences between NN and MLR, TCA seems to be well predictable by a linear combination of the large-scale variables. However, a closer inspection of the frequency distributions of the respective predictions in Figure 3a reveals that TCA_{NN} approximates the observed TCA distribution more closely than TCA_{MLR}, yet the mode of both distributions is at a smaller value of TCA than observed.

For ROME, the correlation of $ROME_{NN}$ to ROME in the test set is 0.62 and for $ROME_{MLR}$, it is 0.53. Thus, the predictions of ROME are generally less accurate than for TCA, but $ROME_{NN}$ has a slightly higher correlation than $ROME_{MLR}$. Figure 2b shows that the predictions by NN and MLR are more correlated to each other (by 0.81) than they are to the observations of ROME, however $ROME_{NN}$ predicts higher values of ROME more accurately. When considering the total data set, the correlation of $ROME_{NN}$ to observed ROME is 0.70. The mean and standard deviation of $ROME_{NN}$ is (157 ± 101) km^2 and for $ROME_{MLR}$, it is (157 ± 80) km^2. With the target of (157 ± 149) km^2, the mean is again well matched by both the NN and MLR, but the variance is larger and closer to the observed value for $ROME_{NN}$ than for $ROME_{MLR}$. The mean squared error for $ROME_{NN}$ is also lower than for $ROME_{MLR}$, with 11,348 km^4 instead of 15,847 km^4. The distributions across all NN- and MLR-predictions

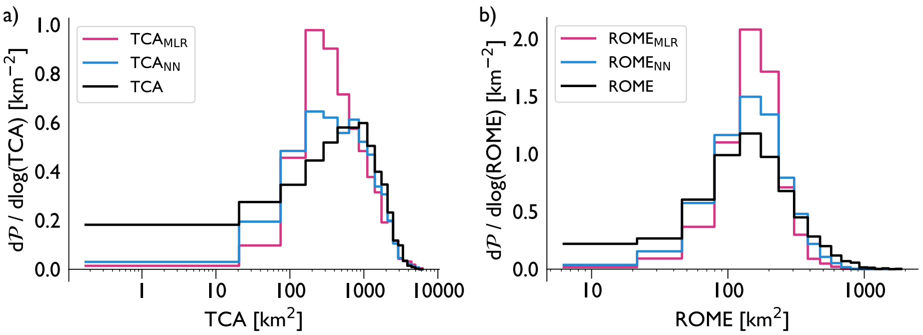

Figure 3. Histogram over the total time series of (a) total convective area (TCA) and (b) Radar Organisation Metric (ROME) and their predictions by the neural net (TCA_{NN} and $ROME_{NN}$) and multiple linear regression (TCA_{MLR} and $ROME_{MLR}$), with probability density on the *y*-axis.

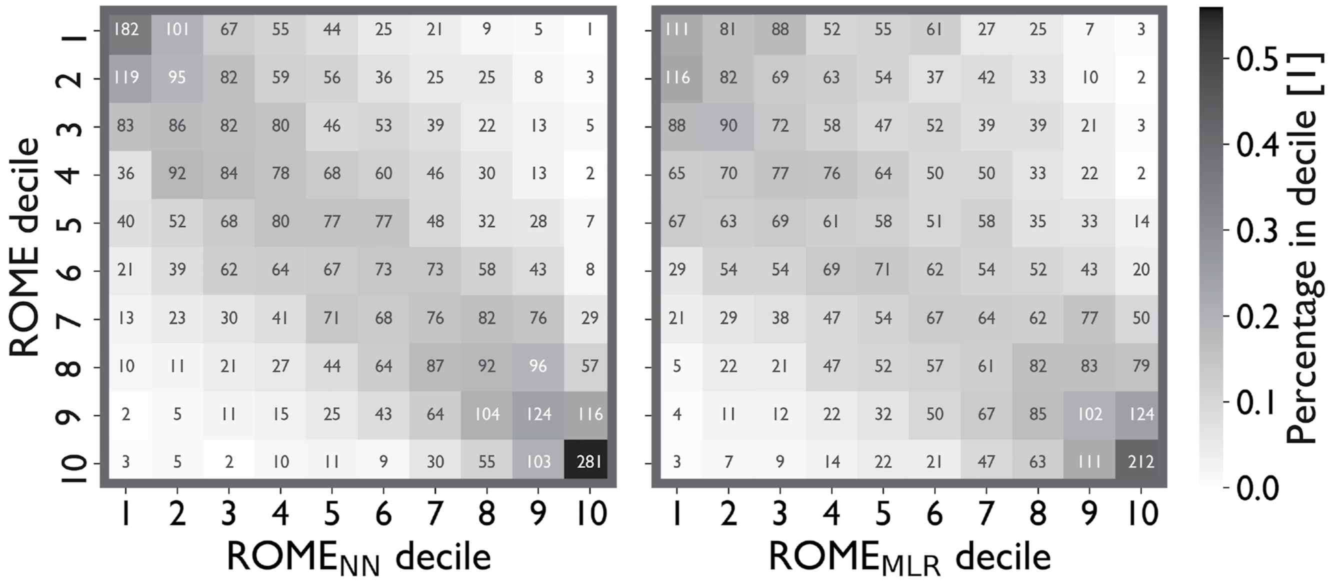

Figure 4. Confusion matrix of the total time series of Radar Organisation Metric (ROME) and its predictions by the neural net (ROME$_{NN}$, on the left) and multiple linear regression (ROME$_{MLR}$, on the right). The *y*-axis is binned into deciles of observed ROME and the *x*-axis is binned into deciles of the respective predictions.

in Figure 3b also show that the predictions by NN align better with observed ROME, because the distribution of ROME$_{NN}$ is broader than the distribution of ROME$_{MLR}$ and thereby matches the observed distribution more closely at high and low values of ROME. Thus the distributions indicate that the NN-predictions not only predict the mean well, but also have less error for both high and low ROME.

The differences between the NN- and MLR-predictions are also visible in the confusion matrix between deciles of observed ROME and deciles of its predictions (Figure 4). Instead of assessing the absolute error between prediction and target, the confusion matrix shows how often the respective deciles of prediction and target match each other. Thus, the confusion matrix is not affected by the different variances of ROME$_{NN}$ and ROME$_{MLR}$. Both ROME$_{NN}$ and ROME$_{MLR}$ most often predict a value in the same respective decile when ROME is either in its highest or lowest decile. However for ROME$_{NN}$, the number of matches is higher than for ROME$_{MLR}$ at both high and low ROME. For the other deciles, especially around the median of ROME, the predicted values of ROME$_{NN}$ and ROME$_{MLR}$ do not match with the observed decile of ROME as often. Yet a decile of ROME$_{NN}$ is mostly close to the observed decile, that is, the deciles are more concentrated along the diagonal in Figure 4, whereas a decile of ROME$_{MLR}$ is much less related to the observed decile. Therefore, even when we neglect the greater variance of ROME$_{NN}$, which may contribute to its reduced error at high ROME, and compare the confusion matrices, the predictions of ROME$_{NN}$ are better aligned with observed ROME than the predictions of ROME$_{MLR}$.

From the above comparison, we conclude that predictions using MLR are comparably accurate to those of the NN for TCA, but the NN is more accurate for predictions of ROME. Thus, a nonlinear approach is not necessary for a prediction of TCA, but it improves a prediction of ROME.

In the next section, we introduce the attribution methods which we will apply to the NN-predictions. To maintain comparability between TCA and ROME, we also apply these methods to TCA$_{NN}$, even though a nonlinear approach might not be absolutely necessary to predict TCA.

3.2. The Attribution Methods

In this section, we describe how we identify the input variables with the highest relevance to the predicted values of TCA$_{NN}$ and ROME$_{NN}$. We provide a summary of the methods here with more detail on the proposed percentage-backtracking implementation of LRP available in Appendix A2.

3.2.1. Integrated Gradients

The method of Integrated Gradients evaluates how the output of an NN changes, between a specified "baseline input" and the actual input of a sample, and it then attributes a portion of this change to each input node. Suppose our NN may be expressed as the function $F : \mathcal{R}^n \rightarrow \mathcal{R}$, which encodes the output (a single output value in our case) based on the input vector $\vec{x}$, and we further define $\hat{\vec{x}}$ to be the baseline input. Then, for an NN with ReLU as the activation function it holds that

$$\sum_{i=1}^{n} \mathrm{IG}_i = F(\vec{x}) - F(\hat{\vec{x}}) \tag{2}$$

where IG_i is the attribution to input i in the input vector $\vec{x}$. To obtain the attribution for x_i we compute the gradient of the output with respect to the input variable, that is, the sensitivity $\frac{\partial F}{\partial x_i}$, integrate those gradients over all inputs that fall on the straight-line path between the input and the baseline vector, and multiply these integrated gradients with the difference of x_i to the baseline:

$$\mathrm{IG}_i = (x_i - \hat{x}_i) \cdot \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F\left(\hat{x} + \frac{k}{m}(x - \hat{x})\right)}{\partial x_i} \tag{3}$$

with m steps for the Riemann-approximated integral. The baseline input needs to be selected such that it "conveys a complete absence of signal, so that the features that are apparent from the attributions are properties only of the input, and not of the baseline (Sundararajan et al., 2017)". It is recommended that the baseline has a near-zero output and a common choice is the all-zero input vector, which we will also use in this study and which corresponds to the time-mean state, because our inputs are defined with the mean subtracted.

3.2.2. LRP-$\alpha\beta$

The basic idea of LRP is to decompose the output value of a neural net into contributions, called "relevances", from each individual input. To compute the relevances, we start at the output value of an NN and calculate how much each node in the last layer of the network contributed to this output value. We then repeat this decomposition for all nodes in all previous, that is, leftward, layers, step by step, until we reach the nodes in the input layer. Thus, contributions to the output value are propagated backwards through the network. The main principle of LRP is based on the fact that for each computed node-value during a prediction, that is, in a forward-pass of an NN, the scalar product is taken between the node-values in the previous layer and a set of unique weights. Thus, when computing this scalar product, each summand by itself already is a contribution from a specific node in the previous layer.

LRP is based on the idea that one may define a "relevance" of a given node that represents its contribution to the values of the nodes in the next (rightward) layer. Specifically, the relevance of a single node i in layer l, R_i^l, is given by the sum of its contributions to each node k in the layer $l + 1$:

$$R_i^l = \sum_{k=1}^{n} R_{ik}^l \tag{4}$$

where layer $l + 1$ has a total of n nodes and the contributions R_{ik}^l are given by,

$$R_{ik}^l = \frac{v_i^l w_{ik}^l}{\sum_{h=1}^{m} v_h^l w_{hk}^l} \cdot R_k^{l+1} \tag{5}$$

Here, layer l has a total of m nodes and v_i is the node-value after a prediction, that is, forward-pass, and w_{ik} is the fixed weight between nodes i and k (also compare Equation 16 in Bach et al. (2015)). However, the ratio in Equation 5 becomes numerically unstable when the denominator approaches zero. To avoid a division by zero, Bach et al. (2015) introduce "rules", which modify the above implementation.

Perhaps the most common rule in LRP is the $\alpha\beta$-rule (e.g., Dobrescu et al., 2019; Mamalakis et al., 2021; Montavon et al., 2018; Toms et al., 2020) which separates the positive summands from the negative summands in

Equation 5 and divides the contribution of node i to node k only by the respective sum of the positive/negative summands. A positive relevance is weighted by α and a negative relevance by β. Because a contribution from a node i to node k can only either be positive or negative, the total relevance for node i now consists of two sums

$$R_i^l = \alpha \cdot \sum_{k=1}^{n^{\text{pos}}} \frac{\left(v_i^l w_{ik}^l\right)^+}{\sum_{h=1}^{m^{\text{pos}}} \left(v_h^l w_{hk}^l\right)^+} R_k^{l+1} - \beta \cdot \sum_{k=1}^{n^{\text{neg}}} \frac{\left(v_i^l w_{ik}^l\right)^-}{\sum_{h=1}^{m^{\text{neg}}} \left(v_h^l w_{hk}^l\right)^-} R_k^{l+1} \tag{6}$$

where node i contributes positively to n^{pos} nodes and negatively to n^{neg} nodes in layer $l+1$. $()^+$ and $()^-$ denote the sign of a term in the scalar product of the weights and nodes in layer l and there are a total of m^{pos} positive terms and m^{neg} negative terms. The two parameters α and β need to be selected such that $\beta > 0$ and $\alpha - \beta = 1$, which is necessary to conserve the total amount of back-propagated relevance. Throughout this study, we use values of $\alpha = 2$ and $\beta = 1$.

3.2.3. LRP-p

Both attribution methods presented above require manually selecting parameters, which for IG is the baseline input and for LRP-$\alpha\beta$ are α and β. Depending on the choice of parameters the attribution results will differ. Thus, additionally to the two methods described above, we also propose a rule for LRP which works without having to choose any parameters. Because the rule only considers percentages during the back-propagation process, we will refer to it as percentage-backtracking, or LRP-p in short.

LRP-p is based on the same principle as LRP-$\alpha\beta$, but instead of separating the positive and negative part to avoid numerical instability, we reduce the chances of dividing by zero by adding the bias b of node k in the denominator, such that a contribution from node i to node k becomes

$$R_{ik}^l = \frac{v_i^l w_{ik}^l}{\sum_{h=1}^{m} v_h^l w_{hk}^l + b_k^{l+1}} \cdot R_k^{l+1} \tag{7}$$

By adding the bias, a division by zero is avoided as long as *either* the scalar product or the bias are nonzero. Adding the bias does not rule out the possibility of dividing by zero, but the possibility is greatly reduced compared to Equation 5. Equation 5 becomes numerically unstable each time the scalar product in the denominator approaches zero, for example, for a nearzero input vector. When the inputs are normalized variables like in our case, a nearzero input vector can occur regularly, as it would represent the average state in all input variables. In Equation 7, the denominator approaches zero when the scalar product offsets the bias: $\sum_{h=1}^{m} v_h^l w_{hk}^l \simeq b_k^{l+1}$. However, an offset between scalar product and bias occurs much less frequently, because the bias is trained to adjust the magnitude of the node value consistently and not to be offset by the trained weights. Thus, by using the bias as a stabilizing term in the denominator, there is no requirement to select a specific parameter to apply LRP.

As the rules of LRP are about back-propagating relevance through an NN, the question arises what the relevance at the starting point of the back-propagation is, that is, at the output node. For LRP-$\alpha\beta$, it is the actual value of the output node. Because this "amount" of relevance is conserved during back-propagation, the attributed relevances at the input nodes add up to the predicted value. That means that if the predicted value is high, the attributed relevance will be high likewise. To enable comparability across attributions for multiple predictions, independent of the predicted value, for LRP-p we take 100 as the starting relevance. As a result, the relevance becomes the percentage with which a node contributed to the following node-values *before* applying their activation function, that is, it is a percentage of "pre-activations". The percentages obtained through LRP-p differ from percentages obtained through simply dividing the relevances from LRP-$\alpha\beta$ by the predicted value, because the α and β parameter weigh the relevances differently, depending on their sign. A detailed description of LRP-p and the differences to LRP-$\alpha\beta$ is provided in Appendix A2.

3.2.4. An Example

We next illustrate the results from the attribution methods using a simple example, which has strong linear relations between the input and output. In this example, we construct a sample of 125,000 predictands, y, based on random samples from the normal distribution $\mathcal{N}(0, 1)$. Given the target series y, we create three input variables x_1, x_2, and x_3, which are designed so that x_1 has a correlation of 0.3, x_2 a correlation of 0.8 and x_3 a correlation of 0 to y. The input variables are computed from y via $x_i = y \cdot r + z \cdot \sqrt{1 - r^2}$, where r is the desired correlation coefficient, that is, 0.3, 0.8, and 0 in our case, and z is a noncorrelated series to y. z is sampled from the same underlying distribution as y and the noncorrelation between z and y is ensured by taking the PC series as the z and y series after applying principal component analysis (PCA) to two sample series of $\mathcal{N}(0, 1)$. We train two MLPs with identical setups (ReLU-activation, mean squared error loss, 10^{-3} learning rate, 15 epochs, 70% training data), but one MLP has one hidden layer with only two nodes (compare also to Figure A1 in the Appendix) and one MLP has two hidden layers with 150 nodes each. We note that for this simple example with its constructed strong relations between input and output, the size of the small MLP is sufficient to capture the relations and achieve accurate predictions with a mean absolute error of $4{\cdot}10^{-6}$. The purpose of the large MLP is to investigate if the MLP's size affects the attribution methods in any way.

We apply IG, LRP-$\alpha\beta$ and LRP-p to each prediction. For IG, we choose the all-zero input vector as the baseline and for LRP-$\alpha\beta$ we choose $\alpha = 2$ and $\beta = 1$, denoted LRP-$\alpha_2\beta_1$, because LRP was found to perform well when both positive and negative contributions are deemed important (Dobrescu et al., 2019; Montavon et al., 2018).

The methods described above provide a quantitative attribution of each of the inputs x_i to a given prediction of y. We will refer to these attribution values as the "contribution" of a given input to the output, irrespective of the particular attribution method. A-priori we expect x_2 to be the most contributing input because it is most strongly correlated to y and therefore a good predictor. We expect x_1 to contribute less than x_2, and x_3 to have very small, if any, contribution. Figure 5 shows box-whisker distributions of the contributions of each input for the small and large MLP (top and bottom row).

In general all three methods confirm our expectations. They attribute x_3 with a contribution of almost zero, and x_2 is identified as the most influential input. For IG, the median contribution from both x_2 and x_1 is zero as well, that is, the same as for x_3, but the variability of contributions from x_2 is the largest, indicating a potential large influence on the magnitude of the predicted value. For LRP-$\alpha\beta$ and LRP-p, the contribution from x_2 is mostly positive, which means that x_2 increases the magnitude of y_{NN}, whereas the contribution for the weaker correlated input x_1 is mostly negative and also much smaller. Thus, overall x_1 contributes less than x_2 and because its contributions are mostly negative, x_1 decreases the magnitude of y_{NN}. x_1 shows the same range in its contributions as x_2, because in the given example the range of contributions for x_1 and x_2 are related to each other, as the percentages for all inputs add up to about 100% for LRP-p or the output value for LRP-$\alpha\beta$. For LRP-p, the contributions from x_2 are centered near 125%. The exact percentage of LRP-p is not related to the correlation coefficient of the input to the target. However, for our 3-input-model, percentages in the order of 100% are plausible, because there is only a single input with a strong relation to the output, which we expect to be captured by the NN and revealed by LRP-p. In general, a high percentage acts as a "pointer" to inputs which are likely to have a strong relationship to the target, be it linear or nonlinear. For LRP-$\alpha\beta$, a high predicted value will automatically yield higher contributions, but the contributions also change with the size of the NN. For the larger NN (Figure 5e) the magnitudes of both contributions from x_2 and x_1 increase compared to the smaller NN (Figure 5b). However, the contributions stay constant relative to each other and thus also very similar to LRP-p.

From the simple example, we see that an influential input to a neural net is characterized by a high mean/median contribution and/or by a large range in its contributions. By using these attribution methods for studying the relationship of convection to the larger scales, the respective influence from the large-scale inputs can be compared, both for predictions of TCA and ROME. We have a total of 5,089 predictions and thus obtain 5,089 different contributions for each input. We include the data used for training in the 5,089 contributions, because we are using the NN to infer strongly contributing variables and not as a purely predictive tool. Those contributions are presented in the next section.

4. Results

4.1. Contributions to Total Convective Area (TCA)

Several recent studies have shown that there is a relationship between TCA and the large-scale state of the atmosphere, in particular with vertical motion and humidity (Davies et al., 2013; Louf et al., 2019; Peters et al., 2013). In this section, we apply the MLP and attribution methods described above to explore if the XAI approach is able to reproduce these well-known relationships.

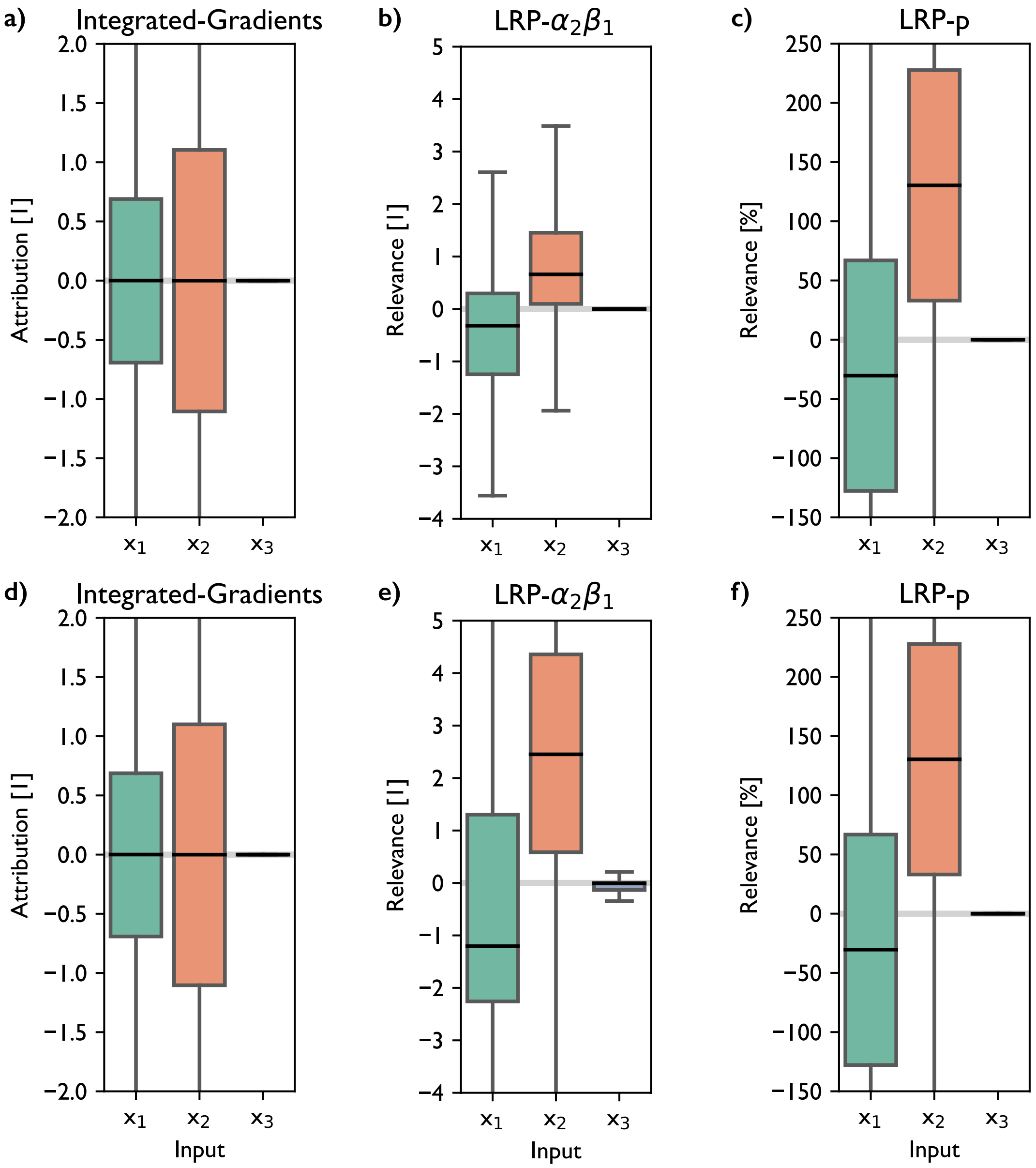

Figure 5. Median and interquartile range of the 125,000 attributions to the input of a neural net, for three different attribution methods. These methods are Integrated Gradients, Layer-wise Relevance Propagation (LRP)-$\alpha\beta$ with $\alpha = 2$ and $\beta = 1$ and LRP-p. (a–c) show the attributions for the small neural network (NN) and (d–f) show the attributions for the large NN. Each NN is trained with the same inputs and target, where by construction x_1 is slightly correlated to the target, x_2 strongly correlated, and x_3 uncorrelated.

To assess which input variables are the most influential, we have to rank them. In the following, we base this ranking on the results of the attribution method LRP-p, because the percentages can be compared to the input percentages for ROME in the following sections. But we later show that the ranking according to other attribution methods (IG with all-zero input as the baseline and LRP-$\alpha_2\beta_1$, i.e., with $\alpha = 2$ and $\beta = 1$) give similar rankings for the most contributing variables (Table 2). We increase the robustness of our results in two ways. First, instead of using a single NN, we train an ensemble of NNs with an identical setup of the nets but different initial random values. We report the first six most contributing input variables, as they are the same across the ensemble. Second, we initially focus LRP-p on the highest decile of TCA and those TCA_{NN} that have less than a 30% error. We refer to this subset of predictions as the "high-and-well-subset". We then rank the input variables according to the most positive contribution inside the high-and-well-subset, as determined by the median. We later relax

Table 2
The 6 Most Contributing Input Variables to the Neural Net and Multiple Linear Regression When Predicting High 6hr-Average TCA, According to the Three Different Attribution Methods Described in the Text

IG	LRP-$\alpha_2\beta_1$	LRP-p	MLR
ω 1. EOF	ω 1. EOF	ω 1. EOF	ω 1. EOF
PW	SH	PW	PW
OLR	OLR	OLR	OLR
drdt 2. EOF	PW	ω 3. EOF	ω 3. EOF
ω 3. EOF	ω 3. EOF	drdt 2. EOF	s 1. EOF
SH	s 5. EOF	SH	drdt 2. EOF

Note. For each attribution method, the ranking of the most contributing variables was based on the median contribution across the high-and-well subset. EOF, empirical orthogonal function; PW, precipitable water; SH, sensible heat flux; OLR, outgoing longwave radiation.

the "high-and-well" assumption to compare the results of this subset to those from the whole data set. Figure 6 shows the distribution of contributions inside the high-and-well-subset (in green) alongside the distribution of contributions for the total 5,089 predictions (in yellow) for the six most influential variables, when applying LRP-p.

The most contributing large-scale variable to predictions of TCA in the high-and-well subset is the PC time series of the first EOF-pattern of vertical velocity, ω_1 (see Figure 7c). The median of the high-and-well subset for ω_1 is at about 50%, which is a much stronger contribution than any of the other influential variables, which all have medians below 10%. Those variables are precipitable water (PW), outgoing longwave radiation (OLR), the tendency of water vapor mixing ratio (drdt) and sensible heat flux (SH). It is worth noting that our results simply represent strong relationships and cannot easily be interpreted to be causal. For example, the relationship between convection and large-scale vertical motion is known to be strongly coupled and is not the simple result of one variable "driving" the other. Additionally, it is well-known that clouds associated with convection strongly influence OLR, especially when the convection is wide-spread (as it is by definition in the top TCA decile). We nevertheless retain OLR in the data set as we wish to investigate its relationship to convective organisation in the following section. We note that removing OLR from the input data does not affect the outcome for the other five variables to be considered most influential (not shown).

When moving from the top TCA decile to the full set of predictions, ω_1 remains the variable with the largest range of contributions. Recall from the simple example presented in Section 3.2 that a large range in the influence identifies a variable which has a large variability in its contributions and thereby is influential for the predicted value. Interestingly, the median of the contributions for the full data set shifts to negative values, when it was positive for the high-and-well data set. This can be understood by the fact that, in our data set, the vast majority of times experience subsiding motion associated with little or no convection. This results in a negative sign in the time series of the first EOF of ω, resulting in negative contributions for most of the observation times.

To further investigate the utility of the attribution methods in identifying physically meaningful relationships, we examine how the six most influential variables relate to TCA by binning them into deciles of TCA (Figure 7). For ω, strong ascending motion throughout the atmosphere is present in the highest decile of TCA, with the highest velocities at mid-levels, whereas for the low and middle decile of TCA, vertical motion is weak and descending at lower levels (Figure 7a). This vertical structure is also seen in the first and third EOF of ω (Figure 7c). We note that it is the PC time series associated with these EOF-patterns that serves as input to the NN, and it is these

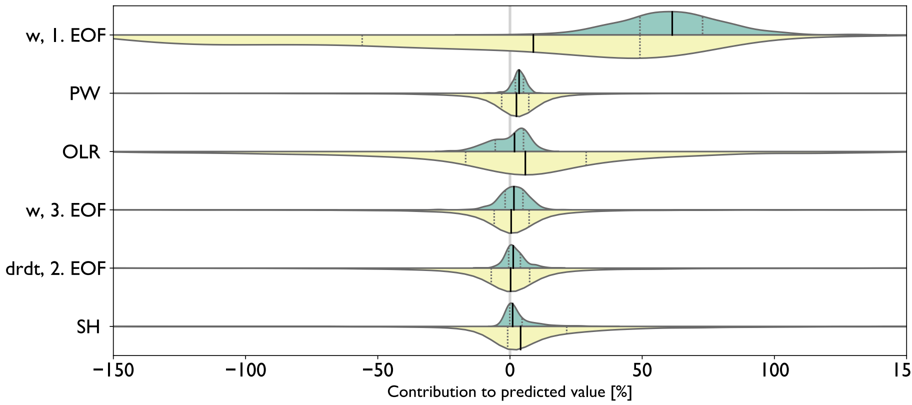

Figure 6. Contribution distributions of large-scale input variables (on y-axis) to predicted total convective area (TCA). The green distribution samples the contributions to TCA when it is larger than its 90th percentile and the predictions have an error less than 30% (sample size 431). The yellow distribution samples all predictions (sample size 5,089). Input variables are ranked by the median in the green distribution.

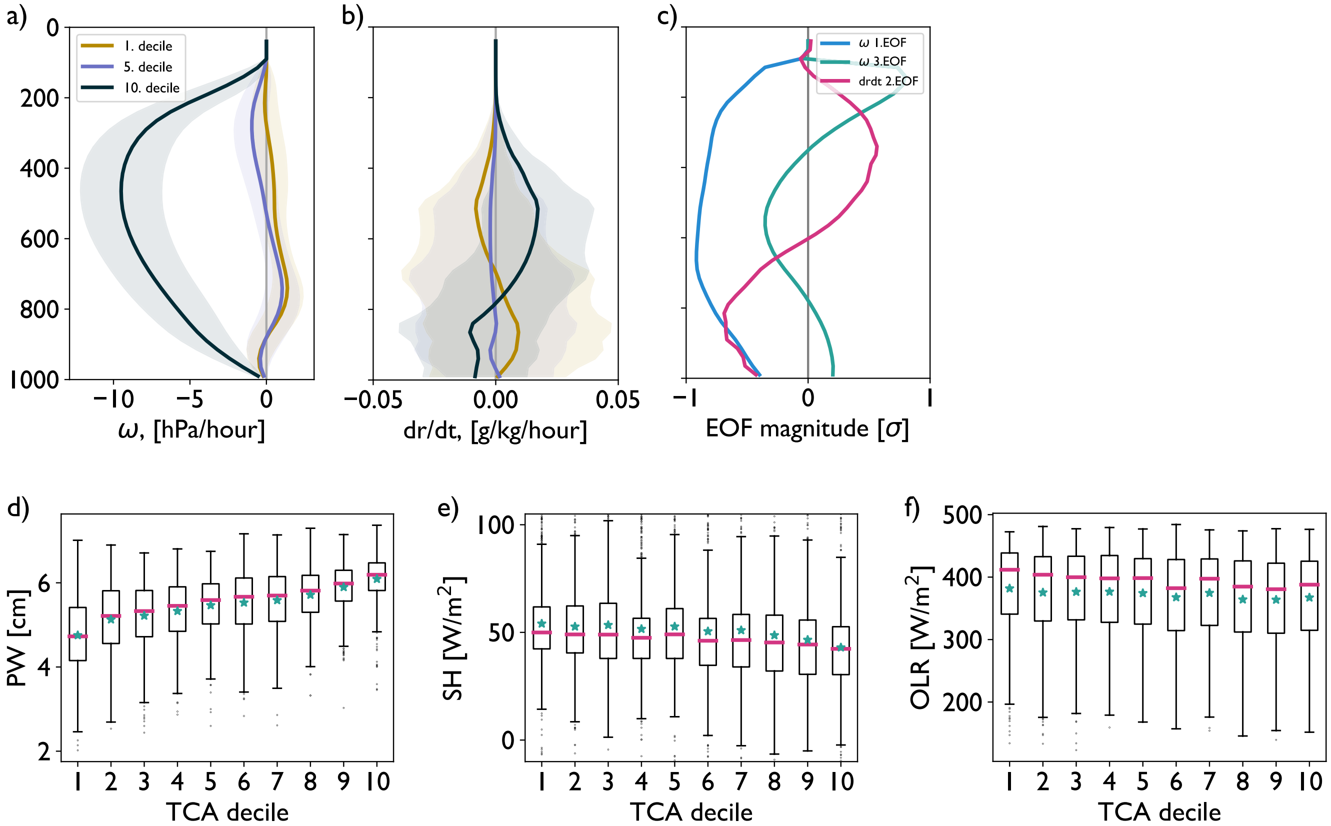

Figure 7. Profiles of (a) ω and (b) drdt from the large-scale data set for first, fifth, and tenth decile of total convective area (TCA). The shading shows the standard deviation inside each decile. (c) Empirical orthogonal function (EOF)-profiles for which the corresponding principal component time series are influential to predict TCA. The EOFs are scaled to units of standard deviation σ of the large-scale quantities. (d) Precipitable water (PW), (e) sensible heat flux, and (f) outgoing longwave radiation across the deciles of TCA. The pink bar is the median, the star is the mean and the sample size is 638 in each decile.

patterns that are most influential to the NN predictions. The first EOF of ω is negative throughout the troposphere and, depending on the sign of the PC series, therefore represents deep ascent or subsidence. The third EOF of ω, which has a much smaller influence on the predictions (Figure 6), shows a tri-pole pattern. Its small positive contribution to the high-and-well predictions indicates upward motion at mid-levels when TCA is high.

Similarly to the ω-pattern, the second EOF-pattern of drdt (Figure 7c) which is among the six most influential inputs according to the LRP-p attribution method, aligns well with the profile of drdt in the highest TCA-decile, for which a moistening tendency is present at upper levels and a drying tendency at lower levels (Figure 7b). For the lowest TCA-decile, the signs in the relation between upper and lower levels becomes reversed. In the interpretation of the drdt profile, it is worth remembering that this quantity represents the combined action of large-scale and convective-scale processes and constitutes a residual between them, rather than the action of either of these scales alone. Precipitable water shows a strong difference between small and large TCA, with a higher atmospheric moisture content associated with larger convective areas (Figure 7d). This confirms a well-established relationship between humidity and convective activity (Bretherton et al., 2004). The relations of TCA to sensible heat flux and OLR are much weaker than those for PW, with a high overlap of the distributions between the different TCA-deciles. However, both the sensible heat flux and OLR decrease for larger TCA. When applying a Student's *t*-test, the average value in the highest TCA-decile is lower than in the first decile on a statistically significant level for both these variables ($p \leq 0.0002$).

Table 2 shows the top six variables, ranked by their median contribution in the high-and-well subset, when applying IG and LRP-$\alpha_2\beta_1$ instead of LRP-p. We additionally list the most contributing variables for MLR in the high-and-well subset, when multiplying the input values with their regression coefficients. All methods agree on five of the six most contributing variables, only the second EOF of drdt is not present among the top six in LRP-$\alpha_2\beta_1$ and SH is not present in MLR. However, all methods attribute the first ω-EOF with having the strongest

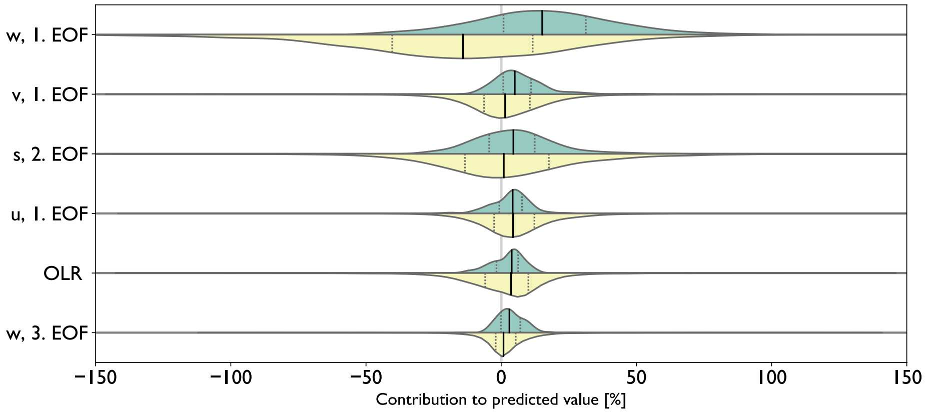

Figure 8. Contribution distributions of large-scale input variables (on *y*-axis) to predicted Radar Organisation Metric (ROME). The green distribution samples the contributions to ROME when ROME is larger than its 90th percentile and the predictions have an error less than 30% (sample size 218). The yellow distribution samples all predictions (sample size 5,089). Input variables are ranked by the median in the green distribution.

contributions by far (not shown). The overall agreement on the set of influential variables when using the different attribution methods increases our confidence in the results.

Two quantities that are often considered to be closely related to convective activity are notable by their absence from the most influential variables for the prediction of TCA, even though they are included in the input data set. Those are convective available potential energy (CAPE) and convective inhibition (CIN), which are ranked 33 and 68, out of 68 total predictors, respectively. This result confirms the findings of several recent studies utilizing a data set very similar to that used here. Peters et al. (2013) show that there is only a weak relation of CAPE to deep convective area fraction in our study area and Louf et al. (2019) report of a “poor overall relationship of CAPE and CIN to convective area fraction”.

Applying the NN and attribution methods to predictions of TCA from large-scale state variables and fluxes has revealed well-known and physically meaningful relationships between them. Importantly, the methods have been able to identify the known key influences on and by convection reported in several other studies. This indicates that the methodology can be successfully applied to identify physical relationships, thereby opening the “black box” that NNs are often described as. Encouraged by this success, the next section will apply the same methods to attempt an identification of the much less well-known relationships between convective organization, expressed as ROME, and the large-scale state of the atmosphere.

4.2. Contributions to Radar Organisation Metric (ROME)

Having successfully applied the attribution methods to TCA, we now apply them to the convective organisation metric ROME. Unlike for TCA, the large-scale influences on organisation in our data set have not before been analyzed.

As for TCA, we mitigate against using poor predictions in our assessment of influence by once again choosing a “high-and-well” subset of the data, that is predictions of values in the highest ROME-decile with less than 30% error. We use a similar set of figures as above to illustrate the most influential inputs in the prediction of ROME. Figure 8 shows the contribution distributions of those input variables which consistently contribute the most to the high-and-well subset (in green), according to LRP-p. Once again, those are determined by training multiple neural nets with differing random initial values and the input variables are ranked by their median in the high-and-well subset. As was the case for TCA, the first EOF-pattern of ω is the most contributing input variable. However, the overall contribution of ω is much smaller, with a median contribution of about 15%, instead of about 50% for TCA. This implies that in a relative sense, the importance of ω to predict ROME is reduced compared to predictions of convective area. Furthermore, the other strongly contributing input variables to ROME are also different than those for TCA, with only OLR and ω occurring in both sets. In particular, the large influence

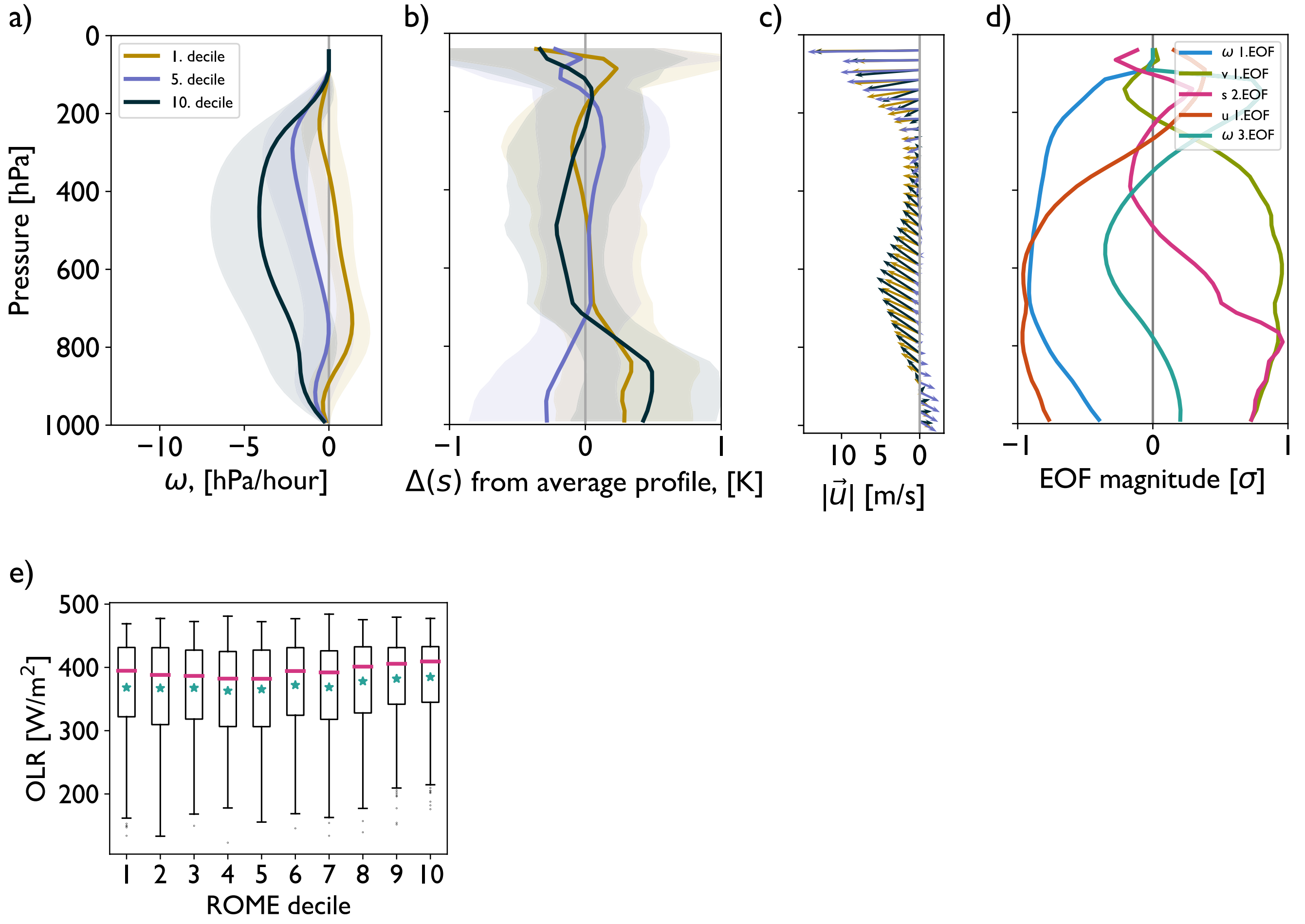

Figure 9. Profiles of (a) ω, (b) dry static energy s, and (c) horizontal wind from the large-scale data set for first, fifth and 10th decile of Radar Organisation Metric (ROME). The shading shows the standard deviation inside each decile. In panel (b), dry static energy s is presented as the difference from the all-time average for easier visualisation. In panel (c), the length of the arrows are the magnitude of the wind, with scale given on the x-axis, and the direction of the arrows gives the cardinal direction of the wind with northward being upwards. (d) Empirical orthogonal function (EOF)-profiles for which the corresponding principal component time series are influential to predict ROME. The EOFs are scaled to units of standard deviation σ from the temporal mean of the large-scale quantities. (e) Outgoing longwave radiation across the deciles of ROME. The pink bar is the median, the star is the mean and the sample size is 636 in each decile.

of moisture-related variables such as PW and drdt on TCA is replaced by circulation related variables, such as horizontal winds (u and v) in addition to the second EOF profile of dry static energy (s).

To gain more insights into the nature of the relationships identified as important by LRP-p, we investigate differences in the input variables by ROME decile (Figure 9). As for TCA, we find that the strongest ascending motion is present in the highest decile of ROME (Figure 9a). However, consistent with the lower median influence in Figure 8, the differences to the lower deciles of ROME are much smaller than they were for TCA. This confirms the suitability of LRP-p to not only identify qualitative relationships, but to quantify their strength.

Dry static energy, s shows a maximum at low levels and a minimum at midlevels when ROME is in its upper decile (Figure 9b), indicating larger than average instability in those situations. Interestingly, dry static energy is also high at low levels when the degree of convective organization is lowest, with the minimum at higher levels absent in those cases. This indicates that while warmer than average boundary layers are present in the study region when convection is organized, they do not help predict its existence.

The wind profiles show westerly low-level winds veering to southeasterly at mid-levels for both the highest and lowest ROME-decile (Figure 9c). For both the zonal and meridional wind, the important EOF-pattern for the NN is their first pattern (Figure 9d). When the input sign of the zonal wind PC time series is positive, indicating easterly winds, it results mostly in positive contributions to ROME$_{NN}$. Similarly, when the meridional winds are

Table 3
The 6 Most Contributing Input Variables to the Neural Net and Multiple Linear Regression When Predicting High 6hr-Maximum ROME, According to the Three Different Attribution Methods Described in the Text

IG	LRP-$\alpha_2\beta_1$	LRP-p	MLR
ω 1. EOF	ω 1. EOF	ω 1. EOF	ω 1. EOF
v 1. EOF	v 1. EOF	v 1. EOF	s 2. EOF
s 2. EOF	u 1. EOF	s 2. EOF	v 1. EOF
u 1. EOF	s 2. EOF	u 1. EOF	OLR
OLR	OLR	OLR	s 1. EOF
ω 3. EOF	ω 3. EOF	ω 3. EOF	PW

Note. For each attribution method, the ranking of the most contributing variables was based on the median contribution across the high-and-well subset. EOF, empirical orthogonal function; PW, precipitable water; OLR, outgoing longwave radiation.

southerly (positive) they contribute postively to $\mathrm{ROME_{NN}}$. Thus high values of ROME are related to south-easterlies, matching the midlevel wind direction present in the highest decile of ROME.

The final influential variable in the high-and-well set of ROME predictions is OLR, which was also in the influential input set for TCA. However, the relationship of OLR to ROME deciles (Figure 9e) is quite different than that found for the TCA deciles (Figure 7f). While OLR decreases as TCA increases, its relationship to ROME is not linear. Here, high OLR is found for both high and low ROME and the lowest OLR occurs for intermediate ROME. This result is consistent with earlier findings by Tobin et al. (2012), which associate more organized convective states with higher OLR due to the presence of clear-sky around the organized systems. Applying a Student's *t*-test shows that the average OLR in the fifth and tenth ROME decile is statistically different ($p \simeq 4 \cdot 10^{-7}$), while the difference in average OLR between the first and fifth ROME decile is not.

As for TCA, we compare the variables found by LRP-p to be influential for predicting ROME with the other attribution methods and MLR. Table 3 shows the top six variables, ranked by their median contribution in the high-and-well subset, when applying IG and LRP-$\alpha_2\beta_1$ and also lists the most contributing variables for MLR in the high-and-well subset, when multiplying the input values with their regression coefficients. All attributions methods applied to the NN agree on the same six most contributing variables, only LRP-$\alpha_2\beta_1$ swaps the order of u and s. For MLR however, u and the third EOF of ω are not among the six most contributing variables, instead they are replaced by the first EOF of s (additionally to its second) and PW. Because the predictions of MLR are not as good as by the NN, we do not necessarily expect the same variables.

The results presented in this section indicate that the combination of neural network predictions with attribution methods from XAI can identify key quantities related to convective characteristics in the Darwin region. The algorithms were able to reproduce well-known relationships between TCA and larger-scale variables. Interestingly, the influential variables identified for ROME, an indicator for convective organization, are quite different from those for TCA. Perhaps most strikingly, moisture-related variables are absent from the set of variables related to organization. We will investigate potential reasons and implications for this behavior in the following section.

5. Discussion

The previous section highlighted the ability of attribution methods applied to an MLP to identify large-scale variables that strongly relate to tropical convection. It also showed that the set of the most influential quantities to predict the overall amount of convection (TCA) is different to the set of quantities for predicting organization (ROME). Notably, the horizontal wind field appears to be more influential to ROME than to TCA, while atmospheric moisture content, perhaps somewhat unexpectedly, does not seem to be influential to predict ROME. In this section, we will further investigate the connection between moisture and organization as well as relate the strong influence of the atmospheric wind profiles to different regimes of the Australian monsoon.

5.1. Relative Humidity at 500 hPa

We will investigate the relationship of humidity to TCA and ROME following a similar approach to Peters et al. (2013) and Louf et al. (2019) by examining the two variables in the two-dimensional phase space spanned by vertical motion and relative humidity in the middle troposphere (515 hPa in our data set). For brevity, we refer to these two quantities as ω_{500} and RH_{500}.

As we determined the most influential variables for the subset of the data representing the 90th percentile of TCA and ROME, respectively, we once again focus our investigation on this percentile. Figure 10 shows the likelihood for ROME (left) and TCA (right) to exceed their 90th percentile value in joint bins of ω_{500} and RH_{500}. Just focusing on the outline of the phase-space in both graphs, it is evident that ω_{500} and RH_{500} are not entirely independent with strongly ascending ω associated with higher values of RH_{500}.

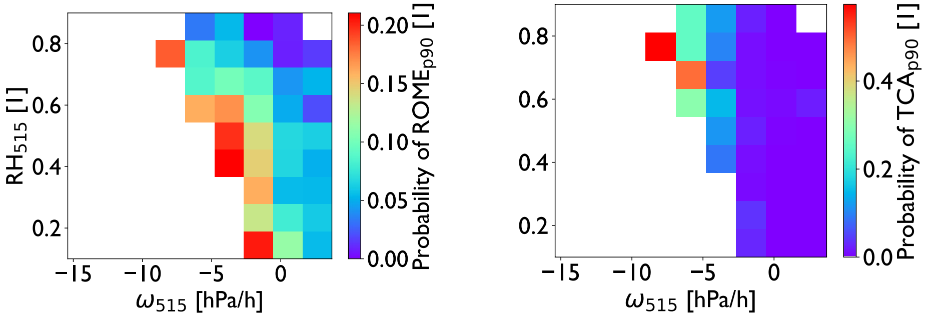

Figure 10. The likelihood for Radar Organization Metric (ROME; left) and total convective area (TCA; right) to be higher than their respective 90th percentiles (ROME_{p90} and TCA_{p90}), binned by ω_{500} and RH_{500}. Only bins containing at least 55 samples (about 1% of total sample size) are shaded.

The likelihood for TCA to be above its 90th percentile follows a well-established relationship that has the highest TCA values occurring in a strongly ascending atmosphere with moist midlevels (Peters et al., 2013). This relationship is very different for the degree of convective organization (ROME). Here, high ROME does not only occur for moist midlevels. Instead, high likelihoods for finding ROME values above its 90th percentile exist across a wide range of RH_{500}, from about 20% to 80%. Thus, RH_{500} by itself does not have a clear relationship to ROME, which likely explains the absence of a humidity related predictor amongst the highly influential set for ROME. However, a clearer relationship does emerge when holding ω_{500} fixed. In each of the bins of ω_{500}, and in particular in the presence of ascent, the likelihood for high ROME increases as RH_{500} decreases. In other words, if midlevel moisture is "as low as possible" for a given ω, high ROME is most likely. Combining this with the knowledge that TCA is strongly related to ω_{500} (the correlation in our data set is −0.82), we can conclude that for a given overall strength of convection the likelihood of the convection to be in a more organized state increases for a drier midlevel atmosphere. This finding is in accordance with Tobin et al. (2012), who show decreased free-tropospheric RH for higher degrees of organization in events with similar convective rainfall, and is also consistent with for example, Holloway et al. (2017) and Wing et al. (2017).

5.2. Wind Regimes

The horizontal wind components were identified by our algorithm to be influential in predicting ROME, but not TCA. Given the relationship of ROME to ω_{500} and RH_{500} discussed above, it is worth considering whether this is a result of organization depending on the different regimes of the Australian monsoon, which have distinct wind and humidity characteristics. To investigate this further, we make use of an objective identification of monsoon regimes by Pope et al. (2009). They used radiosonde observations to objectively classify the state of the environment at Darwin on a daily basis during the wet-season (October to April). They identified five distinct regimes, each associated with specific flow characteristics. They are the *dry east* (DE) regime corresponding to the trade-wind environment; the *east* (E) regime, a transition environment between the trade winds and monsoon season; the *moist east* (ME) regime that characterizes break monsoon conditions; the *deep west* (DW) regime, found during active monsoon phases; and the *shallow west* (SW) regime, also found in monsoon breaks.

Compositing our data set using the five regimes allows us to compare the influential wind profiles for the NN to the regimes' wind profiles and, in doing so also relate ROME and TCA to the wet-season regimes. We do so by dividing each decile of TCA and ROME into the five regimes (Figure 11). Note that our method implies that the five values for each decile add to 100%, allowing the overall occurrence of the five regimes to be reflected in the resulting graphs. The ME regime is the most frequent regime in the data set with an occurrence of 40% of the time. Hence it is also the most populated regime across the deciles. The SW and DW regimes are the second most populated (25% and 19% respectively).

Significant differences in the decile distributions with and within regimes emerge for TCA and ROME, with the largest differences appearing in the active monsoon regime (DW) and the break regime (SW). In the DW regime, there is a clear increase in frequency of the TCA deciles with decile number. In other words, TCA is frequently large during active monsoon conditions, consistent with the large area-average convective rainfall rates in that regime. In contrast, the ROME decile frequency distribution in the DW regime shows a peak in the near-median

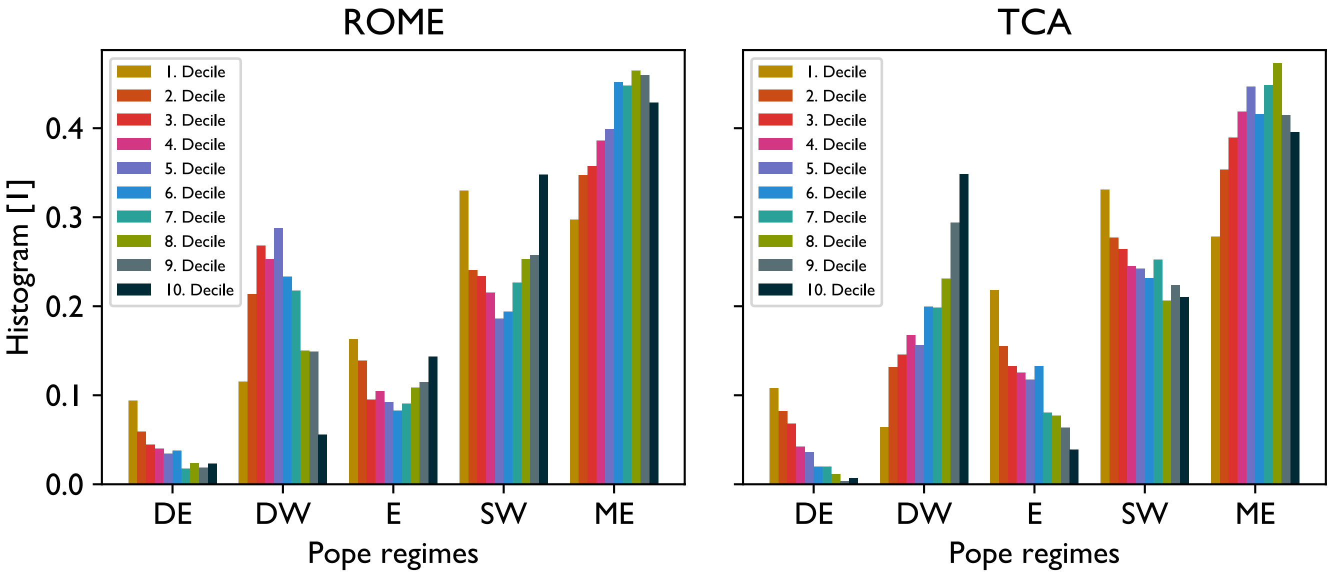

Figure 11. Distribution of each Radar Organisation Metric (ROME)-decile (left) and total convective area (TCA)-decile (right) across Pope-Regimes. Each color, that is, decile, adds up to 1 across the regimes.

deciles, indicative of moderate values of ROME occurring in this regime. As the DW regime is the most humid of the regimes (Pope et al., 2009), this result is consistent with the humidity analysis in the previous subsection.

The TCA decile distribution in the SW regime shows a decrease from smallest to largest decile, indicating the overall convectively suppressed nature of this regime. In contrast, both the lowest and highest decile of ROME are equally common in the SW regimes, with a minimum of ROME frequency near its median. This indicates that when convection occurs in this regime, which is characterized by low midtropospheric humidity (Pope et al., 2009), it has a relatively high probability of being strongly organized. A similar behavior is found in the E regime, which characterizes the transition from relatively dry trade-wind conditions to the moister ME and DW regimes.

The difference in the behavior of TCA and ROME in this wet-season regime analysis provides a possible explanation why wind-related variables were identified as influential for ROME, but not for TCA. The highest decile of ROME increases in its occurrence in regimes dominated by easterly winds at midlevels, that is, the ME and SW regimes, while the highest decile of TCA is found in equal proportion between the ME and DW regimes, which have opposing midlevel winds. In short, the wind direction has only a limited influence on the behavior of TCA, whereas it strongly affects ROME in our study region. The strong influence of the wind-profiles on ROME-predictions suggests that the local conditions may matter to convective organization. The convection in our study region is influenced by its coastal location as well as the varying conditions found during the North Australian monsoon. This makes it typical for coastal locations in the Maritime Continent, but very likely atypical for tropical regions over the open ocean or inland of major continents, which have been shown to show different relationships between humidity and rainfall (Bergemann & Jakob, 2016).

6. Summary

The purpose of this study was to investigate whether an XAI approach can be used to identify the key large-scale atmospheric parameters that determine the overall strength of convection and its state of organization in a region equivalent to the size of a GCM grid-box. To achieve this, we applied attribution methods to a neural net (NN), namely Integrated Gradients (IG) and Layer-wise Relevance Propagation with the $\alpha\beta$-rule (LRP-$\alpha\beta$) and a newly proposed rule which we call percentage-backtracking (LRP-p).

We related the large-scale atmospheric state around Darwin, Australia, to two measures of deep convective activity derived from radar observations, the TCA and the degree of convective organization expressed by the ROME index. In addition to the neural net, we used multiple linear regression (MLR) to connect the large-scale quantities with the convective measures. While MLR is sufficient to predict TCA and infer the influential variables, predictions of ROME are more accurate when applying the NN, indicating a potentially larger role for nonlinear interactions with the larger scales to matter for convective organization. By applying attribution methods to the NN, and LRP-p in particular, we determined those input variables which are consistently the most influential to its predictions.

The single most influential variable to predictions of both ROME and TCA is vertical velocity ω, which is not surprising, given its known strong two-way interaction with convection in general. However, we identified that its link to TCA was significantly stronger than that to ROME. Identifying a much reduced relative importance of ω to convective organization than to overall convective activity is an important finding of this study and suggests a more complex influence of the large-scale onto convective organization than onto TCA and, by association, area-mean convective rain rate.

In addition to the reduced influence of ω for organization, the attributions identified a different set of influential input quantities between TCA and ROME. In particular, the TCA-set included two humidity-related quantities, confirming the well-known relationship between humidity and convection, while the ROME-set did not include humidity as an influential parameter and instead highlighted horizontal-wind variables as playing an important role. Further investigation revealed that the weak relation between ROME and midlevel relative humidity is a result of an intricate interplay of vertical motion and humidity. While TCA is expected to be highest in a strongly ascending atmosphere with high midlevel relative humidity, no such relationship emerges for organization. Instead, we discovered that a higher degree of convective organization is more likely for drier midlevels within bins of vertical motion, but not across them, which explains the absence of humidity as a strong overall predictor for convective organization in the neural net.

Analyzing TCA and ROME composited into regimes representing different states of the North-Australian wet season revealed that the more active and humid regimes of the monsoon (DW and ME) are associated with large TCA, while high ROME can often be found in the suppressed and drier regimes (SW, E). As TCA can be high for both westerly and easterly flow, while high ROME is predominantly associated with easterly flow, wind-related variables are identified as influential to ROME by the neural net. Furthermore, in the more suppressed wet-season regimes, ROME can be both high and low, rendering the conditions occurring in those regimes more powerful in its prediction.

It is important to note that Darwin is a coastal location in a strongly monsoonal climate. Hence, while we can consider our results as somewhat typical for the Maritime Continent regions of strong coastal influence, they are not easily generalized to other regions of the tropics. Another important caveat to our results is that the NN predictions can identify relationships within the data, but they cannot identify causality. This is particularly important to note with regard to predictors such as ω and OLR, which are known to be influenced by convection, as well as potentially influencing it. Nevertheless, our results are useful in pointing toward likely important influences on convective organization for further detailed and process-based studies.

We set out with the aim to demonstrate the usefulness of (a) employing a neural net to capture possibly nonlinear relations in atmospheric phenomena and (b) employing algorithms to understand the inner workings of a neural net with the goal to provide pointers at physical relationships to be further explored. While for TCA a linear approach works well, for ROME we consider our results a success on both accounts. We have shown that machine learning algorithms are not only able to capture relationships in a "black box" prediction approach but can also be used to identify potentially key insights into the nature of these relationships. The development and application of the attribution methods in particular has delivered important insights into the relevant relations between convection and the large-scale atmosphere and has the potential to do so for other applications in atmospheric science.

Appendix A

A1. Predictions With a Multilayer Perceptron

The structure of an MLP is based on an input layer, a series of hidden layers, and an output layer. The hidden layers are comprised of a set of nodes whose values depend on those of the previous layer and act as inputs to the next layer. Consider the example NN in Figure A1, with a single hidden layer. To calculate the value v at one node in an MLP, given a set of inputs (i_1, i_2, and i_3), the first part of the calculation is to take the scalar product between

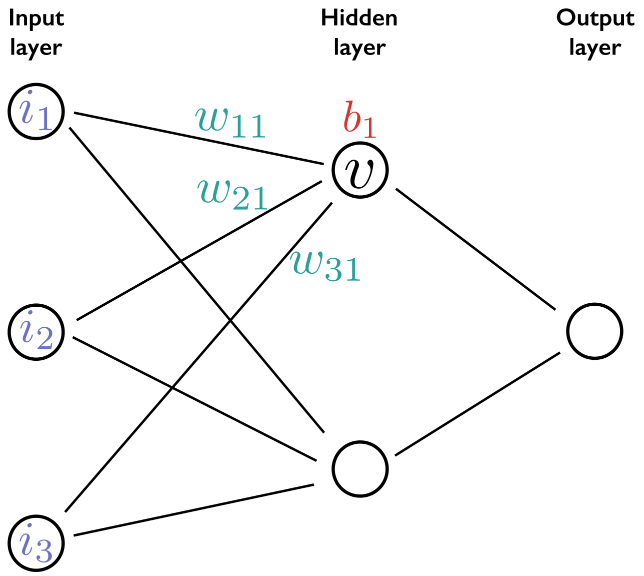

Figure A1. Example of a small multilayer perceptron, with three inputs i_1, i_2, and i_3, one hidden layer and one output node.

the input and a set of weights (w_1, w_2, w_3) and add a bias b_1. Each node has its own unique set of weights and bias; they are determined by an iterative process during the training of the NN described below. The final value of v is then evaluated by applying a nonlinear function f known as the "activation function". This final step is the only nonlinear aspect in an MLP. Symbolically, the value of v in Figure A1 is given by

$$v = f((i_1w_1 + i_2w_2 + i_3w_3) + b_1). \tag{A1}$$

The activation function f can be in principle any nonlinear function. In practice, there are commonly only a few functions that are used. We use the Rectified Linear Units function (ReLU), which is one of the common activation functions, given by

$$f : R \to R, \quad f(x) = \begin{cases} 0, & \text{for } x \leq 0 \\ x, & \text{for } x > 0 \end{cases} \tag{A2}$$

The weights and biases of a neural net are adjusted iteratively. Starting with an arbitrary set of weights and one set of inputs, the predicted value is compared with an error metric to the true output value, the "target", which must be known in advance. We use the squared difference between prediction and target as the error metric. Next, the respective error for each weight is computed and the weights are adjusted slightly, according to the "learning rate". The adjusting, that is, training, of the weights and biases is repeated for each prediction of the "training data set", which is input data for which a target is known. The "validation data set" is also data for which there is a known target to each input, but the validation data is not used for training. Rather, the validation data set may be used to estimate the error of the neural net when applied to unknown data and find the set of "hyper-parameters", which results in the smallest error. The hyper-parameters of an NN include its size, the training rate, the activation function and the number of times the training data is iterated over, called "epochs". Once the hyper-parameters are set, the "test data set", for which there is also a known target, can be used to estimate the error of the final NN on data which has neither been used for training nor validation.

Table A1
Variables of the Large-Scale Data Set Not Used as Input to the Neural Net to Predict 6h-Maximum of ROME or 6h-Average of TCA

Profile variables		Scalars	
Variable	Abbreviation	Variable	Abbreviation
Horizontal wind divergence	div	Average surface pressure	p_srf_aver
Vertical s advection	conv(s)	Center surface pressure	p_srf_center
Water vapor mixing ratio	r	2m air temperature	T_srf
Vertical r advection	conv(r)	2m relative humidity	RH_srf
Temperature	T	10m wind speed	wspd_srf
Advection of T	adv(T)	10m zonal wind	u_srf
Vertical T advection	conv(T)	10m meridional wind	v_srf
Tendency of T	dTdt	Surface net radiation	rad_net_srf
Water vapor mixing ratio	r	Longwave top of atmosphere net flux	lw_net_toa
		Shortwave top of atmosphere net flux	sw_net_toa
		Low cloud fraction from satellite	cld_low
		Middle cloud fraction from satellite	cld_mid
		High cloud fraction from satellite	cld_high
		Total cloud fraction from satellite	cld_tot
		Cloud liquid water path from MW radiometer	LWP
		Local tendency of column water	dh2odt_col
		Column water advection	h2o_adv_col
		Surface evaporation	evap_srf
		Local tendency of column s	dsdt_col
		Column s advection	s_adv_col
		Column radiative heating	rad_heat_col
		Column latent heating	LH_col
		2m r	r_srf
		2m s	s_srf
		500 hPa downward CAPE	D_CAPE
		Surface upwards longwave	lw_up_srf
		Surface downwards longwave	lw_dn_srf
		Surface upwards shortwave	sw_up_srf
		Surface downwards shortwave	sw_dn_srf

A2. The Percentage-Backtracking-Rule and $\alpha\beta$-Rule of LRP

Here, we discuss the differences between LRP-p and LRP-$\alpha\beta$. Both methods are aimed at attributing a given input's contribution to the output of a neural net. But as seen in Equations 6 and 7, their form is different. LRP-$\alpha\beta$ avoids a division by zero by separating positive from negative terms and thereby avoiding an offset to zero between them. LRP-p reduces the likelihood of a division by zero by adding a bias in the denominator. Because during the calculations of LPR-p, a division by zero or very small numbers might occur, one single prediction might yield unreasonably large or undefined values. This is a shortcoming of percentage-backtracking, yet, when tracing back multiple predictions and evaluating the distributions of contributions as is done in this work, such outliers should not change the overall result.

As mentioned above, both LRP-$\alpha\beta$ and LRP-p attribute the input nodes with values which represent their contributions to the output. Therefore, it is possible to simply divide the attributions to the inputs from LRP-$\alpha\beta$ by the output value, which results in a percentage as the attributed value as well. However, due to the different weighting with α and β for the positive/negative part of the scalar product in LRP-$\alpha\beta$, these percentages differ from the percentages obtained by LRP-p. To see this, we can consider a scalar product of three summands. Let $\{9, 5, 4\}$ be "node values" and $\{1, 1, -1\}$ the respective "weights", then the "output" is $9{\cdot}1 + 5{\cdot}1 + 4{\cdot}(-1) = 10$, if we assume the bias of the output is zero and has a ReLU-activation. The percentage for the "first node" from LRP-p can directly be calculated as 9/10, that is, 90%. If we assume $\alpha = 2$, then the attribution to the first node from LRP-$\alpha_2\beta_1$ will be $\frac{9}{9+5} \cdot 2 \cdot 10 = 12.9$, which would be 129% of 10 as the output value. Depending on the choice for α and β this percentage will be different.

While, the main principle of both LRP-$\alpha\beta$ and LRP-p is the same, the differences between them can be summarized as follows:

1. LRP-$\alpha\beta$ traces back the value of the output node. Beginning at the output node, the value of this (output) node is taken as the "relevance" and then is distributed backwards through the neural net. LRP-p only considers the relative contribution to pre-activation values, that is, percentages of node values before their activation function is applied.
2. To avoid the numerical instability which can arise through a division by zero, LRP-$\alpha\beta$ splits the scalar product in the denominator into two parts and introduces a positive and negative part. Because the positive and negative part cannot offset each other when treated separately, a division by zero is avoided. LRP-p adds the bias to the scalar product in the denominator and thereby greatly reduces the chance of a division by zero.
3. LRP-$\alpha\beta$ introduces two integer parameters, α and β, which are multiplied with the positive and negative part, respectively. It is required that $\alpha - \beta = 1$, to ensure that the amount of relevance, that is, the output node value, is conserved when redistributing its value to the nodes in the previous layer. However, if either the positive or negative part does not exist, because all summands of the scalar product have the same sign, then there is no conservation of relevance. For LRP-p, a node can obtain percentages above ±100%, that is, $|p| > 1$, and this implies that other nodes in the same layer, or the biases of the next layer, offset this node's contribution, such that the total percentage of all nodes and the biases adds up to 100%. In general, the bias terms $\frac{b_k^{l+1}}{\sum_h v_h^l w_{hk}^l}$ (compare Equation 5) are comparatively small, and the sum of the percentage contributions across all nodes in a layer is close to 100%.

Another difference between LRP-p and LRP-$\alpha\beta$ is that the obtained values by LRP-$\alpha\beta$ do not represent percentages but absolute values and therefore carry the units of the prediction. Thus, when for example, predicting an area, the relevance which is back-propagated and assigned to each node also has the units of area. This can make the interpretation of the LRP-$\alpha\beta$ easier. However, a prediction of a high value will automatically result in higher relevance assigned to the input nodes, because the relevance usually is conserved. Thus, a high relevance of an input node does not necessarily mean that it contributed much to the prediction. This can make the interpretation of LRP-$\alpha\beta$ less intuitive.

It is worth noting, that the calculations of LRP are linear, since the denominator consists only of a scalar product (plus a node-specific bias in the case of LRP-p), without applying a nonlinear activation f. Yet, the node-values v which go into the scalar product are values from nodes with f being applied. Thus, as an example, if f is the ReLU activation function, a negative node value becomes 0. With $v = 0$, this node also contributes 0 to the following nodes. Hence, LRP is able to trace back a nonlinear mapping of the input to the output of an MLP.

Data Availability Statement

Primary data and scripts used in the analysis and other supporting material that may be useful in reproducing the authors work are archived at monash.figshare.com (https://bridges.monash.edu/articles/dataset/Retsch_etal_2021/14736132).

Acknowledgments
The authors are grateful to the three anonymous reviewers for their detailed and helpful suggestions. The authors acknowledge support from the Australian Research Council through grants CE170100023 (MR, CJ, MS) and DE190100866 (MS).

References

Ahmed, F., & Neelin, J. D. (2018). Reverse engineering the tropical precipitation–buoyancy relationship. *Journal of the Atmospheric Sciences*, *75*, 1587–1608. https://doi.org/10.1175/JAS-D-17-0333.1

Arakawa, A. (2004). The cumulus parameterization problem: Past, present, and future. *Journal of Climate*, *17*(13), 2493–2525. https://doi.org/10.1175/1520-0442(2004)017<2493:ratcpp>2.0.co;2

Arakawa, A., & Schubert, W. H. (1974). Interaction of a cumulus cloud ensemble with the large-scale environment, part I. *Journal of the Atmospheric Sciences*, *31*(3), 674–701. https://doi.org/10.1175/1520-0469(1974)031<0674:ioacce>2.0.co;2

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Becker, T., & Wing, A. A. (2020). Understanding the extreme spread in climate sensitivity within the radiative-convective equilibrium model intercomparison project. *Journal of Advances in Modeling Earth Systems*, *12*(10), e2020MS002165. https://doi.org/10.1029/2020ms002165

Bergemann, M., & Jakob, C. (2016). How important is tropospheric humidity for coastal rainfall in the tropics? *Geophysical Research Letters*, *43*(11), 5860–5868. https://doi.org/10.1002/2016gl069255

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M., & Gentine, P. (2021). Machine learning for clouds and climate (invited chapter for the AGU geophysical monograph series "clouds and climate"). *Earth and Space Science Open Archive*, *27*. https://doi.org/10.1002/essoar.10506925.1

Bony, S., Stevens, B., Coppin, D., Becker, T., Reed, K. A., Voigt, A., & Medeiros, B. (2016). Thermodynamic control of anvil cloud amount. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(32), 8927–8932. https://doi.org/10.1073/pnas.1601472113

Bretherton, C. S., Blossey, P. N., & Khairoutdinov, M. (2005). An energy-balance analysis of deep convective self-aggregation above uniform SST. *Journal of the Atmospheric Sciences*, *62*(12), 4273–4292. https://doi.org/10.1175/jas3614

Bretherton, C. S., Peters, M. E., & Back, L. E. (2004). Relationships between water vapor path and precipitation over the tropical oceans. *Journal of Climate Change*, *172*(7), 1517–1528. https://doi.org/10.1175/1520-0442(2004)017<1517:rbwvpa>2.0.co;2

Byers, H., & Union, T. A. G. (1948). The use of radar in determining the amount of rain falling over a small area. *EoS Transactions*, (29), 187–196. https://doi.org/10.1029/tr029i002p00187

Davies, L., Jakob, C., May, P., Kumar, V. V., & Xie, S. (2013). Relationships between the large-scale atmosphere and the small-scale convective state for Darwin, Australia. *Journal of Geophysical Research: Atmospheres*, *118*(20), 11534–11611. https://doi.org/10.1002/jgrd.50645

Dobrescu, A., Giuffrida, M. V., & Tsaftaris, S. A. (2019). Understanding deep neural networks for regression in leaf counting. In *2019 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)* (pp. 2600–2608). https://doi.org/10.1109/CVPRW.2019.00316

Doneaud, A., Ionescu-Niscov, S., Priegnitz, D. L., & Smith, P. L. (1984). The area-time integral as an indicator for convective rain volumes. *Journal of Climate and Applied Meteorology*, *23*(4), 555–561. https://doi.org/10.1175/1520-0450(1984)023<0555:tatiaa>2.0.co;2

Emanuel, K. A., Neelin, J. D., & Bretherton, C. S. (1994). On large-scale circulations in convecting atmospheres. *Quarterly Journal of the Royal Meteorological Society*, *120*, 1111–1143. https://doi.org/10.1002/qj.49712051902

Fovell, R. G., & Tan, P. H. (1998). The temporal behavior of numerically simulated multicell-type storms. Part II: The convective cell life cycle and cell regeneration. *Monthly Weather Review*, *126*(3), 551–577. https://doi.org/10.1175/1520-0493(1998)126<0551:ttbons>2.0.co;2

Hitchcock, S. M., Lane, T. P., Warren, R. A., & Soderholm, J. S. (2021). Linear rainfall features and their association with rainfall Extremes near Melbourne, Australia. *Monthly Weather Review*, *149*(10), 3401–3417. https://doi.org/10.1175/mwr-d-21-0007.1

Holloway, C. E., & Neelin, J. D. (2009). Moisture vertical structure, column water vapor, and tropical deep convection. *Journal of the Atmospheric Sciences*, *66*(6), 1665–1683. https://doi.org/10.1175/2008jas2806.1

Holloway, C. E., Wing, A. A., Bony, S., Muller, C., Masunaga, H., L'Ecuyer, T. S., et al. (2017). Observing convective aggregation. *Surveys in Geophysics*, *38*(6), 1199–1236. https://doi.org/10.1007/s10712-017-9419-1

Holloway, C. E., & Woolnough, S. J. (2016). The sensitivity of convective aggregation to diabatic processes in idealized radiative-convective equilibrium simulations. *Journal of Advances in Modeling Earth Systems*, *8*(1), 166–195. https://doi.org/10.1002/2015ms000511

Houze, R. A. (1977). Structure and dynamics of a tropical squall-line system. *Monthly Weather Review*, *105*(12), 1540–1567. https://doi.org/10.1175/1520-0493(1977)105<1540:sadoat>2.0.co;2

Houze, R. A. (2014). Chapter 6—Nimbostratus and the separation of convective and stratiform precipitation. In R. A. Houze (Ed.), *Cloud dynamics* (Vol. 104, pp. 141–163). Academic Press. https://doi.org/10.1016/b978-0-12-374266-7.00006-8

Igel, M. R., & van den Heever, S. C. (2015). The relative influence of environmental characteristics on tropical deep convective morphology as observed by CloudSat. *Journal of Geophysical Research: Atmospheres*, *120*(9), 4304–4322. https://doi.org/10.1002/2014jd022690

Kadoya, T., & Masunaga, H. (2018). New observational metrics of convective self-aggregation: Methodology and a case study. *Journal of the Meteorological Society of Japan*, *96*(6), 535–548. https://doi.org/10.2151/jmsj.2018-054

Keenan, T., Morton, B. R., Manton, M. J., & Holland, G. J. (1989). The Island Thunderstorm Experiment (ITEX)—A study of tropical thunderstorms in the Maritime continent. *Bulletin of the American Meteorological Society*, *70*(2), 152–159. https://doi.org/10.1175/1520-0477(1989)070<0152:titeso>2.0.co;2

Leary, C. A., & Houze, R. A. (1979). The structure and evolution of convection in a tropical cloud cluster. *Journal of the Atmospheric Sciences*, *36*(3), 437–457. https://doi.org/10.1175/1520-0469(1979)036<0437:tsaeoc>2.0.co;2

Louf, V., Jakob, C., Protat, A., Bergemann, M., & Narsey, S. (2019). The relationship of cloud number and size with their large-scale environment in deep tropical convection. *Geophysical Research Letters*, *46*(15), 9203–9212. https:10.1029/2019GL083964

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2021). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *IEEE Transactions on Neural Networks and Learning Systems*.

May, P. T., Mather, J. H., Vaughan, G., Jakob, C., McFarquhar, G. M., Bower, K. N., & Mace, G. G. (2008). The tropical warm pool international cloud experiment. *Bulletin of the American Meteorological Society*, *89*(5), 629–646. https://doi.org/10.1175/bams-89-5-629

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G. E., Elmore, K. L., Homeyer, C. R., & Smith, T. (2019). Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, *100*(11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Muller, C. J., & Held, I. M. (2012). Detailed investigation of the self-aggregation of convection in cloud-resolving simulations. *Journal of the Atmospheric Sciences*, *69*(8), 2551–2565. https://doi.org/10.1175/jas-d-11-0257.1

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). *KerasTuner*. Retrieved from https://github.com/keras-team/keras-tuner

Peters, K., Jakob, C., Davies, L., Khouider, B., & Majda, A. J. (2013). Stochastic behavior of tropical convection in observations and a multicloud model. *Journal of the Atmospheric Sciences*, *70*(11), 3556–3575. https://doi.org/10.1175/jas-d-13-031.1

Pope, M., Jakob, C., & Reeder, M. J. (2009). Regimes of the North Australian wet season. *Journal of Climate*, *22*(24), 6699–6715. https://doi.org/10.1175/2009jcli3057.1

Redelsperger, J. L., & Lafore, J. P. (1988). A 3-dimensional simulation of a tropical squall line - convective organization and thermodynamic vertical transport. *Journal of the Atmospheric Sciences*, *45*(8), 1334–1356. https://doi.org/10.1175/1520-0469(1988)045<1334:atdsoa>2.0.co;2

Retsch, M. H., Jakob, C., & Singh, M. (2020). Assessing convective organization in tropical radar observations. *Journal of Geophysical Research: Atmospheres*, *125*(7), e2019JD031801. https://doi.org/10.1029/2019jd031801

Schiro, K. A., Ahmed, F., Giangrande, S. E., & Neelin, J. D. (2018). GoAmazon2014/5 campaign points to deep-inflow approach to deep convection across scales. *Proceedings of the National Academy of Sciences*, *115*(18), 4577–4582. https://doi.org/10.1073/pnas.1719842115

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). *Deep inside convolutional networks: Visualising image classification models and saliency maps*. Retrieved from https://arxiv.org/abs/1312.6034v2

Steiner, M., Houze, R. A., & Yuter, S. E. (1995). Climatological characterization of three-dimensional storm structure from operational radar and rain gauge data. *Journal of Applied Meteorology*, *34*(9), 1978–2007. https://doi.org/10.1175/1520-0450(1995)034<1978:ccotds>2.0.co;2

Sundararajan, M., Taly, A., & Yan, Q. (2017). *Axiomatic attribution for deep networks*. Retrieved from https://arxiv.org/abs/1703.01365

Tan, J., Jakob, C., Rossow, W. B., & Tselioudis, G. (2015). Increases in tropical rainfall driven by changes in frequency of organized deep convection. *Nature*, *519*(7544), 451–454. https://doi.org/10.1038/nature14339

Tobin, I., Bony, S., & Roca, R. (2012). Observational evidence for relationships between the degree of aggregation of deep convection, water vapor, surface fluxes, and radiation. *Journal of Climate*, *25*(20), 6885–6904. https://doi.org/10.1175/jcli-d-11-00258.1

Tompkins, A. M., & Semie, A. G. (2017). Organization of tropical convection in low vertical wind shears: Role of updraft entrainment [Journal Article]. *Journal of Advances in Modeling Earth Systems*, *9*(2), 1046–1068. https://doi.org/10.1002/2016ms000802

Toms, B. A., Barnes, E. A., & Ebert-Uphoff, I. (2020). Physically interpretable neural networks for the geosciences: Applications to Earth system variability. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2019MS002002. https://doi.org/10.1029/2019ms002002

Weickmann, K. M., & Khalsa, S. J. S. (1990). The shift of convection from the Indian Ocean to the western Pacific Ocean during a 30–60 day oscillation. *Monthly Weather Review*, *118*(4), 964–978. https://doi.org/10.1175/1520-0493(1990)118<0964:tsocft>2.0.co;2

White, B. A., Buchanan, A. M., Birch, C. E., Stier, P., & Pearson, K. J. (2018). Quantifying the effects of horizontal grid length and parameterized convection on the degree of convective organization using a metric of the potential for convective interaction. *Journal of the Atmospheric Sciences*, *75*(2), 425–450. https://doi.org/10.1175/jas-d-16-0307.1

Wing, A. A., Emanuel, K., Holloway, C. E., & Muller, C. (2017). Convective self-aggregation in numerical simulations: A review. *Surveys in Geophysics*, *38*(6), 1173–1197. https://doi.org/10.1007/s10712-017-9408-4

Wing, A. A., & Emanuel, K. A. (2014). Physical mechanisms controlling self-aggregation of convection in idealized numerical modeling simulations. *Journal of Advances in Modeling Earth Systems*, *6*(1), 59–74. https://doi.org/10.1002/2013ms000269

Xie, S., Cederwall, R. T., & Zhang, M. (2004). Developing long-term single-column model/cloud system–resolving model forcing data using numerical weather prediction products constrained by surface and top of the atmosphere observations. *Journal of Geophysical Research*, *109*(D1), D01104. https://doi.org/10.1029/2003jd004045

Xu, K.-M., & Emanuel, K. A. (1989). Is the tropical atmosphere conditionally unstable? *Monthly Weather Review*, *117*(7), 1471–1479. https://doi.org/10.1175/1520-0493(1989)117<1471:ittacu>2.0.co;2

Zhang, M. H., & Lin, J. L. (1997). Constrained variational analysis of sounding data based on column-integrated budgets of mass, heat, moisture, and momentum: Approach and application to ARM measurements. *Journal of the Atmospheric Sciences*, *54*(11), 1503–1524. https://doi.org/10.1175/1520-0469(1997)054<1503:cvaosd>2.0.co;2

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. In *2016 IEEE conference on Computer Vision and Pattern recognition (CVPR)*. https://doi.org/10.1109/Cvpr.2016.319

Zipser, E. J. (1977). Mesoscale and convective-scale downdrafts as distinct components of squall-line structure. *Monthly Weather Review*, *105*(12), 1568–1589. https://doi.org/10.1175/1520-0493(1977)105<1568:macdad>2.0.co;2