npj | climate and atmospheric science

Article

Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-025-01154-1

How key features of early development shape deep convective systems

Check for updates

Sophie Abramian[1] ✉, Caroline Muller[2], Camille Risi[1], Thomas Fiolleau[3] & Rémy Roca[3]

Deep Convective Systems (DCSs) reaching scales of 100–1000 km play a pivotal role as the primary precipitation source in the tropics. Those systems can have large cloud shields, and thus not only affect severe precipitation patterns but also play a crucial part in modulating the tropical radiation budget. Understanding the complex factors that control how these systems grow and how they will behave in a warming climate remain fundamental challenges. Research efforts have been directed, on one hand, towards understanding the environmental control on these systems, and on the other hand, towards exploring the internal potential of systems to develop and self-aggregate in idealized simulations. However, we still lack understanding on the relative role of the environment and internal feedbacks on DCS mature size and why. The novel high-resolution global SAM simulation from the DYAMOND project, combined with the TOOCAN Lagrangian tracking of DCSs and machine learning tools, offers an unprecedented opportunity to explore this question. We find that a system's growth rate during the first 2 h of development predicts its final size with a Pearson correlation coefficient of 0.65. Beyond this period, growth rate emerges as the strongest predictor. However, in the early stages, additional factors–such as ice water path heterogeneity, migration distance, interactions with neighboring systems, and deep shear–play a more significant role. Our study quantitatively assesses the relative influence of internal versus external factors on the mature cloud shield size. Our results show that system-intrinsic properties exert a stronger influence than environmental conditions, suggesting that the initial environment does not strictly constrain final system size, particularly for larger systems where internal dynamics dominate.

Deep convective systems (DCSs) exert a profound influence on the tropical water and energy cycle[1]. These systems refer to organized deep cloud systems that span scales larger and last longer than an individual convective cell. The largest of these systems (reaching mesoscales, and known as mesoscale convective systems; here mesoscale refers to scales of 100s km, i.e., between the scale of individual convective clouds ~1 km and the synoptic scale ~1000 km) contribute to over 50% of precipitation tropicswide[2,3]. The disproportionate impact of large long lasting systems on extreme rainfall[4] underscores the need to unravel the factors governing their development[5]. Modeling studies[6,7] have further suggested that the spatial distribution of deep convection, especially the degree of clustering of deep clouds, could also impact tropospheric humidity and cloud coverage, and thus the radiative balance of the Earth, which has been confirmed by a recent study based on observational data[8]. Understanding the complex factors that control how these systems organize and how they will behave in a warming climate remain fundamental challenges[9,10].

At the core of deep convection and DCSs formation lie three fundamental ingredients: moisture, instability, and a lifting mechanism[11,12]. Humidity supplies the water necessary for cloud formation while minimizing the detrimental impact from dry air entrainment. Instability reflects the atmosphere's potential for vertical motion. Lifting mechanisms initiate the upward motion that triggers convection. These three pillars are influenced by both internal feedbacks within DCSs and external environmental factors. However, distinguishing between internal and external drivers is often challenging due to their complex interactions and the limitations of data that capture both the large-scale and finer-scale dynamics.

Given these constraints, past research has often approached internal and external processes independently. On one hand, observational studies

[1] Laboratoire de Météorologie Dynamique, IPSL, CNRS, Ecole Normale Supérieure, Sorbonne Université, PSL Research University, Paris, France. [2] Institute of Science and Technology Austria (ISTA), Klosterneuburg, Austria. [3] Université de Toulouse, Laboratoire d'Etudes en Géophysique et Océanographie Spatiales, (CNRS/CNES/IRD/UT3), Toulouse, France. ✉e-mail: sophie.abramian@gmail.com

and global modeling efforts have largely concentrated on how large-scale environmental factors influence the behavior of DCS[13,14]. Among these efforts, a recent study[15] highlights that short-lived systems exhibit weak regional variability, while long-lived systems show strong variability, implying that external processes can't explain the diversity of systems observed and suggesting that internal dynamic might play a major role in this. However, observational data alone have not been sufficient to conclusively validate this hypothesis. On the other hand, idealized studies, such as radiative-convective equilibrium simulations, offer a simplified framework to explore internal feedback mechanisms in greater depth, particularly the self-aggregation of deep convection into larger cloud system[16,17]. Within those idealized controlled environments, four main physical processes have been identified as playing a key role in organizing deep convection, namely radiative feedbacks, turbulent entrainment at the edge of clouds, cold pools and waves[18]. But the relevance of these idealized studies to the organization of clouds in the real tropics is still debated.

Although the underlying processes remain complex and not fully understood, DCSs nonetheless exhibit a remarkably systematic life cycle. Typically, DCSs exhibit a linear phase of growth, characterized by a rapid expansion in size, followed by a linear phase of decay, where the system gradually dissipates. This simple life cycle can be effectively captured by a model with three key parameters: maximum area (A_{max}), lifespan (D), and duration of the growth phase (t_{max}) (see Materials and Methods, Fig. 6). Notably, approximately 60% of cases investigated here exhibit a nearly symmetrical life cycle, where t_{max} is equal to half of the lifespan. In our dataset (described in the next paragraph), DCDs extend in mean approximately 115 km in one direction ($\sqrt{A_{max}}$), with a standard deviation of 45 km and have an average duration of 7.5 h (D) with a standard deviation of 2.7 h. This consistent life cycle may be due to the fact that the growth rate encapsulates much of the information about the various internal processes acting on the system. The growth rate dA/dt itself can be described by a simplified mass balance equation[19]

$$\frac{dA}{dt} = A_{c,src} - \frac{1}{\rho}\frac{dM_c}{dz} - \frac{1}{\rho}\frac{dM_s}{dz} - \frac{A}{\tau} \qquad (1)$$

where A represents the cloud shield area, $A_{c,src}$ is the convective area (summed over the convective updrafts within the system), and the subscript "src" refers to the fact that convection acts as a source for the total cloud shield. Additionally, M_c and M_s correspond to the convective and stratiform mass flux, respectively. Parameters ρ and τ stand for the atmospheric density and cloud shield area decay timescale, taken in the mid-troposphere.

Recent advances, notably using Lagrangian tracking of systems, have enabled more comprehensive investigations of DCS life cycles, allowing for a more holistic study of both internal and external influences. High-resolution (4 km) global simulation using the SAM model within the DYAMOND project[20–22], combined with a sophisticated storm tracking method called TOOCAN[23], now provide over 100,000 tropical (30S to 30N, see Fig. 1A) DCSs in August and September 2016 as an extensive dataset to explore. A rigorous comparison of DCSs properties with observations[24] indicates that global SAM is well representative of the current global Cloud-Resolving Models (CRMs) generation's ability to represent organized DCSs. Although our analysis focuses exclusively on SAM, these findings are likely applicable to other global CRMs of similar class. The reliability of the TOOCAN tracking algorithm is further substantiated by evidence showing minimal model-observation differences in cloud shield properties compared to other trackers[24] (for a comparison of DCSs in the DYAMOND simulations and in satellite observations, see Supporting Information Text $.4 and Fig. S9).

In this study, we leverage machine learning algorithms–random forest, multilinear regression, and neural network multi-layer perceptron–to predict the maximum upper-level cloud shield extension that a DCS reaches during its lifecycle. These predictions are based on the system's early development stages and initial environmental conditions from high-resolution global simulations. Our main objective is to determine whether the fate of DCSs–specifically their maximal area–is predetermined by their initial stage, and if this holds true for the largest systems exceeding 100 km. Achieving these objectives will help clarify the relative roles of internal and external processes in controlling the size of the system.

In summary, DCSs are central to extreme precipitation events and play a significant role in modulating the Earth's radiative energy budget. Accurately predicting their mature size is therefore essential, but what is even more crucial is gaining a better understanding of the processes that govern

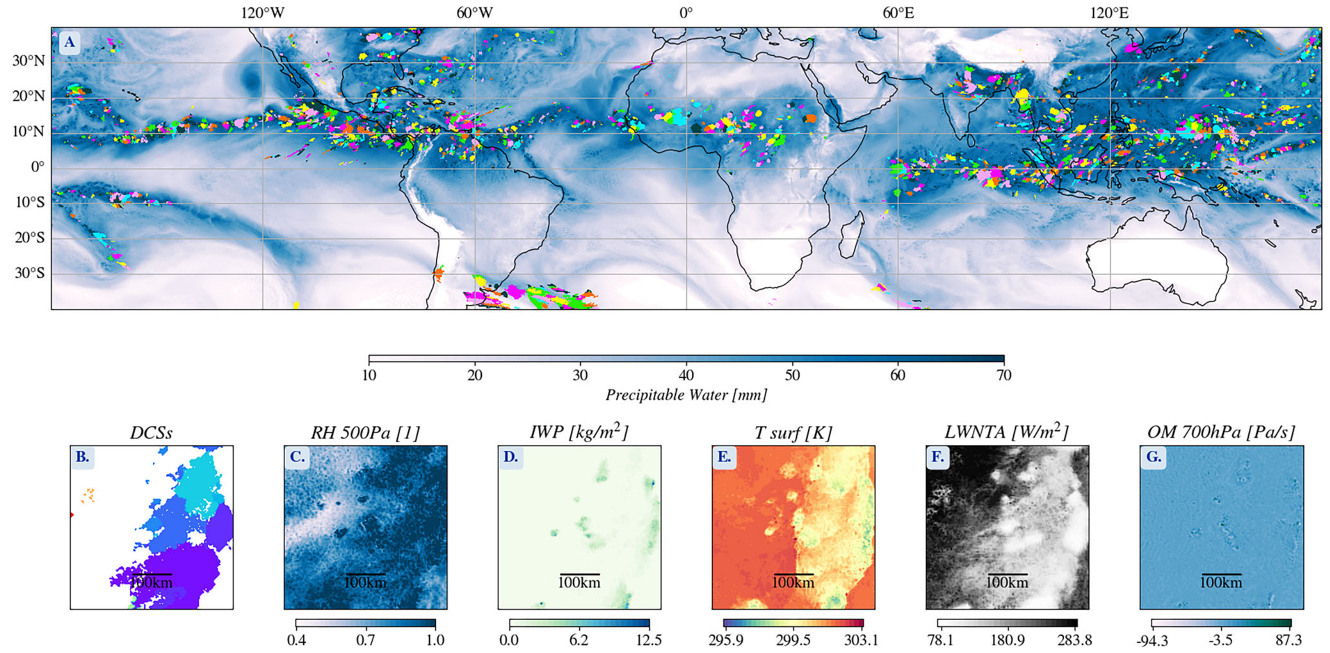

Fig. 1 | Tropicwide overview of deep convective systems in DYAMOND-SAM tracked by TOOCAN. **A** Snapshot from the DYAMOND-SAM summer simulation (4 km horizontal resolution) during August 23, 2016 at 23 h 30 of the Precipitable Water with on top the Deep Convective Systems tracked by TOOCAN focusing only on the tropics. bottom. Example of considered physical fields for a given DCS at 1 h of development : **B** the system and its neighbors. **C** Relative humidity field. **D** Ice water path. **E** Surface temperature. **F** Long wave emission. **G** Vertical velocity at 700 hPa (note the saturated colorbar to ease visualization).

their evolution and identifying the factors that drive these changes. Given the relatively straightforward life cycle of DCSs, it is likely that their early developmental stages, particularly their growth rate, largely determine their maximum extent. This study seeks to test this broader hypothesis by addressing three key questions:

- Can the maximum area of a DCS be reliably predicted based on its early growth rate?
- Does the accuracy of predictions improve when using a broader set of physical features, without explicitly accounting for growth rate? Which types of systems show better predictive performance under this approach, and what factors contribute to this accuracy?
- Which physical features, those associated with the DCS itself (internal) or its surrounding environment (external), have a stronger impact on the prediction?

Machine learning algorithms are used to investigate the relationship between the onset of DCS growth and its maximum area. The DCSs analyzed, the machine learning pipelines applied, and the two experiments (one based solely on growth rate area and the other incorporating additional features) are outlined in the section Methods. The subsequent three sections address each of the key questions in turn.

Results

Prediction of maximal size with growth rate only

Focusing first on the results when the learning relies solely on the initial evolution of the growth rate of the area, we begin by examining the impact of the observation period of the system on the final prediction of its maximum extension ($\mathcal{L}_{max} = \sqrt{A_{max}}$ where A_{max} denotes the maximum area of the DCS, see Materials and Methods). As mentioned in the introduction, the average maximal extension is 115 km with a standard deviation equal to 45 km. Figure 2A shows the evolution of the mean squared error and the R-Squared index for each trained model (random forest, linear regression (lasso), and neural network multilayer perceptron (mlp)) based on the observation period of the growth rate, ranging from 30 min (frequency of outputs from the DYAMOND-SAM simulation for the two-dimensional variables used here) up to 5 h. Firstly, we can see that increasing the observation period improves the models performance, such that after 5 h, they can predict the final extension of the system with an average accuracy of 10 to 15 km. However, since the systems on average last 7.5 h and reach their maximum area after around 3 h, beyond this point, the task becomes too easy, and the models detect it effortlessly. Secondly, with 1.5 h of observation, all three models predict the maximum system size with a score of ~0.5 and an average error of about 35 km. This experiment demonstrates a strong relationship between the initial evolution of the growth rate and the maximum extension of the system.

Finally, the similarity in performance across the three models suggests a near-linear relationship, as this pattern is effectively captured by the multilinear model within the first three hours. These results indicate that the explosiveness of the system partly determines its maximum size, which aligns with previous findings[25–27]. This raises questions about whether the simple growth rate theoretical model described in the introduction[19] is optimal to predict DCSs maximum area, whether the score of 0.5 can be improved by adding features, or whether adding features makes it possible to predict the maximum system size from as early as one hour and a half. To address these questions, we will examine the results of the second experiment in the next section.

Prediction with physical features

In this section, we predict the maximum size of DCSs by analyzing key physical features observed during their early development, aiming to identify the primary drivers of growth. To that end, we retrain the model without the growth rate but with physical features. These features, extracted from high-resolution SAM-Dyamond simulations, include humidity, atmospheric instability, vertical lifting, and wind shear, derived from both 2D and 3D data outputs (Table 1). To differentiate internal feedback processes from external influences, we compute scalar metrics–mean and standard deviation–within the DCS and within a 5° by 5° surrounding region. This distinction enables us to capture both system-specific features and environmental context. Additionally, we incorporate metrics related to neighboring convective systems, such as their number, size, and proximity, which can affect moisture distribution and lifting mechanisms[28] (Table 2). We also account for system-level attributes, including migration distance, geographic position, and local time, to refine our understanding of DCS growth (Table 2).

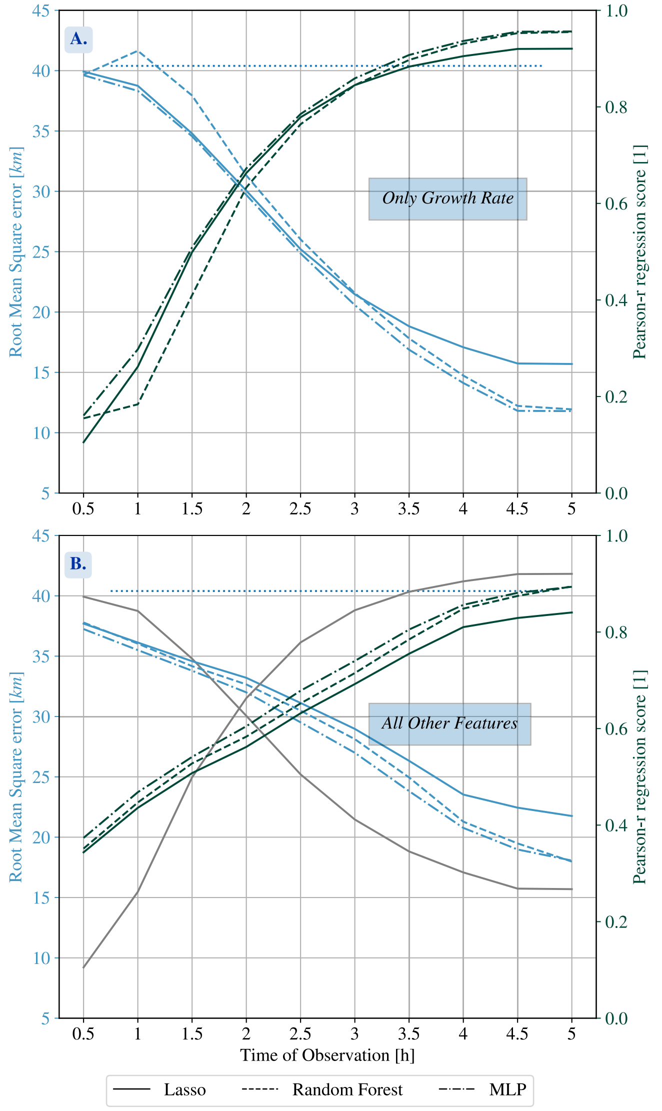

Fig. 2 | Machine learning model performance in predicting system size over varying observation periods. **A** Evolution of Pearson-r regression score in green and mean square error in blue for the estimate of maximum extension $\mathcal{L}_{max}$ in the three machine learning models – Linear Regularized Regression in solid line, Multi Layer Perceptron in dash-dotted line and Random Forest in dashed line – trained on the evolution of the growth rate during a considered observed period. **B** Same with physical features considered, which include system shape, physical field mean and standard deviation, trajectory and neighbor influence. The Multi Linear model performances from the top panel are repeated in the lower panel in gray to ease comparison (ascending is pearson-r, descending is rmse). The horizontal blue dotted line indicates the standard deviation of the dataset, representing the performance of a random prediction.

The first results of this second experiment are shown in Fig. 2B. As before, this panel shows the evolution of the root mean squared error and the R-Squared index for each trained model (random forest, linear regression, and neural network) based on the observation period of features, ranging from 30 min up to 5 h. We again observe an increasing performance, as the observation period increases. Compared to the previous experiment, we find lower quantitative error from the initial 30 min, and after 1.5 h of observation, we reach a score of 0.55 for the neural network model and an average error of less than 35 km. With 2 h of observations, the score reaches 0.6 for the three models and the error decreases to less than 30 km. We note in passing that the accuracy of the prediction is not greatly influenced the duration of systems (see Supplementary Fig. S3). Our findings indicate that after 1.5 h of observation, the growth rate emerges as the most reliable predictor of system size. However, prior to this point, a combination of additional features appears to be more influential. Our current focus is on identifying which features are critical in these early stages and understanding the mechanisms behind their predictive power, as detailed in the section “Feature importance and dimensionality reduction”.

Before investigating the key features that allow to predict the size of the systems, we aim to identify which systems are most sensitive to the inclusion of new features, by analyzing how predictions from the multilinear model, which includes all features, vary across different DCS sizes. Although other models offer slightly better predictive performance (see Supplementary Figs. S4A–D and S5A–D), we focus on the multilinear model for its clearer interpretability of individual features. Figure 3 presents a one-to-one comparison of the model’s predictions (y-axis) with the ground truth (x-axis). Note that our goal here is not necessarily to optimize the prediction, but rather to investigate the factors impacting the prediction, and to clarify how much of the DCS fate is written from the start of its lifecycle. We therefore focus on the early 1.5 h and 2 h of observation, and investigate the sensitivity of the prediction to the period of observation used, to the system size, and to the variables included in the learning. Figure 3A shows results where all other features during the first 1.5 h are provided for the training of the model, and Fig. 3B shows the same for 2 h. We can see that adding times of observation has led to an improvement in prediction for both small systems (smaller than ~ 100 km) and larger systems (larger than ~100 km). Notably, a positive bias is observed for smaller and more frequent systems; however, this bias is not attributed to overfitting, as our analysis confirms (not shown).

Table 1 | Physical fields used from SAM-Dyamond in the second experiment

Physical Variables from SAM-Dyamond used for feature extraction	2D or 3D
Precipitable Water	2D
Relative Humidity at 500 hPa	2D
Relative Humidity at 700 hPa	2D
Surface Temperature	2D
Ice Water Path	2D
U-component of Wind at 10 m	2D
V-component of Wind at 10 m	2D
Land Mask	2D
Omega at 500 hPa (Vertical Velocity)	2D
Omega at 700 hPa (Vertical Velocity)	2D
Omega at 850 hPa (Vertical Velocity)	2D
Wind Shear	3D
Deep Wind Shear	3D
Wind Shear (longitudinal)	3D
Deep Wind Shear (longitudinal)	3D
Difference in Moist Static Energy between Mid-Troposphere and Boundary Layer	3D

The ‘2D or 3D’ column indicates whether the variables are directly available from 2D outputs or derived from 3D data outputs.

We can quantify this observation by looking at Fig. 3C, which compares the evolution of the mean square error as a function of the maximal size of DCSs for the multilinear model. Results are shown for the experiment with all features, as a function of the observation period (curves with different shades). For small systems and large systems alike, increasing the observation period used for training improves accuracy. With every additional 30 min of observation, the error is reduced by about 5 km.

Prediction accuracy is highest for systems smaller than 100–120 km, while performance declines for larger systems, with regression scores remaining below 0.5 even after two hours of observation (not shown). This reduced skill may result from two factors: statistical bias, as large systems are less frequent, or additional processes influencing large systems that are not captured by early growth alone. However, since large systems (≥120 km) still comprise nearly 25% of the dataset (see Percentile Rank in Fig. 3C), sampling limitations are unlikely to fully explain this discrepancy. We note in passing that restricting training to mesoscale convective systems would introduce a survival bias: the model may overlook features that appear uniform within this subset but actually differentiate large systems from smaller or less mature ones. Consequently, key variables that explain the transition from small to large systems could be underrepresented, limiting the model’s ability to capture essential developmental processes.

This highlights an inherent trade-off: training across a wide range of systems enhances generality but may reduce accuracy for rare, extreme cases. To test the robustness of our interpretation, we introduce a weighted loss function that emphasizes larger systems by increasing their contribution to the error. While this adjustment leads to a modest decline in overall performance (not shown), the main conclusions remain consistent. This sensitivity analysis lends confidence to the stability of the results, particularly in the interpretation presented in the following section, where we examine what the model learns and the relative importance of environmental versus intrinsic system properties (notably, results with this new loss function confirm the list of most important variables listed below Fig. 4 and the dominance of internal variables for large systems, not shown).

Feature importance and dimensionality reduction

Focusing solely on the 1.5-h prediction, we now seek to identify the features that contribute to predicting the maximum extension. As mentioned in the introduction, from the literature we expect variables related to lift, instability, and moisture to matter, as well as land/ocean contrasts. The former can be quantified with large-scale ascendance/subsidence, the second with CAPE, which we approximate with the MSE difference between the boundary layer and free troposphere. The third is quantified with RH, which matters through its effect on entrainment, and the fourth is identified

Table 2 | List of features associated with surrounding systems and system-level attributes

Features characterizing the surrounding DCSs	Features characterizing the system of interest
Number of Surrounding Systems	Eccentricity (core and envelope)
Average and Maximal Age of Surrounding Systems	Geographic Position (latitude and longitude)
Average and Maximal Size of Surrounding Systems	Local Time (for diurnal cycle)
Average and Maximal Distance of Surrounding Systems	Migration Distance
Average and Maximal Distance of Surrounding Systems Weighted by Their Size	

The left column lists features related to the surrounding DCSs, while the right column describes system-level attributes. We evaluate neighboring DCSs by examining either the average or maximum value of specific characteristics (e.g., size).

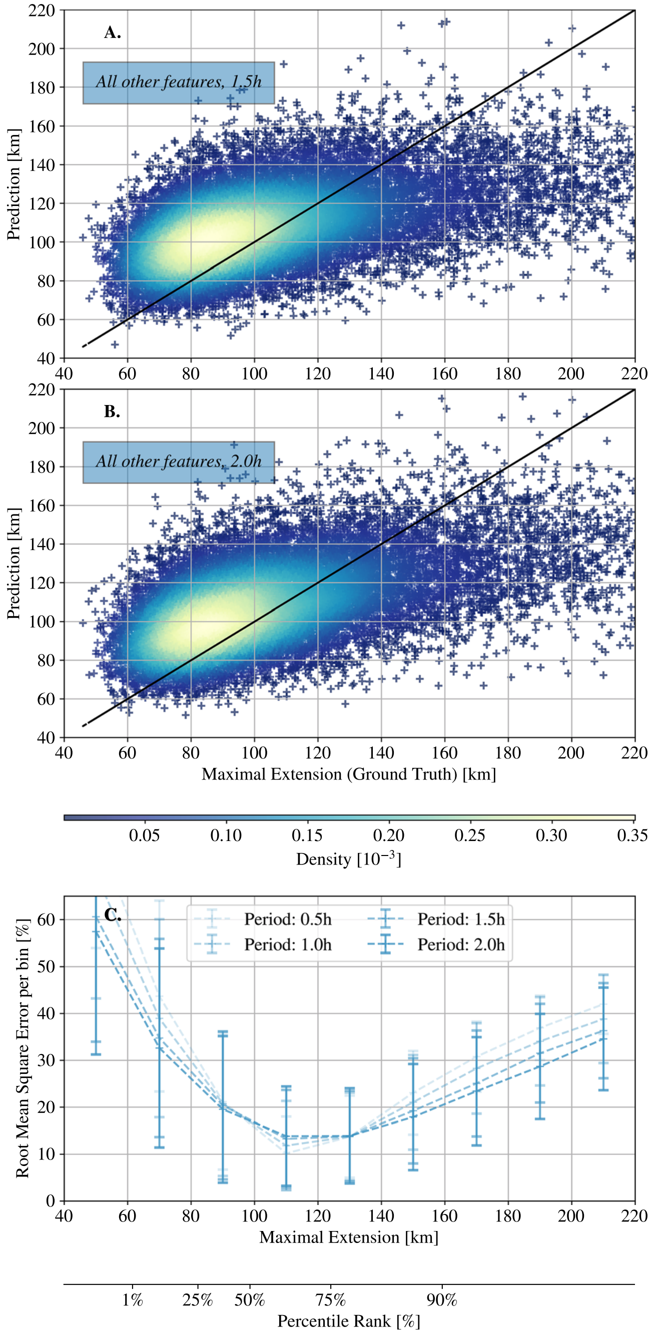

Fig. 3 | One-to-one comparison of predicted and observed system sizes using physical features in a multi-linear model. One-to-one diagrams for the prediction and the target for the multi-linear model trained with all features, including shape, physical fields, migration distance and neighboring systems influence, based on **A** 1.5 h of observation and **B** 2 h of observation. The color of the scatters represents the density of the points, calculated as the joint distribution of (x, y). **C** Evolution of the relative mean square error for all systems of a given maximal extension for different periods of training. The percentile corresponding to a given maximal extension is indicated below the x-axis.

with the landmask fraction. We further classify these processes into internal and external contributions, by averaging the relevant variables under the cloud system only, or in the environment (5°x 5° domain centered at the DCS, lower panels of Fig. 1).

Since all three models exhibit similar performance, we focus here on the linear Lasso model, which is easier to interpret. We determine the important features for the linear model with the amplitude of coefficients assigned to each variable. Figure 4A shows the top 15 important variables for this model. We note in passing that we can also identify the important features of the random forest, using the average Gini index score over all decision trees for each variable. Although they do not appear in exactly the same order, we find the same first variables of importance for both models to be consistent (see Supplementary Fig. S6).

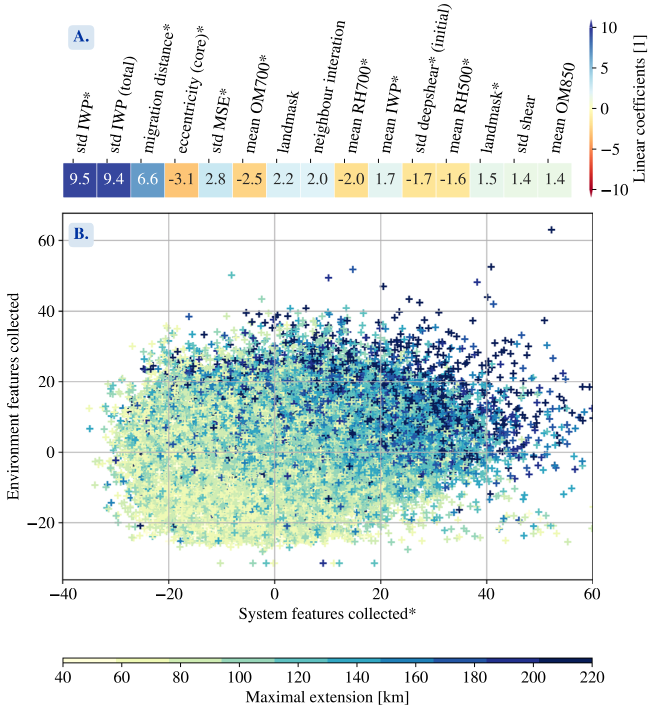

Fig. 4 | Feature contributions to multi-linear prediction of maximum system extension. A Coefficients optimized for the multi-linear model to solve the supervised task of predicting the maximal extension based on the first 1.5 h evolution of DCS growth rate and development features. The coefficients (denoted c_i in [(4)]) are sorted with their absolute value. There are 15 features displayed, other features are considered negligeable (coefficient are below 0.1). These coefficients are applied to normalized input (denoted f_i in [(4)]). Features with stars are associated with the system's own characteristic, and those without stars are associated with the environment. **B** This figure represents the dependence of the final maximal extension of a given DCS with respect to the initial components from its own characteristics (x-axis, example : eccentricity, migration distance) and the ones from its initial environment (y-axis, example :neighbor temperature, mean vertical velocity). It shows that the multi-linear regression is sensitive to both the environment and the system to predict the maximal extension, with a dominant effect from the system itself (stronger gradient in the horizontal direction).

Among these variables, we identify the heterogeneity of the ice water path (std IWP) distribution both within and outside the system, the migration distance, the eccentricity of the DCS core, the average IWP within the cloud anvil, the standard deviation of integrated moist static energy (MSE) associated with the system, the land mask under the system and outside the system, the mean vertical velocity at 700 hPa and 850 hPa, the mean longwave radiation emitted by all DCSs in the environment, relative humidity at 500 hPa and 700 hPa, and the initial deep shear within the system. These 15 variables collectively form a robust set for predicting the maximum extension of the system. Importantly, these variables are not independent (see supplementary Fig. S7 for the correlation matrix). But we note that the correlations between the different variables are always smaller than 0.5.

Among the variables of importance, we recover the expected contributions from lift, instability, moisture, and landmask. The lift is captured by the mean vertical velocity omega in the system and in its environment. Instability is reflected through the system's MSE (which is indicative of CAPE) and the ice water path, which is associated with the altitude reached by deep convection. Ice formation thus serves as a proxy for both convective velocities and the vertical extent of convection. Additionally, the standard

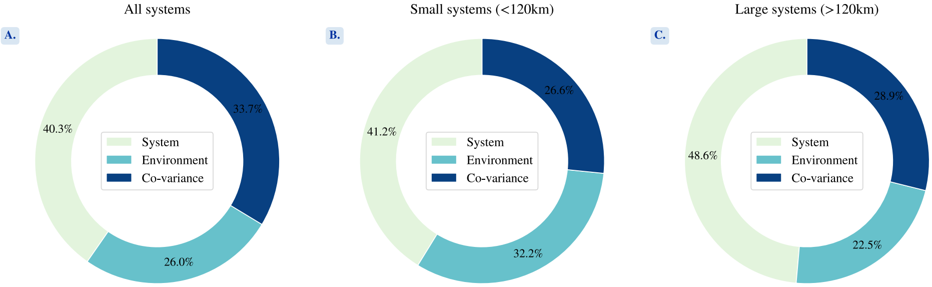

Fig. 5 | Partitioning variance in dcs maximal extension between system and environmental factors. Fraction of the total variance of the DCS maximal extension prediction explained by the system or environmental conditions for **A** all systems, **B** systems smaller than 120 km and **C** larger than 120 km.

deviation of ice water path (IWP) emerges as a strong predictor of system growth, highlighting the role of spatial heterogeneity in ice content. We interpret the variability of IWP as an indicator for the number of intense convective sources within the system. This aligns with the idea that multiple convective sources can enhance the overall growth rate, as described in Eq. (1). This interpretation is further supported by a statistical analysis of 4000 randomly sampled systems, which reveals a positive correlation between the standard deviation of IWP and the number of convective clusters (not shown). Moisture, with relative humidity (RH) being the key variable, is critical as it governs the role of entrainment in limiting the development of deep convection. Landmask, both within the system and its environment, also plays a significant role. Additionally, the role of shear, particularly deep-layer shear, in organizing convection is highlighted, consistent with recent studies[29,30]. This pronounced influence of shear may be partially linked to its effect on squall line organization[31,32], which could also be related to the migration distance and eccentricity variables, that emerge as significant factors[33]. The final contribution, which, to the best of our knowledge, has not been emphasized in DCS studies, pertains to the influence of neighboring systems. One expected impact of nearby systems is the reduction in available CAPE, as the potential instability is distributed among all the systems triggering convection. However, even with CAPE included in our variables, the presence of neighboring systems remains a significant factor, with a positive coefficient indicating a favorable impact on the growth. This may be linked to conditions that promote aggregation (as evidenced by the presence of surrounding systems) which would provide a positive feedback on the growth and warrants further investigation.

Having identified the key predictors of DCS size, we now further examine the relative contributions of internal system dynamics versus external environmental factors in shaping system size. To do so, we decompose the variability in DCS size into two primary axes: one representing the system's intrinsic characteristics (features shown with an *) and the other reflecting environmental conditions (other features). Our goal is to determine whether these two axes account for the primary sources of variation in system size, if they have a relatively balance role and to understand how they relate to both small and large systems. In summary, we seek to establish whether DCS size is predominantly influenced by internal dynamics, environmental factors, or a combination of both, and whether the relative importance of these drivers differs between small and large systems. To do so, in Fig. 4, we have projected the DCSs onto these two axis, where the x-axis encapsulates the system's internal conditions (which have stars in Fig. 4A), while the y-axis represents the external environmental conditions. To be more precise, we can express this as follows:

$$\mathcal{L}_{max} = \mathcal{L}_0 + \ell(f_1, f_2, ..f_n) \sim \mathcal{L}_0 + \ell(f_1, f_2, ..f_{15}) \tag{2}$$

$$\sim \mathcal{L}_0 + c_1 f_1 + c_2 f_2 + ... + c_{15} f_{15}, \tag{3}$$

where $\mathcal{L}_{max}$ denotes the maximum extension of the system, $\mathcal{L}_0$ the average DCS extension, ℓ is the linear form trained on all features, f_i are the features in order of importance (normalized by removing the mean and dividing by the standard deviation), and c_i are the associated coefficients. We can then separate the features into those related to the system (indicated by stars) and those related to the environment:

$$\mathcal{L}_{max} = \mathcal{L}_0 + \sum_{i \in sys*} c_i f_i^* + \sum_{j \in env} c_j f_j. \tag{4}$$

We can calculate these two terms for all DCSs and observe how the maximum extension of the system varies in this phase diagram, similar to a principal component analysis (PCA, see Supplementary Fig. S8 for a PCA analysis which shows a similar clustering of small and large systems as in Fig. 4; here we focus on the decomposition of Fig. 4 as it allows to directly interpret the axes as internal and external variables). The result is shown in Fig. 4B. The color of the markers represents the maximum extension of the system. We observe that the maximum extension increases linearly with the increase in both axes. For very low values of x, we see that the system will be small regardless of the value of y. Beyond this, both the environment and the system jointly contribute to the prediction.

We further quantify the relative roles of the environment and of the system in the prediction of the maximal extension by computing the explained variance, following the same method as described in ref. 34. Applying the variance -a non-linear operator- to eq. [(4)], it writes,

$$V(\mathcal{L}_{max} - \mathcal{L}_0) = V\left(\sum_{i \in sys*} c_i f_i^*\right) + V\left(\sum_{j \in env} c_j f_j\right) + 2COV\left(\sum_{i \in sys*} c_i f_i^*, \sum_{j \in env} c_j f_j\right),$$

where V denotes the variance over the systems, and COV the co-variance, also along the system set.

Figure 5A shows the three contributions of the right-hand side as percentages of the total variance explained for all DCSs (for the prediction using all features during the first 1.5 h of the system). As noted in Fig. 4, both the system and the environment contribute to the prediction, and Fig. 5A further shows that internal processes play a predominant role, accounting for 40.3% of the variance in maximal extension across all systems, while environmental factors account for 26%.

Figure 5B, C show these same contributions separately for relatively small systems (<120 km) and relatively large systems (>120 km) respectively (the 120 km cutting scale was used as it corresponds to the scale beyond which the score falls by a factor of about 2, not shown). We analyze small and large systems separately, anticipating that larger systems develop distinct internal dynamics and are thus more strongly governed by internal feedbacks than by environmental conditions. It is well established that DCSs reaching mesoscale develop complex internal circulations that sustain convection and moisture inflow[35]. Our analysis confirms that internal processes increasingly dominate in larger systems. In contrast, smaller systems are more influenced by environmental factors, the variance being impacted by the greater variability of the ice water path (IWP) in their environment.

Consistent with these expectations, for small systems, the fraction of the variability in their maximal extension explained by their initial environment is larger than for all systems (32.2%); conversely for the larger systems, the maximal extension seems to depend largely on the system's characteristics. More precisely, the ratio between the system and the environment contribution to the variance reaches 42% at 80 km (not shown) and keeps increasing as the size of systems does. These results are based on the first 1.5 h. Sensitivity tests (not shown) indicate that longer observation periods increase variability, highlighting a need for further analysis. But overall, this suggests that large systems have the potential to create their own conditions and internal feedbacks favoring cloud shield growth. Given the strong societal and climate impacts of DCSs that reach mesoscales, these results open new interesting research avenues to address the nature and strength of internal feedbacks leading to DCS growth beyond that predicted from initial environmental conditions.

Discussion

The objective of this study is to conduct a data-driven analysis of storms, with a focus on clarifying the relationship between DCS and associated variables. Previous literature has identified a range of commonly considered factors (e.g., lift, instability, etc.). In this work, we incorporate additional variables into machine learning models to allow the data to reveal potentially overlooked or unexpected factors. The ultimate aim is to identify the key processes that influence the extent of DCS and to highlight new, promising research directions that warrant further investigation.

To that end, we investigated the relationship between DCSs early stage of development, and their maximal size. We utilized global high-resolution simulation with the model SAM from the DYAMOND project, coupled with the storm tracking algorithm TOOCAN. To predict the maximum extension of DCSs, we compared three different machine learning models based on early development stages and initial environmental conditions of the systems. Interestingly, we find that the initial growth rate of the system area strongly anticipates its eventual maximum extension. By analyzing growth rates during the first two hours of development, we achieved a regression score of 0.65, regardless of the system's ultimate lifespan. As all three models exhibited similar performances, this shows a near-linear relationship between growth rate and system size. However, with only one hour of data, the score dropped to around 0.25, suggesting that the growth rate itself is the result of earlier underlying processes. Incorporating additional variables – including physical fields, features of surrounding DCSs, and characteristics of the system itself – into a neural network model improved the regression score from 0.25 to 0.47 within the first hour. This increase in predictive accuracy enabled us to identify key factors influencing DCS size, offering new insights into the roles of both environmental and internal processes.

Several key factors emerged as important for predicting the behavior of DCSs, reflecting both well-known and less expected influences. As expected, lift, instability, moisture, and land-ocean differences play major roles. Lift is captured by vertical velocity, while instability is represented by moist static energy (MSE)–a proxy for CAPE–and the ice water path, which reflects the depth of convection and, through its standard deviation, the number of convective updraft sources. Moisture is mainly controlled by relative humidity, which affects entrainment, and the land mask highlights the different behavior of convection over land and ocean. Our analysis also confirms the importance of deep-layer shear, a factor that helps organize convective systems and supports the development of squall lines. Although shear has been recognized in past studies, our statistical analysis further highlights its relevance.

More unexpectedly, features associated with surrounding systems also emerge as significant. Even more surprising is that these features appear alongside our proxy for CAPE, suggesting that the influence of neighboring systems extends beyond merely modifying available energy, possibly by promoting aggregation. These additional factors could only be uncovered through a data-driven approach, in which the model was supplied with a broader set of variables beyond the conventional drivers of deep convection.

Several directions could help improve the performance of data-driven models. One promising approach is to use causal inference techniques–such as those proposed by Beucler et al.[36]–to identify true cause-and-effect relationships, rather than relying only on correlations. Although the models include many input variables, their skill scores after 1.5 h of prediction remain moderate (around 0.5–0.6 for linear and neural network models), suggesting that important factors may still be missing. A better understanding of boundary conditions, including processes during the growth phase that are not tied to the initial state, could help capture more of the system's behavior. It may also be important to rethink how spatial information is used. Instead of relying on averages and standard deviation, applying convolutional neural networks could help detect spatial patterns and interactions between fields, potentially improving prediction. These improvements could make the models more accurate and allow the approach to be extended to other datasets, including satellite observations[37].

Although further improvements to the modeling framework are possible–particularly in feature selection and the treatment of spatial information–the current analysis already provides significant insight. We find that both internal and external drivers contribute to predicting the maximal size of DCSs, with contributions of 40.3% and 26%, respectively. This indicates that a system's initial environment does not necessarily limit its eventual size. This effect is especially pronounced for larger systems, whose growth is more significantly influenced by their internal characteristics (48.6%). This aligns with the well-documented internal circulation that DCSs can develop, enabling them to expand to mesoscales[35]. This observation is particularly intriguing when considering systems developing within a cohort of DCSs, where CAPE is likely low (consumed by convection within the DCS and its neighbors). Understanding the organization between mesoscale convective systems and the processes that allow a system to thrive among many others warrants further in-depth research.

Methods

Global CRM coupled with storm tracking algorithm

Current satellite data does not provide complete information on the vertical structure of variables, notably dynamical fields, therefore, we have turned to simulation data for our study. To investigate the life cycle of DCSs, we rely on high-resolution (4 km) global simulation with the model System for Atmospheric Model (SAM[20]), an anelastic model of fluid dynamics with parametrized microphysics (part of the DYAMOND project[21]). This global-storm resolving simulation ran for a 40-day period (1 August–10 September 2016), outputting two-dimensional variables (which we used in this study) every half hour. By resolving the transient dynamics of convective storms in the tropics, global storm-resolving models eliminate the need to parameterize tropical deep convection, leading to a more robust representation of the climate system and a more natural connection to high-resolution data from satellite-borne sensors.

Most DYAMOND models, including SAM, accurately capture essential DCS characteristics such as lifetime, cloud shield area, and volume of rainfall[38]. Simulated DCS movement speeds over the ocean generally agree with observations, but over land, some models produce faster speeds, possibly indicating stronger cold pool intensities that promote DCS movements[32]. To detect, track, and measure the evolution of DCSs, we leverage the capabilities of the cloud tracking algorithm TOOCAN[23], which relies on a definition of a DCS consisting in a combination of a convective core, characterized by low brightness temperature, associated to an anvil cloud, characterized by relatively higher brightness temperature than convective cores. These components evolve over time, and the TOOCAN algorithm links the convective cores with their respective anvil clouds within a spatio-temporal domain to identify individual convective systems. This is achieved by processing a spatio-temporal volume of infrared images through an iterative process of detection and dilatation of convective seeds in three dimensions. This process continues until it encounters the outer limits of the high cold cloud shield, demarcated by a brightness temperature threshold of 235K (calibration details can be found in the Supplementary section Tracking algorithm calibration). As a result, TOOCAN is able to identify and track individual DCSs in a single process, and to partition the high cold cloud shield into DCS components. The DCSs identified through this process exhibit a broad spectrum of convective organization, from short-lived, small and isolated systems, to long-lived and large systems[39], similar to observations (see Supplementary section, Assessing realistic properties of MCSs in DYAMOND-SAM simulation and Fig. S8). The initial dataset comprises 287,031 simulated systems during August and September 2016 worldwide, which were then pre-processed prior to analysis as described next.

Pre-processing of data

In the following paragraphs, we describe the pre-processing steps applied to the TOOCAN dataset prior to using machine learning algorithms. From the systems tracked by TOOCAN, we focus solely on tropical DCSs, restricting our analysis to within ±30 degrees of latitude. Our particular interest lies in relatively long-lasting and large-scale systems, defined as those with a minimum lifespan of 5 h and a minimum extent of at least 40 km. From the initial systems, we then focus on 107,582 DCSs that fit these criteria.

During the analysis of system life cycles, we noticed that some take time to dissipate, as shown in Supplementary Fig. S10, and although they last for more than 5 h, their effective duration is shorter. In a smaller proportion of cases, a system may also experience a delayed growth phase. In both instances, we chose to concentrate on the *active* life cycle of the systems, namely those with significant growth and decay rates of area, at the boundary of the life cycle (see Fig. S10A for an illustration). In other words, these boundaries are meant to represent the start and end points of the system's active life cyle. To determine these thresholds, we aim to strike a balance between maximizing the correlation between maximum area and lifespan, as expected from observations, while minimizing the number of systems removed (often active cycles fall below the 5-hour minimum lifespan threshold). These thresholds are described in more detail in the supplementary Fig. S10B, which depicts the evolution of the first criterion (correlation between area and lifespan) and the second criterion (number of systems removed) based on different thresholds applied to the growth and decay rates. An optimal compromise is found at 1000 km^2/h. It is worth noting that the growth rate of area threshold remains fixed and independent of the maximum area of the systems, ensuring there are no a priori biases or information embedded in this threshold.

Finally, the DCS life cycle, and size, lifespan, and growth time, are illustrated in Fig. 6A for a few DCSs, alongside the joint distribution of lifespan and maximum extension shown in Fig. 6B. This dataset of 68,913 systems, coupled with high-resolution physical fields, is the dataset used for the application of machine learning. In the next section, we will describe the protocols utilizing machine learning algorithms to address the overarching question of this study: What determines, during the early stages of growth, whether a system will become large, and why?

Implementation of Machine Learning Pipelines : models, input, output and error quantification

Handling all this data is challenging, especially considering all the dynamic and thermodynamic fields accessible for all systems. The machine learning approach can be seen as an initial step towards developing a physical model. Learning occurs when a program solves tasks without being explicitly programmed for them. In the case of supervised learning, as is the focus here, programs primarily create a model that minimizes the average statistical error with respect to the target task.

Our method aims to first understand statistically what leads to significant extensions of the systems and then to comprehend the physical interpretation of this learning. The protocol involves training a learning model on a subset of DCSs (the train and validation datasets represent 85% of all systems) to predict the maximum extension of a DCS based on its early growth information. Subsequently, we assess its performance on new systems, separately in a test set (the remaining 15% of DCSs, randomly chosen). In the following, we define the model used and precisely specify the input characteristics of the model. We compare three methods: a multilinear model, a neural network, and a random forest.

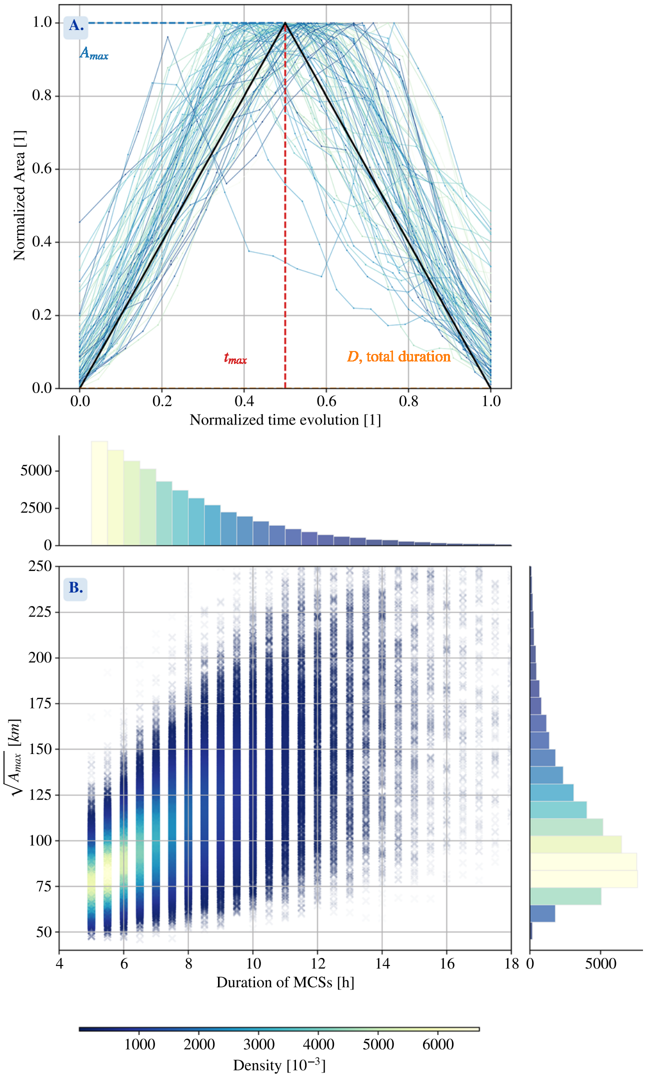

Fig. 6 | Life cycle evolution and size-duration relationship of deep convective systems. A Superimposed evolution of the life cycle of tracked Deep Convective Systems (arbitrarily chosen) - i.e., the normalized area evolution as a function of the normalized time for 100 systems. The solid black line represent the simple theoretical model proposed in ref. 15 which captures the life cycle with 3 parameters, the maximal area, the duration and the time of maximal area which is most of the time close to half of the duration. **B** Joint distribution of the duration and the maximal extension for all considered systems. The color indicates the density of points. Through post-processing of data, see Supplementary Fig. S9, correlation between maximal extension and duration is equal to 0.67 consistent with observations of deep convective systems.

Machine Learning models. The multilinear model optimizes the weights of a linear form that maps the input vector of system characteristics to the target maximum size. Here, we employ a lasso model, which includes weight regularization, requiring the model to minimize the prediction error and also minimize the L1 norm of the linear form. For the neural network, the principle is the same, except there are non-linear activation thresholds in the layers. Lastly, the random forest consists of a collection of decision trees. Each decision tree describes a set of possible outcomes (also known as leaves), each representing the consequence of a logical decision made at each tree node. The decision is determined by a threshold applied to an input variable. The order of variables and the threshold are optimized in the training process to minimize the prediction error. The iteration involves promoting variables with a high Gini impurity score, which represents a variable's ability to be correlated with the output compared to a random variable. The random forest aggregates predictions from the collection of trees, allowing for a more robust prediction. The advantage is that the average Gini score for each variable can be known, facilitating result interpretation. All our analysis uses the Sklearn Python Package[40].

Output variable and error quantification. We trained these three machine learning models to predict the square root of maximum area of a DCS (we made the arbitrary choice to work with square root of area instead of area, although similar results are obtained with area prediction, not shown), based on the evolution of the early stage of the system, observed at intervals ranging from 30 min up to 5 h (the minimum duration of all systems). To assess the performance of each model, we used the Pearson correlation coefficient, denoted as r, which is defined as:

$$r = \frac{cov(y_{pred}, y_{target})}{\sigma(y_{pred})\sigma(y_{target})} \tag{5}$$

where y_{pred} represents the set of model-predicted square root of maximum areas, y_{target} is the true target square root of maximum areas, cov denotes covariance, and σ denotes variance. This score measures how well the model's predictions align with the target values, accounting for any recurrent biases. It is important to note that a high correlation does not necessarily mean the model is accurate, as systematic overestimation or underestimation could still lead to a high correlation.

To evaluate if the model achieves its intended goal, we consider a second criterion, the Root Mean Squared Error (*rmse*), defined as:

$$rmse = \sqrt{E[(y_{pred} - y_{target})^2]} \tag{6}$$

where E represents the first moment operator, and y_{pred} and y_{target} are defined as above. The mean squared error is the criterion minimized during model training, meaning that the correlation coefficient still incorporates this information, but does not include it explicitly.

Input variables. Regarding the input data, as described in the introduction, we designed two experiments. In the first experiment, we input only the evolution of the growth rate of the area to predict the maximum size. More precisely, for each DCS we compute the following array

$$\left\{\frac{dA}{dt}(t = t_0 + \Delta t), \frac{dA}{dt}(t = t_0 + 2\Delta t), .., \frac{dA}{dt}(t = t_0 + n\Delta t)\right\}, \tag{7}$$

where A is the area of the cloud shield, t_0 is the birth time of the system, Δt is equal to 30 min, n is an integer less or equal to 10 (which correspond to 5 h of development), and dA/dt is estimated by finite difference between the timestep and the previous one. We conduct 10 cases, by varying the observed period of evolution of the system from 30 min to 5 h. We note in passing that in the theoretical model described in Fig. 6 the growth rate is constant in the increasing phase, but in the data there are some small variations in area growth rate which we retain. This protocol serves as our baseline and will investigate if the growth rate already incorporates environmental parameters.

In the second experiments, we predict the maximum size of DCSs by analyzing key physical features observed during their early stages of development, derived from high-resolution SAM-Dyamond simulations. Humidity is represented by relative humidity at 500 hPa and 850 hPa, obtained from the 2D outputs of the simulations. Atmospheric instability is quantified by the difference in moist static energy (MSE) between the mid-troposphere and the boundary layer, a measure derived from the 3D outputs, offering a computationally feasible alternative to more complex indices such as CAPE and CIN. Low-level instability, often linked to cold pool activity, is approximated using 2m-temperature data available in the 2D outputs. Lifting mechanisms, critical for convective development, are captured through vertical velocity fields and the ice water path (IWP), which reflects the role of ice loading within the system. Wind shear, a key factor not provided as a 2D field, is calculated from 3D velocity data by determining the difference in horizontal wind speeds between the upper and lower atmospheric levels (respectively 4000 m and 1000 m). In addition to these primary variables, supplementary 2D fields from the SAM-Dyamond simulations, as outlined in Table 1, are incorporated to investigate if they enhance the predictive capability of our data-driven methods.

Scalar features, namely the mean and standard deviation of these variables, are computed for each system and serve as inputs to our models. To distinguish between internal feedback processes and external influences, these metrics are calculated both within the DCS and within a 5° by 5° region centered on the system's barycenter. Bottom panels of Fig. 1, as well as supplementary examples in supplementary Fig. S1, show examples of physical fields for a given DCS at 1.5 h of development. This dual approach provides a comprehensive view of values within the system and in the environment surrounding the system, allowing us to better understand how internal and external conditions influence the size of mature systems. Additionally, we also include features that account for the influence of surrounding convective systems -number, age, and proximity of these neighboring systems, as detailed in the Table 2- since they may impact moisture distribution, CAPE or enhance lifting through gravity wave interactions[28]. One key features, called *interaction power* denotes the influence of each neighboring system size (see Fig. 1B), quantified by weighting its distance to the center of the system of interest (using $\exp(-(d/d_0)^2)$, where d is the distance to the center, and $d_0 = 50$ km), and then averaging over all neighboring systems to obtain the average influence factor. For the maximum influence factor, we took the maximum value instead of the average (nearest neighbor). Finally, system-level attributes, such as eccentricity, geographic position (latitude and longitude), local time, and migration distance and whether it forms over land or ocean are also incorporated into the analysis, as described in Table 2.

At each time step of the system's evolution (every 30 min), these features are calculated and concatenated into a single vector, which is then provided as input to the model. To ensure all input features are weighted equally, each variable is standardized by removing the mean and scaling to unit variance. We first apply a filter to exclude scalar features with correlations greater than 85% (except for land mask). Then, to accelerate training, we use a genetic algorithm to preselect 70% of the input data based on its statistical correlation with the target prediction. We proceed to train three models–random forest, lasso, and neural network–incorporating these selected features.

Accession codes

This study utilizes the global Cloud-Resolving Model (CRM) SAM from the DYAMOND project, which is stored alongside other global CRMs at the *Deutsches Klimarechenzentrum* (DKRZ). SAM is an open acces model (http://rossby.msrc.sunysb.edu/SAM/). The tracking algorithm TOOCAN source code and ressources can be found here https://data.ipsl.fr/catalog/srv/eng/catalog.search#/metadata/9e924ac9-7e43-4c9a-ba2e-188b0309a783. The resulting TOOCAN-SAM-DYAMOND dataset is accessible here: https://data.ipsl.fr/catalog/srv/eng/catalog.search#/metadata/0d567c47-3318-4767-8365-e5416b256aff. The processed data supporting the findings of this study are archived at the World Data Center for Climate (WDCC). These include the system's environmental fields, the train and test feature dataset, available at https://www.wdc-climate.de/ui/entry?acronym=DeepFate. The various algorithms related to the machine learning component and the main script developed in this article are freely available on the Zenodo DeepFate project repository (https://doi.org/10.5281/zenodo.12588221).

Data availability

This study utilizes the global Cloud-Resolving Model (CRM) SAM from the DYAMOND project, which is stored alongside other global CRMs at the Deutsches Klimarechenzentrum (DKRZ). SAM is an open acces model (http://rossby.msrc.sunysb.edu/SAM/). The tracking algorithm TOOCAN source code and ressources can be found here https://data.ipsl.fr/catalog/srv/eng/catalog.search#/metadata/9e924ac9-7e43-4c9a-ba2e-188b0309a783. The resulting TOOCAN-SAM-DYAMOND dataset is accessible here: https://data.ipsl.fr/catalog/srv/eng/catalog.search#/metadata/0d567c47-3318-4767-8365-e5416b256aff. The processed data supporting the findings of this study are archived at the World Data Center for Climate (WDCC). These include the system's environmental fields, the train and test feature dataset, available at https://www.wdc-climate.de/ui/entry?acronym=DeepFate. The various algorithms related to the machine learning component and the main script developed in this article are freely available on the Zenodo DeepFate project repository (https://doi.org/10.5281/zenodo.12588221).

Received: 7 May 2025; Accepted: 24 June 2025;
Published online: 08 July 2025

References

1. Stephens, G. et al. The First 30 years of GEWEX. *Bull. Am. Meteorol. Soc.* **104**, 126–157 (2022).
2. Nesbitt, S. W., Cifelli, R. & Rutledge, S. A. Storm morphology and rainfall characteristics of trmm precipitation features. *Mon. Weather Rev.* **134**, 2702–2721 (2006).
3. Roca, R., Aublanc, J., Chambon, P., Fiolleau, T. & Viltard, N. Robust observational quantification of the contribution of mesoscale convective systems to rainfall in the tropics. *J. Clim.* **27**, 4952–4958 (2014).
4. Roca, R. & Fiolleau, T. Extreme precipitation in the tropics is closely associated with long-lived convective systems. *Commun. Earth Environ.* **1**, 1–6 (2020).
5. Schiro, K. A. et al. Environmental controls on tropical mesoscale convective system precipitation intensity. *J. Atmos. Sci.* **77**, 4233–4249 (2020).
6. Khairoutdinov, M. & Emanuel, K. Rotating radiative-convective equilibrium simulated by a cloud-resolving model. *J. Adv. Model. Earth Syst.* **5**, 816–825 (2013).
7. Mauritsen, T. & Stevens, B. Missing iris effect as a possible cause of muted hydrological change and high climate sensitivity in models. *Nat. Geosci.* **8**, 346 (2015).
8. Bony, S. et al. Observed modulation of the tropical radiation budget by deep convective organization and lower-tropospheric stability. *AGU Adv.* **1**, e2019AV000155 (2020).
9. Yang, Q., Leung, L. R., Feng, Z. & Chen, X. Impact of global warming on us summertime mesoscale convective systems: a simple lagrangian parcel model perspective. *J. Clim.* **36**, 4597–4618 (2023).
10. Bolot, M. et al. Kilometer-scale global warming simulations and active sensors reveal changes in tropical deep convection. *npj Clim. Atmos. Sci.* **6**, 209 (2023).
11. Schumacher, R. S. & Rasmussen, K. L. The formation, character and changing nature of mesoscale convective systems. *Nat. Rev. Earth Environ.* **1**, 300–314 (2020).
12. Galarneau Jr, T. J. et al. Tropical mesoscale convective system formation environments. *Atmos. Sci. Lett.* **24**, e1152 (2023).
13. Yang, Q., Leung, L. R., Feng, Z. & Chen, X. A moist potential vorticity model for midlatitude long-lived mesoscale convective systems over land. *J. Atmos. Sci.* **80**, 2399–2418 (2023).
14. Cheng, Y.-M., Dias, J., Kiladis, G., Feng, Z. & Leung, L. R. Mesoscale convective systems modulated by convectively coupled equatorial waves. *Geophys. Res. Lett.* **50**, e2023GL103335 (2023).
15. Roca, R., Fiolleau, T. & Bouniol, D. A simple model of the life cycle of mesoscale convective systems cloud shield in the tropics. *J. Clim.* **30**, 4283–4298 (2017).
16. Wing, A., Emanuel, K., Holloway, C. & Muller, C. Convective self-aggregation in numerical simulations: a review. *Surv. Geophys*. **38** 1–25 (2017).
17. Wing, A. A. Self-aggregation of deep convection and its implications for climate. *Curr. Clim. Change Rep.* **5**, 1–11 (2019).
18. Muller, C. et al. Spontaneous aggregation of convective storms. *Annu. Rev. Fluid Mech.* **54**, 133–157 (2022).
19. Elsaesser, G. S., Roca, R., Fiolleau, T., Del Genio, A. D. & Wu, J. A simple model for tropical convective cloud shield area growth and decay rates informed by geostationary ir, gpm, and aqua/airs satellite data. *J. Geophys. Res. Atmos.* **127**, e2021JD035599 (2022).
20. Khairoutdinov, M. F. & Randall, D. A. Cloud resolving modeling of the arm summer 1997 iop: Model formulation, results, uncertainties, and sensitivities. *J. Atmos. Sci.* **60**, 607–625 (2003).
21. Stevens, B. et al. DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. *Prog. Earth Planet. Sci.* **6**, 1–17 (2019).
22. Khairoutdinov, M. F., Blossey, P. N. & Bretherton, C. S. Global system for atmospheric modeling: model description and preliminary results. *J. Adv. Model. Earth Syst.* **14**, e2021MS002968 (2022).
23. Fiolleau, T. & Roca, R. An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite. *IEEE Trans. Geosci. Remote Sens.* **51**, 4302–4315 (2013).
24. Feng, Z. et al. Mesoscale convective systems tracking method intercomparison (mcsmip): application to dyamond global km-scale simulations. *Geophys. Res. Atmos.* **130**, e2024JD042204 (2025).
25. Coniglio, M. C., Hwang, J. Y. & Stensrud, D. J. Environmental factors in the upscale growth and longevity of mcss derived from rapid update cycle analyses. *Mon. Wea. Rev.* **138**, 3514–3539 (2010).
26. Fritsch, J. & Forbes, G. Mesoscale convective systems. in *Severe convective storms*, 323–357 (Springer, 2001).
27. McAnelly, R. L. & Cotton, W. R. Meso-β-scale characteristics of an episode of meso-α-scale convective complexes. *Mon. Weather Rev.* **114**, 1740–1770 (1986).
28. Mapes, B. E. Gregarious tropical convection. *J. Atmos. Sci.* **50**, 2026–2037 (1993).
29. Chen, X., Leung, L. R., Feng, Z. & Yang, Q. Environmental controls on mcs lifetime rainfall over tropical oceans. *Geophys. Res. Lett.* **50**, e2023GL103267 (2023).
30. Roca, R., Fiolleau, T. & Netz, L. Scale-dependence of tropical oceanic deep convective systems' cloud shield morphology to environmental conditions. *Geophys. Res. Atmos.* **130**, e2024JD042494 (2024).
31. Rotunno, R., Klemp, J. B. & Weisman, M. L. A theory for strong, long-lived squall lines. *J. Atmos. Sci.* **45**, 463–4 (1988).
32. Abramian, S., Muller, C. & Risi, C. Shear-convection interactions and orientation of tropical squall lines. *Geophys. Res. Lett.* **49**, e2021GL095184 (2022).
33. Carbone, R., Conway, J., Crook, N. & Moncrieff, M. The generation and propagation of a nocturnal squall line. part i: observations and implications for mesoscale predictability. *Mon. Weather Rev.* **118**, 26–49 (1990).
34. Chakraborty, S., Fu, R., Massie, S. T. & Stephens, G. Relative influence of meteorological conditions and aerosols on the lifetime of mesoscale convective systems. *Proc. Natl. Acad. Sci. USA* **113**, 7426–7431 (2016).
35. Houze, R. A. Jr Mesoscale convective systems.*Rev. Geophys* **42**, RG4003 (2004).
36. Beucler, T. et al. Selecting robust features for machine-learning applications using multidata causal discovery. *Environ. Data Sci.* **2**, e27 (2023).
37. Fiolleau, T. & Roca, R. A database of deep convective systems derived from the intercalibrated meteorological geostationary satellite fleet and the toocan algorithm (2012–2020). *Earth Syst. Sci. Data* **16**, 4021–4050 (2024).
38. Feng, Z. et al. Mesoscale convective systems in dyamond global convection-permitting simulations. *Geophys. Res. Lett.* **50**, e2022GL102603 (2023).

39. Fiolleau, T. & Roca, R. A deep convective systems database derived from the intercalibrated meteorological geostationary satellite fleet and the toocan algorithm (2012–2020). *Earth Syst. Sci. Data Discuss.* **2024**, 1–42 (2024).
40. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

C.M. and S.A. gratefully acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Project CLUSTER, grant agreement 805041), and from the PhD fellowship of Ecole Normale Supérieure de Paris-Saclay. DYAMOND data management was provided by the German Climate Computing Center (DKRZ) and supported through the projects ESiWACE and ESiWACE2. The projects ESiWACE and ESiWACE2 have received funding from the European Union’s Horizon 2020 research and innovation program under grant agreements No 675191 and 823988. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project IDs bk1040 and bb1153. The authors express their gratitude to Sophie Cloché and Eileen Hertwig for their assistance in data archival at IPSL and DKRZ, respectively. We also thank Christophe Lampert and Benjamin Fildier for valuable scientific discussions, and acknowledge the thoughtful comments of two anonymous reviewers.

Author contributions

S.A., C.M., C.R., and R.R. designed research; S.A. performed research; S.A., C.M., T.F., and R.R. contributed new analytic tools; S.A., C.M., C.R., and R.R. analyzed data; and S.A., C.M., C.R., and T.F. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-025-01154-1.

Correspondence and requests for materials should be addressed to Sophie Abramian.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.