How key features of early development shape deep convective systems

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

[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
[*]sophie.abramian@gmail.com

ABSTRACT

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 simulation SAM from the DYAMOND project, combined with machine learning tools, present an unprecedented opportunity to address this inquiry. Leveraging these techniques, our study aims to quantitatively assess the influence of different variables — either internal or external — for system cloud shield size at maturity. Utilizing only the systems' growth rate of area within their first 1.5 hour of development to predict the mature cloud shield yields an estimate with a Pearson correlation coefficient of 0.5. When only considering additional features related to the system and to its environment, the accuracy increases to 0.65. Key factors at play include the presence of ice in the system, the long-wave emission by the system, the migration distance and initial values of vertical velocity.

1 Introduction

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 there are three fundamental ingredients: moisture, instability, and a lifting mechanism[11]. Humidity supplies the water necessary for cloud formation, while 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 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 and global modeling efforts have largely concentrated on how large-scale environmental factors influence the behavior of deep convective systems (DCS)[12,13]. Among these efforts, a recent study[14] 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[15,16]. 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[17]. 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. 1). 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, deep convective systems 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 hours (D) with a standard deviation of 2.7 hours. This consistent life cycle is 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 itself can be described by a simplified mass balance equation[18]

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

where A represents the cloud shield area, A_c is the convective area, and the subscript "src" denotes the temporal generation of new convective area. 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.

Recent advances have enabled more comprehensive investigations of DCS life cycles, allowing for a more holistic study of both internal and external influences. High-resolution global simulation using the SAM model within the DYAMOND project[19–21], combined with a sophisticated storm tracking method called TOOCAN[22], 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 deep convective system properties with observations[23] indicates that global SAM is well representative of the current global CRM generation's ability to represent organized deep convective systems. Although our analysis focuses exclusively on SAM, these findings are likely applicable to other GCRMs 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[23] (for a comparison of DCSs in the DYAMOND simulations and in satellite observations, see Supporting Information Text §.3 and Fig. S8).

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 deep convective system (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, deep convective systems (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 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 the 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 deep convective systems 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 following section. The subsequent three sections address each of the key questions in turn.

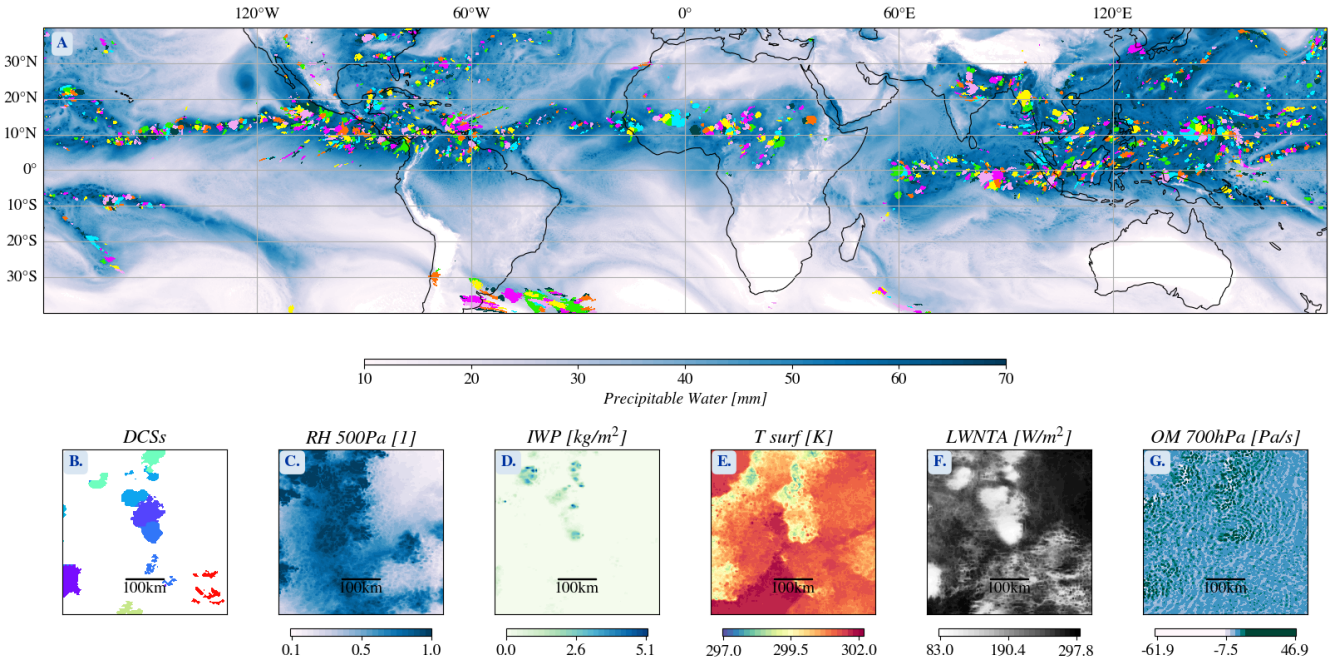

Figure 1. (A) Snapshot from the DYAMOND-SAM summer simulation during August 23 2016 at 23h30 of the Precipitable Water with on top the Deep Convective Systems tracked by TOOCAN focusing only on the tropics. bottom. Example of considered physical fields for a given DCS at 1h of development : (B) the system 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).

2 Results

2.1 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 ($\mathscr{L}_{max} = \sqrt{A_{max}}$ where A_{max} denotes the maximum area of the DCS, see Materials and Methods). As mentioned in the introduction, the average maximal extension is 115km with a standard deviation equal to 45 km. Fig. 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 minutes (frequency of outputs from the DYAMOND-SAM simulation for the two-dimensional variables used here) up to 5 hours. Firstly, we can see that increasing the observation period improves the models performance, such that after 5 hours, they can predict the final extension of the system with an average accuracy of 10 to 15 kilometers. However, since the systems on average last 7.5 hours and reach their maximum area after around 3 hours, beyond this point, the task becomes too easy, and the models detect it effortlessly. Secondly, with 1.5 hour of observation, all three models predict the maximum system size with a score of approximately 0.5 and an average error of about 35 kilometers. 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[24–26]. This raises questions about whether the simple growth rate theoretical model described in the introduction[18] 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.

2.2 Prediction with physical features

In this section, we aim to predict the maximum size of deep convective systems (DCSs) based on physical features observed during their early development, focusing on identifying the key drivers of this predictability. The growth of convection is shaped by both internal feedback mechanisms and external environmental factors, with critical influences such as humidity,

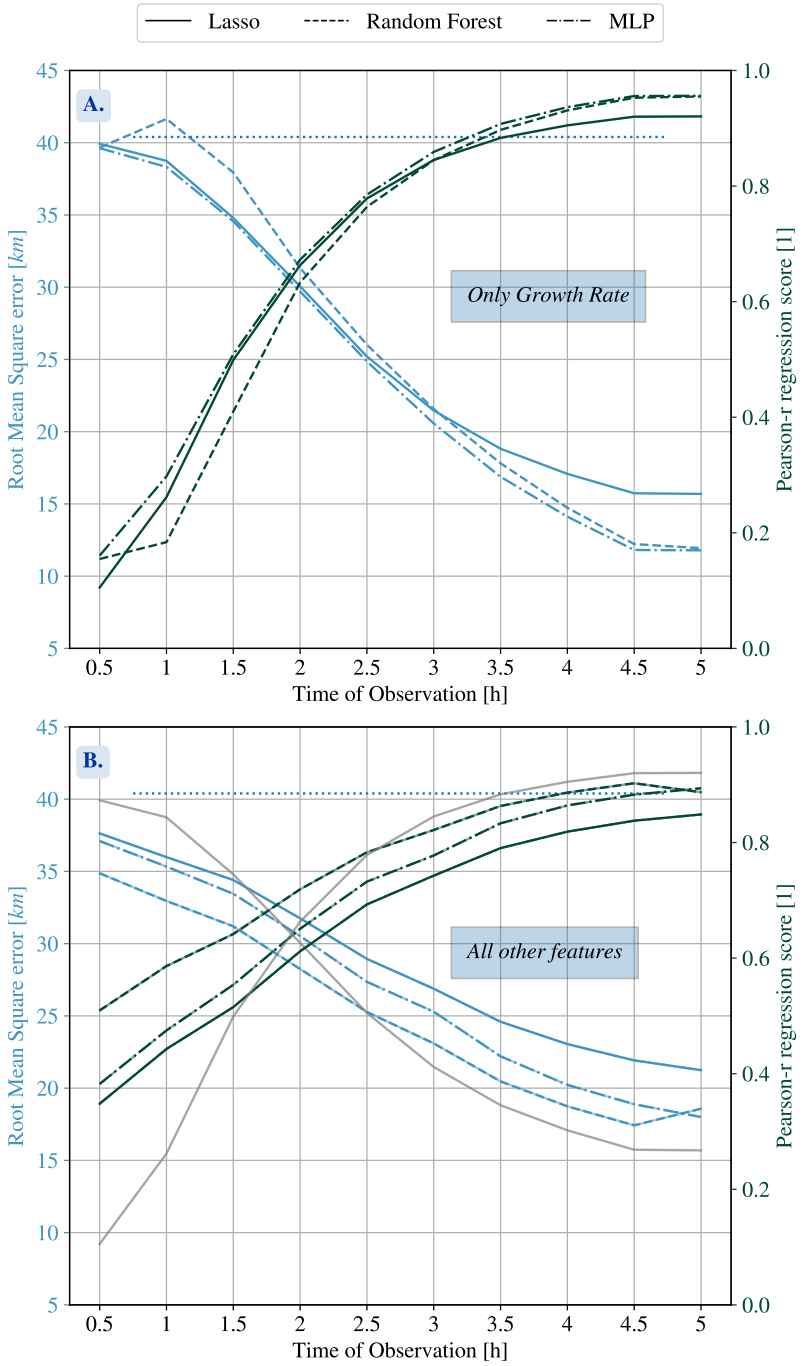

Figure 2. (A) Evolution of Pearson-r regression score in blue and mean square error in green for the estimate of maximum extension $\mathscr{L}_{max}$ in the three machine learning models — Linear Regularized Regression in solid line, Multi Layer Perceptron in dash-dotted line and Random Forest in dashed line — trained on the evolution of the growth rate during a considered observed period. (B) Same with additional features added, 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 grey to ease comparison (ascending is pearson-r, descending is rmse).

atmospheric instability, and vertical lifting processes. For our analysis, we selected a comprehensive set of physical features from kilometer-scale SAM-Dyamond simulations, as described in the next paragraph.

In SAM-Dyamond, 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 surface 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. In addition to these primary variables, supplementary 2D fields from the SAM-Dyamond simulations, as outlined in table 1, are incorporated to enhance the predictive capability of the model through 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. This dual approach provides a comprehensive view of the environmental context 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 or enhance lifting through gravity wave interactions[27]. 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 (see Supplementary Fig. S2 for a detailed overview of the computation process for the input variables).

Long Name Variable	**2D or 3D**
Longwave Net Radiation at Top of Atmosphere	2D
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 10m	2D
V-component of Wind at 10m	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

Table 1. Physical fields used from SAM-Dyamond in the second experiment. The '2D or 3D' column indicates whether the variables are directly available from 2D outputs or derived from 3D data outputs.

Features related to the surrounding DCSs
Number of Surrounding Systems
Average and Maximal Age of Surrounding Systems
Average and Maximal Size of Surrounding Systems
Average and Maximal Distance of Surrounding Systems
Average and Maximal Distance of Surrounding Systems Weighted by Their Size

Table 2. List of Features Related to Surrounding Systems

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 minutes up to 5 hours. 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 minutes, and after one hour, we achieve a R-Squared index coefficient of 0.5 for the multi-linear and random forest models and 0.6 for the neural network. With 1.5 hour of observation, we reach a score of 0.65 for the neural network model and an average error of less than 35 km. We see that at this point, the linear model trained only on the growth rate provides better performance than the linear one trained on the other features. With 2 hours of observations, the score reaches 0.75 for the neural network and the error decreases to less than 30 km. In contrast to the previous experiment, the MLP model consistently outperforms the multi-linear, maintaining a score approximately 0.1 higher in R-Squared. This emphasizes the non-linear relationship between physical factors and system size. We note in passing that the accuracy of the prediction is not greatly influenced the duration of systems (see Supplementary Fig. S3).

To identify which systems are most sensitive to the inclusion of new features, we analyze how predictions from the multilinear model, which includes all features, vary across different DCS sizes. Although other models offer better predictive performance (see Supplementary Fig. S3A-D and Fig. S4A-D), we focus on the multilinear model for its clearer interpretability of individual features. While it may not fully capture the complexity of the feature-DCS size relationship, it still provides valuable insights into key variables and processes. 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 hour and 2 hour 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. Fig. 3A shows results where all other features during the first 1.5 hour are provided for the training of the model, and Fig. 3B shows the same for 2 hours. We can see that adding times of observation has led to an improvement in prediction for both small systems (smaller than $\sim$ 100 km) and larger systems (larger than $\sim$ 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).

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 minutes of observation, the error is reduced by about 5 km.

The regression scores are found to be higher, and thus the predictions are more accurate, for systems smaller than 100-120 km (not shown). For larger systems, even with an observation period of 2 hours, the regression score falls below 0.5. This suggests that additional mechanisms, besides the initial growth, are at play in the evolution of large systems. Given that large DCSs are associated with the largest impacts (extreme weather, larger cloud shields), this deserves further investigation. As a first step in this direction, in the next section we use our results to investigate the key variables that determine the growth of systems in our predictions. Of particular interest is whether the environment, or the characteristics of the system itself, are more influential in this prediction.

2.3 Feature importance and dimensionality reduction

Focusing solely on the 1.5-hour prediction, we now seek to identify the features that contribute to predicting the maximum extension. Since all three models exhibit similar performance, we focus here on the linear and random forest models, which are easier to interpret, and investigate whether they rely on the same variables or not. To that end, we determine the important features for each model. For the linear model, we have the list of coefficients assigned to each variable. For the random forest, we can retrieve the average Gini index scores over all decision trees for each variable. Fig. 4A shows the top 15 important variables for the linear model.

Although they do not appear in exactly the same order, we find the same first variables of importance for both models (see Supplementary Fig. S5A,B), which leads us to focus only on the multi-linear model (both models exhibit equivalent performance but this one is explicit). Among these variables, we identify the standard deviation of the ice water path (IWP) both within and outside the system, the migration distance, the mean long-wave radiation emitted by the system, the eccentricity of the DCS core, the mean IWP, the standard deviation of integrated moist static energy (MSE) within the system, land mask variance, the mean vertical velocity at 850 hPa and 700 hPa, the mean long-wave radiation emitted by all DCSs in the environment, relative humidity at 500 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); for example, a strong correlation is observed between the mean of IWP and the mean of the long-wave radiation across the entire domain.

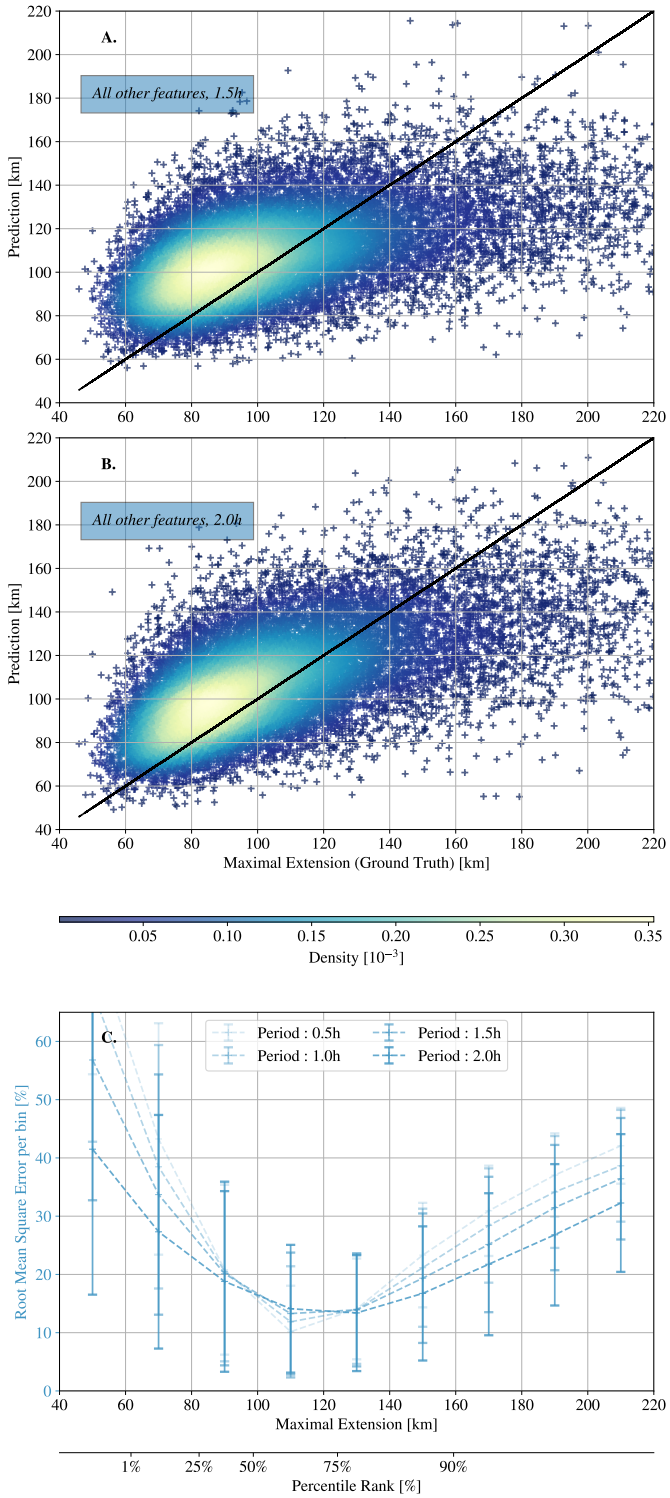

Figure 3. One-to-one diagrams for the prediction and the target for the multi-linear model trained with all features, including shape, physical fields, migration distance and neighboring systems influence, based on (A) 1.5 hour of observation and (B) 2 hour of observation. The color of the scatters represents 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.

Interestingly, we find that key microphysical, dynamic, and thermodynamic factors are well represented in this set. The importance of microphysics, particularly ice loading, has recently been emphasized as a critical determinant in convective system behavior[10], and it continues to be a leading factor in predicting system dynamics. Although the impact of surrounding neighbors may be captured by the LWNTA over the domain, none of the computed features were selected, indicating that interactions between DCSs are not well described by their number, age, or size. Finally, the dynamical variables seem to be crucial in forecasting the future of the system, a capability that is highlighted by the use of high-resolution global models, which provide dynamical fields not available from observational data.

Having identified the key predictors of DCS size, we now 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:

$$\begin{aligned}\mathscr{L}_{max} &= \mathscr{L}_0 + \ell(f_1, f_2, ...f_n) \sim \mathscr{L}_0 + \ell(f_1, f_2, ...f_{15}) &(2)\\ &\sim \mathscr{L}_0 + c_1 f_1 + c_2 f_2 + ... + c_{15} f_{15}, &(3)\end{aligned}$$

where $\mathscr{L}_{max}$ denotes the maximum extension of the system, $\mathscr{L}_0$ 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:

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

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, see Supplementary Fig. S7 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[28]. Applying the variance -a non-linear operator- to the equation [4], it writes,

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

where V denotes the variance over the systems, and COV the co-variance, also along the system set. Fig. 5A shows the three contributions of the right-hand side as percentages of the total variance explained for all DCSs (from now on, we focus on the prediction using all features during the first 1.5 hour 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, particularly those related to growth rate and the presence of ice with play a predominant role, accounting for 40.5% of the variance in maximal extension across all systems. Environmental factors, while also significant (26.3%), are primarily driven by the presence of ice in neighboring systems, highlighting the importance of surroundin

Fig. 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 deep convective systems (DCSs) reaching mesoscale dimensions develop complex internal circulations that sustain convection and moisture inflow[29]. Our analysis confirms that internal processes increasingly dominate in larger systems, not least due to the greater influence of growth rate (not shown). 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, a significant fraction of the variability in their maximal extension is explained by their initial environment; 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 50% at 80 km (not shown) and keeps increasing as the size of systems does. These results are based on the first 1.5 hour. 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.

3 Discussion

In our study, we investigated the relationship between deep convective systems 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. When other features—such as shape, physical fields, trajectory, and the influence of surrounding systems—were incorporated into a neural network model, the regression score improved to 0.70, and even with just one hour of data, the score reached 0.6. In contrast, the linear model consistently yielded lower performance, indicating that the underlying relationship between these variables is nonlinear. This enhanced predictive accuracy therefore led us to identify key factors influencing DCS size, offering new insights into the roles of both environmental and internal processes.

Several noteworthy factors emerged as significant contributors to prediction accuracy, including the standard deviation of the ice water path (IWP) in the system and outside of the system, the migration distance, the mean long-wave emitted solely by the system, the eccentricity of the core of the DCS, the mean of IWP, the standard deviation of the integrated moist static energy (MSE) within the system, land mask and dynamical fields including the initial deep shear. The predictors identified align with established theories regarding growth rate and the relationship between convective strength and system size, and highlights the potential for further research to extend these results beyond the GCRM framework, possibly using satellite data[30]. However, adapting these methods to observational datasets may introduce uncertainties, particularly due to the absence of key dynamical fields in satellite data, which seems to play a critical role in the accurate prediction of system development. In the longer term, the feature selection process could be refined and adapted for observational datasets through the application of causal feature selection methods. These approaches, which identify cause-and-effect relationships from data, could help pinpoint the key drivers directly linked to the forecasted variables, ultimately improving both the robustness and interpretability of the models[31].

This study is unique in its combination of several novel tools, including high-resolution global simulations, a Lagrangian approach to deep convective systems, and machine learning methods. Consequently, there have been few similar studies conducted in the literature, raising the question of whether the results described above are satisfactory or if better or worse outcomes could have been expected.

Previously, the hypothesis that the mature state of the system is strongly correlated with its initial growth rate has been documented in the literature without proposing a quantifiable score or relationship[24–26]. Furthermore, a recent study[10] has precisely highlighted how ice loading is strongly influenced by changes in convection velocities. The use of machine learning models, with no a priori assumptions on the existing relationship between growth rate and maximum extension, nor on the role of ice load, appears to support these hypotheses. However, even with the addition of numerous variables, the predictive score after one hour of observation ranges from 0.5 to 0.6 (respectively for the multi-linear and neural network models), indicating that not all factors contributing to the relationship have been captured.

While the initial conditions of the system seem to play a significant role, contrasting them with the system's boundary conditions, i.e., an event occurring during the growth phase unrelated to the initial state, could provide valuable insights. Efforts were made to correlate the prediction error with variables such as distance traveled, presence of a coastline, or final system lifespan, but these attempts yielded no conclusive signals. Further investigation into this matter would be beneficial.

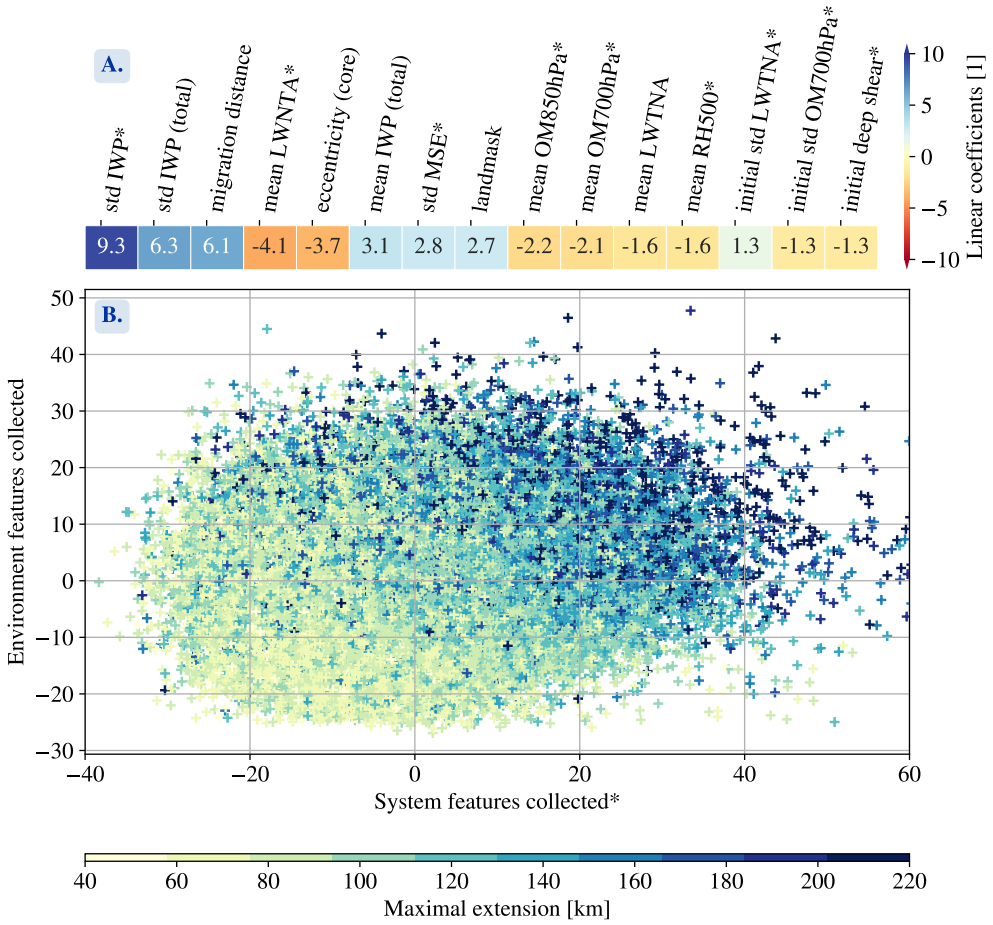

Figure 4. (A) Coefficients optimized for the multi-linear model to solve the supervised task of predicting the maximal extension based on the first 1.5h evolution of DCS growth rate and development features. The coefficients (denoted c_i in [4]) are sorted with their absolute value. There are 15 features 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).

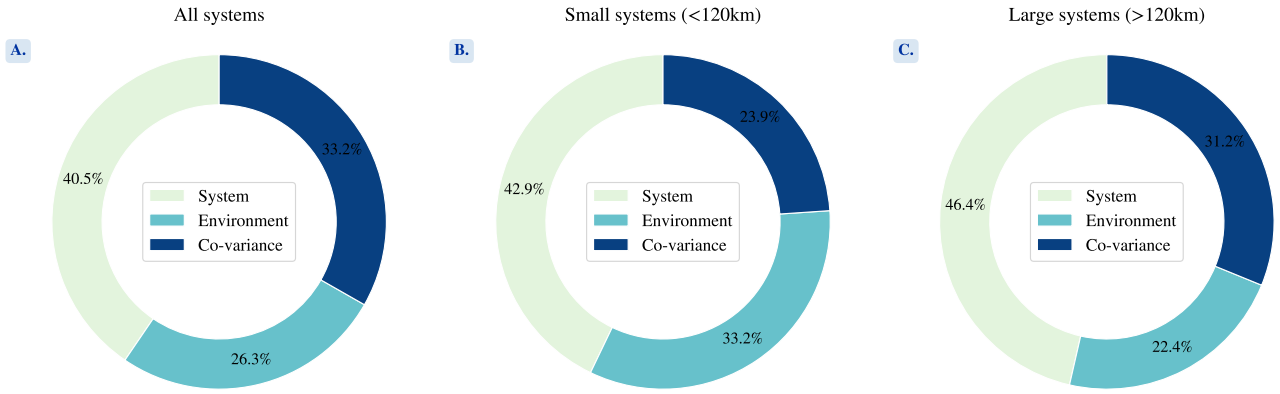

Figure 5. Fraction of the total variance of the DCS maximal extension prediction explained by the system or environmental conditions for (A) all systems, (B) systems smaller than 120 km and (C) larger than 120 km.

One interesting finding is that features based on both the DCS and its environment contribute to the prediction, showing that a given initial environment does not necessarily constrain the size of a system. This is all the more true for large systems, whose fates are more largely determined by their internal characteristics. This is consistent with the well-documented internal circulation that DCSs can develop, allowing them to grow to mesoscales[29]. Between the initial and mature stage, the system and its environment change, as they are impacted by the system[24]. Understanding how the system itself evolves, and how it feeds back on its environment, are thus promising avenues to improve our fundamental understanding of systems growth. This observation is particularly intriguing when considering systems developing within a cohort of deep convective systems, where CAPE (Convective Available Potential Energy) 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.

4 Methods

4.1 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 deep convective systems (DCSs), we rely on high-resolution global simulation SAM (part of the DYAMOND project[20]). 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. For our study, we used the System for Atmospheric Model (SAM,[19]), an anelastic model of fluid dynamics with parametrized microphysics.

Most DYAMOND models, including SAM, accurately capture essential DCS characteristics such as lifetime, cloud shield area, and volume of rainfall[32]. 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[33]. To detect, track, and measure the evolution of DCSs, we leverage the capabilities of the cloud tracking algorithm TOOCAN[22], which relies on a definition of a deep convective system consisting in a combination of a convective core, characterized by low brightness temperature, associated to an anvil cloud, characterized by relatively higher brightness temperature. These components evolve over time, and the TOOCAN algorithm aims to link 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, to partition the high cold cloud shield into DCS components. The deep convective systems identified through this process exhibit a broad spectrum of convective organization, from short-lived, small and isolated systems, to long-lived and large systems[34], 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. In the following paragraphs, we describe the

pre-processing steps applied to the data.

4.2 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 will focus solely on tropical DCSs, restricting our analysis to within $\pm$ 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 hours and a maximum 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. S9, and although they last for more than 5 hours, 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. S9A for an illustration). 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. S9B, 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. 1A for a few DCSs, alongside the joint distribution of lifespan and maximum extension shown in Fig. 1B. 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?

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

Handling all this data is challenging, especially considering that the dynamic and thermodynamic fields are 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). 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.

Machine Leaning 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[35].

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 minutes up to 5 hours (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:

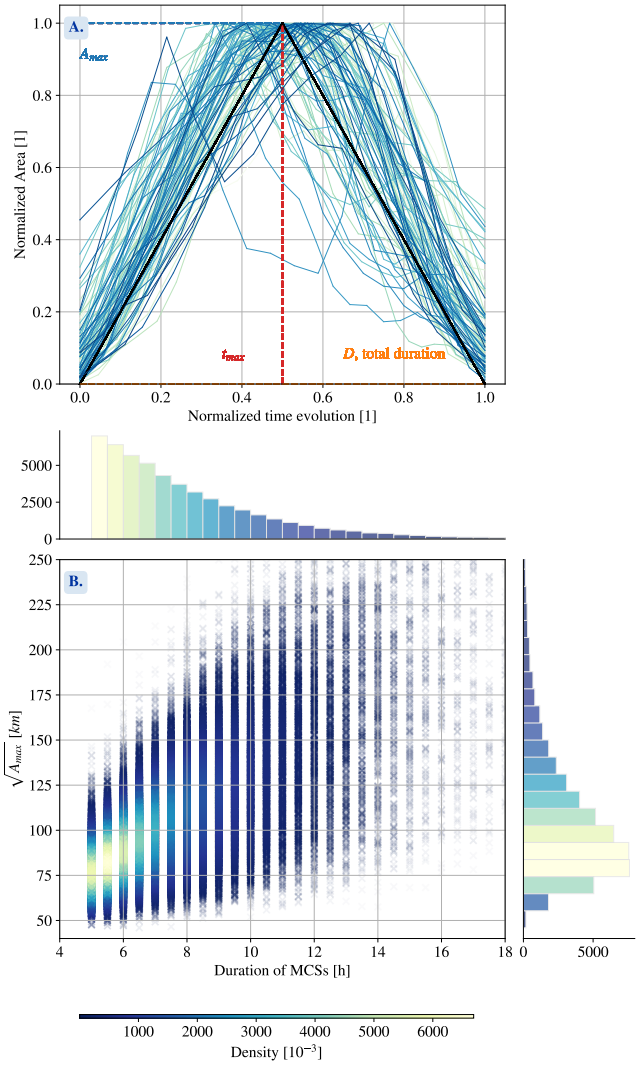

Figure 1. (A) Superimposed evolution of the life cycle of tracked Deep Convective Systems (arbitrarily chosen) - i.e. the normalized area evolution as a function of the normalized time for 100 systems. The solid black line represent the simple theoretical model proposed in[14] which captures the life 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.

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

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 hours 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 hours. We note in passing that in the theoretical model described in Fig. 1 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 phase, we consider new features into the model through a feature engineering approach. These features are categorized into four main groups: cloud morphology, surrounding physical fields in the local environment, system migration, and interactions with neighboring deep convective systems (DCSs). Concerning the 3D variables, we computed shear and deep shear in both the longitudinal and latitudinal directions. Shear is defined as the difference between wind velocities at 900 hPa and 10 m, while deep shear refers to the difference between wind velocities at 400 hPa and the surface. Additionally, we computed moist static energy (MSE) as a proxy for convective available potential energy (CAPE) and convective inhibition (CIN), evaluating the MSE difference between the mid-troposphere (900 hPa to 400 hPa) and the boundary layer (400 hPa to the surface). Given that shear and CAPE are pre-storm variables linked to storm initiation, these variables are only considered in the three hours prior to storm formation, consistent with the available time resolution.

Alongside these 3D variables, we selected 12 key 2D variables: long-wave net radiation at the top of the atmosphere, precipitable water, relative humidity at 700 hPa and 500 hPa, surface temperature at 2 m, ice water path, meridional and zonal wind velocities at 10 m, vertical wind velocities at 850 hPa, 700 hPa, and 500 hPa, and the land mask.

Bottom panels of Fig. 1, as well as supplementary examples in supplementary Fig. S2, show examples of physical fields for a given DCS at 1 hour of development. For each variable, we calculated the mean and standard deviation over a 5-degree by 5-degree window centered around the system's barycenter. Additionally, we computed the mean and standard deviation just beneath the system (by applying the mask) and only outside the system (by multiplying with the complement of the mask). The aim was to distinguish the contributions from the system and its environment (see dataset description in WDCC publication, in Data archival section).

For the interaction with neighboring DCSs, we defined both an average influence factor and a maximum influence factor. To compute these, we again used a 5-degree by 5-degree window centered around the system. We weighted the influence of each neighboring system size (see Fig. 1B) by its distance to the center (using $\exp(-(d/d_0)^2)$, where d is the distance to the center, and $d_0 = 50$ km), and then averaged them to obtain the average influence factor. For the maximum influence factor, we took the maximum value instead of the average.

At each time step of the system's evolution (every 30 minutes), 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%. 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. As in the first experiment, we vary the observation time window used for training.

References

1. Stephens, G. *et al.* The First 30 years of GEWEX. *Bull. Am. Meteorol. Soc.* **104**, 126–157, DOI: 10.1175/bams-d-22-0061.1 (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. Atmospheric 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 advances* **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. Atmospheric 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. Yang, Q., Leung, L. R., Feng, Z. & Chen, X. A moist potential vorticity model for midlatitude long-lived mesoscale convective systems over land. *J. Atmospheric Sci.* **80**, 2399–2418 (2023).
13. 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).
14. 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).
15. Wing, A., Emanuel, K., Holloway, C. & Muller, C. Convective self-aggregation in numerical simulations: A review. *Surv. Geophys.* **38** (2017).
16. Wing, A. A. Self-aggregation of deep convection and its implications for climate. *Curr. Clim. Chang. Rep.* **5**, 1–11 (2019).
17. Muller, C. *et al.* Spontaneous aggregation of convective storms. *Annu. Rev. Fluid Mech.* **54**, 133–157 (2022).
18. 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. Atmospheres* **127**, e2021JD035599 (2022).
19. 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).
20. 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).

21. 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).

22. Fiolleau, T. & Roca, R. An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite. *IEEE transactions on Geosci. Remote. Sens.* **51**, 4302–4315 (2013).

23. Feng, Z. *et al.* Mesoscale convective systems tracking method intercomparison (mcsmip): Application to dyamond global km-scale simulations. *Authorea Prepr.* (2024).

24. 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).

25. Fritsch, J. & Forbes, G. Mesoscale convective systems. In *Severe convective storms*, 323–357 (Springer, 2001).

26. McAnelly, R. L. & Cotton, W. R. Meso-β-scale characteristics of an episode of meso-α-scale convective complexes. *Mon. weather review* **114**, 1740–1770 (1986).

27. Mapes, B. E. Gregarious tropical convection. *J. Atmos. Sci.* **50**, 2026–2037 (1993).

28. 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.* **113**, 7426–7431 (2016).

29. Houze, R. A., Jr. Mesoscale convective systems. *Rev. Geophys.* **42**, RG4003 (2004).

30. 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).

31. Beucler, T. *et al.* Selecting robust features for machine-learning applications using multidata causal discovery. *Environ. Data Sci.* **2**, e27 (2023).

32. Feng, Z. *et al.* Mesoscale convective systems in dyamond global convection-permitting simulations. *Geophys. Res. Lett.* **50**, e2022GL102603 (2023).

33. Abramian, S., Muller, C. & Risi, C. Shear-convection interactions and orientation of tropical squall lines. *Geophys. Res. Lett.* **49**, e2021GL095184 (2022).

34. 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, DOI: 10.5194/essd-2024-36 (2024).

35. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

CJM and SA 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 programme 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 statement

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

Additional information

Accession codes This study utilizes the global cloud-resolving model SAM from the DYAMOND project, which is stored alongside other GCRMs 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)

Competing interests The authors declare no competing interests. The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper.