Supporting information

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

Further comparison between Multi-linear model and Random Forest

On Figure 4 we can see four one-to-one diagrams of prediction and ground truth for the multi-linear model (top) and for the random forest (bottom) for an observation period equal to 1 hr (left column) and 1.5 hrs (right column). Both models overestimate small systems and underestimate large ones, consistent with a strategy to use the mean over systems as a prediction (see also Figure 5). They both fail in providing a good prediction. However, with an extra 30 minutes of data, we observed that both models improve their prediction for systems smaller than 100 km : the scatter is sharper and follows the $y = x$ function. The multi linear model tends to overestimate system sizes with a systematic biais of 10-15 km for these kind of systems (smaller than 100 km). Random forest prediction for 1.5 hrs of observation shows lower performance for small systems but higher performance for large systems.

Figure 5 displays the mean (dashed line) and deviation (error bar) of the mean-square error as a function of the maximal extension of DCSs, for the multi-linear model (lasso, on the top) and the random forest (bottom) with the prediction based only on the area growth rate (left column) and with all features (right column). The black line is the mean-square error value for a constant prediction equal to the mean of the dataset, and it has the typical shape of a 'v'. For 30 min (blue curve), we see that the prediction for both models follows closely the 'v-shape', assessing that models simply predict the average value. At 1 hr (turquoise curve) the error is smaller but is still large (40%) for small (below 60 km) and very large (beyond 160 km) systems. At 1.5 hrs (orange curve), the error of the model is under the 'v-shaped' curve and is flattened. The error is reduced by 15% for very small systems (smaller than 60 km) and reduced by 5% for systems as large as 140 km. For 2 hrs (red curve), the scatter is flattened near below 20%, supporting that the model has learned how to predict the maximal extension. We observe that the mean-square error is increasing for larger systems, as well as the standard deviation, which may be partly due to the fact that there are less systems in this range (and thus less systems to learn from).

Evolution of the prediction error as a function of system duration

Can this method — which relies on physical features and machine learning models to predict system size — effectively operate on short-lived systems? Here, we show evidence supporting that the method is able to predict the maximal size of DCSs even for longer ones. The progression of the the root-mean-square-error (RMSE) versus the systems' lifespan is illustrated in figure 3, for the multilinear model trained for either 1 hr (panel A) or 1.5 hrs (panel B). Mean values for each duration are indicated by dashed lines, accompanied by error bars representing the associated standard deviation. In panel A, the RMSE is approximately 17% for systems lasting 7.5 hours, reaching 20% for those lasting 12.5 hours. To further support this, in panel B, precision improves for systems with a 7.5-hour duration, with a RMSE below 15%, while systems lasting 12.5 hours are predicted with an average error of 20%. The graphical representation highlights a weak correlation between model skill and system duration.

Tracking algorithm calibration

Prior to applying the TOOCAN algorithm to the SAM outputs, it is required to convert the OLR of the SAM model to equivalent brightness temperatures (BT_{equi}). To do this, we compute a relationship between OLR and brightness temperature (BT) from the Scanner for Radiation Budget (ScaRaB) radiometer onboard the Megha-Tropiques satellite over the entire 2012 period. The ScaRaB instrument enables simultaneous observations of the OLR at the Top Of the Atmosphere (TOA) and infrared thermal measurements within the 10.5-12.5 μm window channel. We took advantage of this capability to compute a regression equation estimating an equivalent IR brightness temperature (BT_{equi}) from the OLR at TOA. A flux equivalent brightness temperature (BT_f) is first determined from OLR by:

$$OLR = \sigma BT_f^4. \tag{1}$$

Then, the derived equation regression between BT_{equi} and BT_f is of the form:

$$BT_{equi} = aBT_f^2 + bBT_f + c \tag{2}$$

where $a = 0.0035215$, $b = -0.146688$ and $c = 86.6780$.

The residual bias between BT_{equi} and the observed BT is ~ 0.16 K for a $BT_f < 245$ K, and is of the same order of magnitude as the residual bias calculated with the method proposed by[1] to estimate BT from OLR. We then perform then the detection and tracking of individual DCSs by applying the TOOCAN storm tracking algorithm on the simulated BT_{equi} of the SAM simulation for a 40°S-40°N latitudinal band over the 1 August–10 September 2016 period. To address the cyclic feature of longitudes and guarantee continuity of DCS tracking across the 0° and 360° meridians, a specific preprocessing step is undertaken for each IR image. This involves duplicating a 20° longitude band from 340° to 360° and appending it adjacent to the 0° longitude, and similarly, a 20° longitude band from 0° to 20° is duplicated and placed alongside the 360° meridian. By extending the IR data in this manner, the cloud tracking algorithm can identify and track DCSs which propagate across the 0° and 360° meridians. DCSs identified twice, once on each side of the 0° and 360° meridians, are removed from the resulting deep convective systems database. Only DCSs occurring between 10 August and 10 September 2016 are included in our analysis to avoid the model spin-up period, ensuring the reliability of our results[2,3]. Also, we focus solely on tropical DCSs, restricting our analysis to within +-30° of latitude.

Assessing realistic properties of DCSs in Dyamond-SAM simulation

Fig. 9 A and B show the distributions of DCS lifetime duration and maximum area (Amax) simulated by the SAM model over the tropical belt (+-30° of latitude), and observed by a fleet of geostationary satellites for similar period and region[4]. First, we can notice than the simulated DCSs are more numerous ($\sim$ 184000) than the observed ones ($\sim$ 148000). The lifetime duration distributions, as illustrated in Fig. 9a, are characterized by a mode at 5 hours with quite similar occurrence for both simulated ($\sim$ 43000) and observed DCSs ($\sim$ 47000). Following this, there is a decline in the frequency of occurrence for both distributions. The slope of the distribution for observed DCSs is more pronounced, indicating these systems exhibit shorter durations. The observed systems have a maximum lifetime duration of $\sim$ 41 hrs, while it is approximately 48 hrs for simulated DCSs. The distributions depicting the maximum area reached by the DCSs throughout their life cycles (Fig. 9B) demonstrate a substantial similarity in the cloud surface between the observed and simulated DCS events. Observed DCSs are slightly larger ($\sim$ 720000 km²) than the simulated ones ($\sim$ 360000 km²). Fig. 9C and D illustrate the two-dimensional distribution of both observed and simulated DCSs, as a function of their maximum spatial area and the duration of their lifetime. This phase diagram can be seen as a measure of the organization of deep convective systems[4]. The analysis shows that, similarly to observed DCSs, simulated DCSs exhibit a broad spectrum of organizational structures, ranging from short-lived systems spanning a few thousand km^2 to long-lived systems covering several thousand square kilometers. There is an apparently strong relationship between lifetime duration and the maximum spatial area. However, for both simulated and observed DCSs, a more detailed analysis indicates that, for any specific lifetime duration, the maximum spatial area can vary significantly, covering a spectrum that spans several orders of magnitude. This dispersion appears more significant for the simulated systems. Overall, the DCS events simulated by the SAM model exhibit life cycles that appear realistic and comparable to those observed. The distribution of maximum area aligns closely with observational data. However, the simulated lifetime durations tend to be overestimated.

A recent study[5] conducted a detailed analysis of tropical mesoscale convective system (MCS) characteristics in the DYAMOND models, including DYAMOND-SAM, across both summer and winter phases. Eight feature-tracking algorithms, including TOOCAN, were applied to both simulations and satellite observations, aiming to identify biases in the global convection-resolving models (GCRMs) and the tracking methods used. Results show that the probability density function (PDF) of simulated Brightness Temperature (Tb) from most DYAMOND models agrees reasonably well with observations, with no significant deviation observed in DYAMOND-SAM. However, simulated rain rates exhibit greater spread across models than Tb, with most models overestimating the frequency of heavy rain rates (10–40 mm h^{-1}) while underestimating light rain rates. This variability likely affects MCS identification, and most models (13 out of 19), including SAM, show higher precipitation-to-precipitable water (PW) ratios than observations, with biases ranging from approximately 10% to 33% (SAM's bias is 19% higher). These findings suggest that most DYAMOND models tend to overestimate the sensitivity of precipitation to environmental moisture over tropical oceans. Despite these biases, DYAMOND-SAM appears reasonably consistent with observational data overall. Additionally, the study assessed the models' ability to simulate the cloud and precipitation characteristics of MCSs throughout their lifecycle. Lifecycle composite analysis for MCSs with median lifetimes in each model showed that most DYAMOND models captured the canonical MCS lifecycle, though the magnitude and rate of change varied among the trackers. Model-observation differences in cloud shield properties were smallest with the TOOCAN tracker compared to others.

References

1. Ohring, G., Gruber, A. & Ellingson, R. Satellite determinations of the relationship between total longwave radiation flux and infrared window radiance. *J. Appl. Meteorol. Climatol.* **23**, 416 – 425, DOI: 10.1175/1520-0450(1984)023<0416:

SDOTRB>2.0.CO;2 (1984).

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

3. 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).

4. 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).

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

Name and meaning of features

We use the following abbreviations:

For each variable we compute the mean and the standard deviation and use this abbreviation:

- mean: **mn**
- standard deviation: **std**

Mean and standard deviation are compute in four different manners with the assiciated abbreviation and meaning:

- total subdomain: **ts** - the domain is $5\,\mathrm{deg} \times 5\,\mathrm{deg}$
- under cloud: **uc**
- neighbor systems: **ns** - the variable is averaged only below other systems
- except the system: **es** - the variable is averaged everywhere in the subdomain except below the considered system

Format for Variable Names

- **mn_uc_vX_tY**: Mean under cloud for variable X at timestep Y
- **std_uc_vX_tY**: Standard deviation under cloud for variable X at timestep Y
- **mn_ns_vX_tY**: Mean for neighbor systems for variable X at timestep Y
- **std_ns_vX_tY**: Standard deviation for neighbor systems for variable X at timestep Y
- **mn_es_vX_tY**: Mean except the system for variable X at timestep Y
- **std_es_vX_tY**: Standard deviation except the system for variable X at timestep Y
- **mn_ts_vX_tY**: Mean for total subdomain for variable X at timestep Y
- **std_ts_vX_tY**: Standard deviation for total subdomain for variable X at timestep Y

For the features relative to the system's neighbor, we use the following abbreviations

- **nb_neighbours**: Number of neighbors
- **number_of_active_neighbour**: Number of active neighbors (growth phase)
- **mean_position_neighbour**: Mean distance of neighbors
- **min_position_neighbour**: Minimum distance of neighbors
- **average_age_neighbour**: Average "age" of neighbors
- **max_age_neighbour**: Maximum "age" of neighbors
- **long_lived_neighbour_detected**: Long-lived neighbor detection (greater than 5 hrs)
- **mean_interaction_power**: Mean interaction power - mean distance of neighbor system weighted by their size
- **max_interaction_power**: Maximum interaction power - maximum distance of neighbor system weighted by their size

Format for Variable Names
<abbreviation>_time_Y: Variable description at timestep Y

Index	Variable Name	Long Name Variable	Physical Meaning
2	PW	Precipitable Water	Total column-integrated water vapor
3	RH500	Relative Humidity at 500 hPa	Moisture content at mid-troposphere (500 hPa level)
4	RH700	Relative Humidity at 700 hPa	Moisture content at lower mid-troposphere (700 hPa level)
5	T2m	2m-Temperature	Temperature near Earth's surface (2m above ground)
6	IWP	Ice Water Path	Column-integrated mass of ice particles
7	U10m	U-component of Wind at 10m	Zonal wind speed (east-west) near the surface
8	V10m	V-component of Wind at 10m	Meridional wind speed (north-south) near the surface
9	LANDMASK	Land Mask	Binary indicator for land vs ocean
10	OM500	Omega at 500 hPa (Vertical Velocity)	Vertical motion in mid-troposphere (500 hPa level)
11	OM700	Omega at 700 hPa (Vertical Velocity)	Vertical motion at lower mid-troposphere (700 hPa level)
12	OM850	Omega at 850 hPa (Vertical Velocity)	Vertical motion near boundary layer (850 hPa level)
13	SHEAR	Wind Shear	Difference in wind speed between 1000m and surface
14	DEEPSHEAR	Deep Wind Shear	Difference in U-wind speed between 6000m and surface
15	SHEARV	Wind Shear in V	Difference in V-wind speed between 1000m and surface
16	DEEPSHEARV	Deep Wind Shear in V	Difference in V-wind speed between 6000m and surface
17	INT_FMSE_BL	Integrated Moist Static Energy within Boundary Layer	Energy related to moisture and temperature at boundary layer
18	INT_FMSE_MID	Integrated Moist Static Energy within Mid-Troposphere Layer	Energy related to moisture and temperature at mid-troposphere
19	DIFF_FMSE _MID_BL	Difference in Integrated Moist Static Energy between Mid-Troposphere and Boundary Layer	Instability between mid-troposphere and boundary layer

Table 1. Variable Index, Variable Name, Long Name Variable, and Physical Meaning

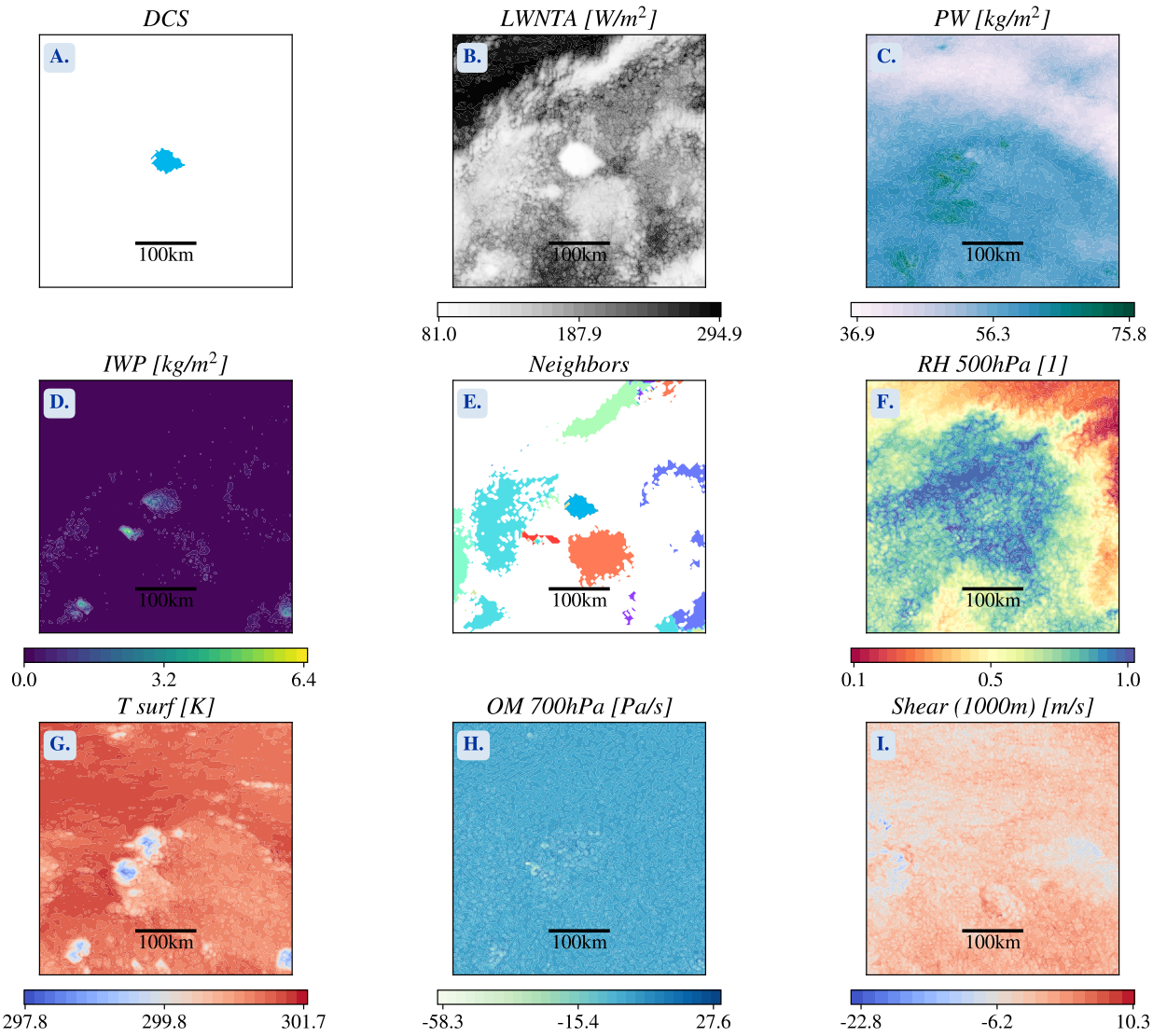

Figure 1. Example of considered physical fields for a given DCS at 1 hr of development : A. the system alone. B. its neighbors. C. long wave top of atmosphere. D. precipitable water. E. ice water path. F. relative humidity at 500 hPa. G. surface temperature. H. vertical velocity at 700 hPa. I. Wind Shear at 1000m

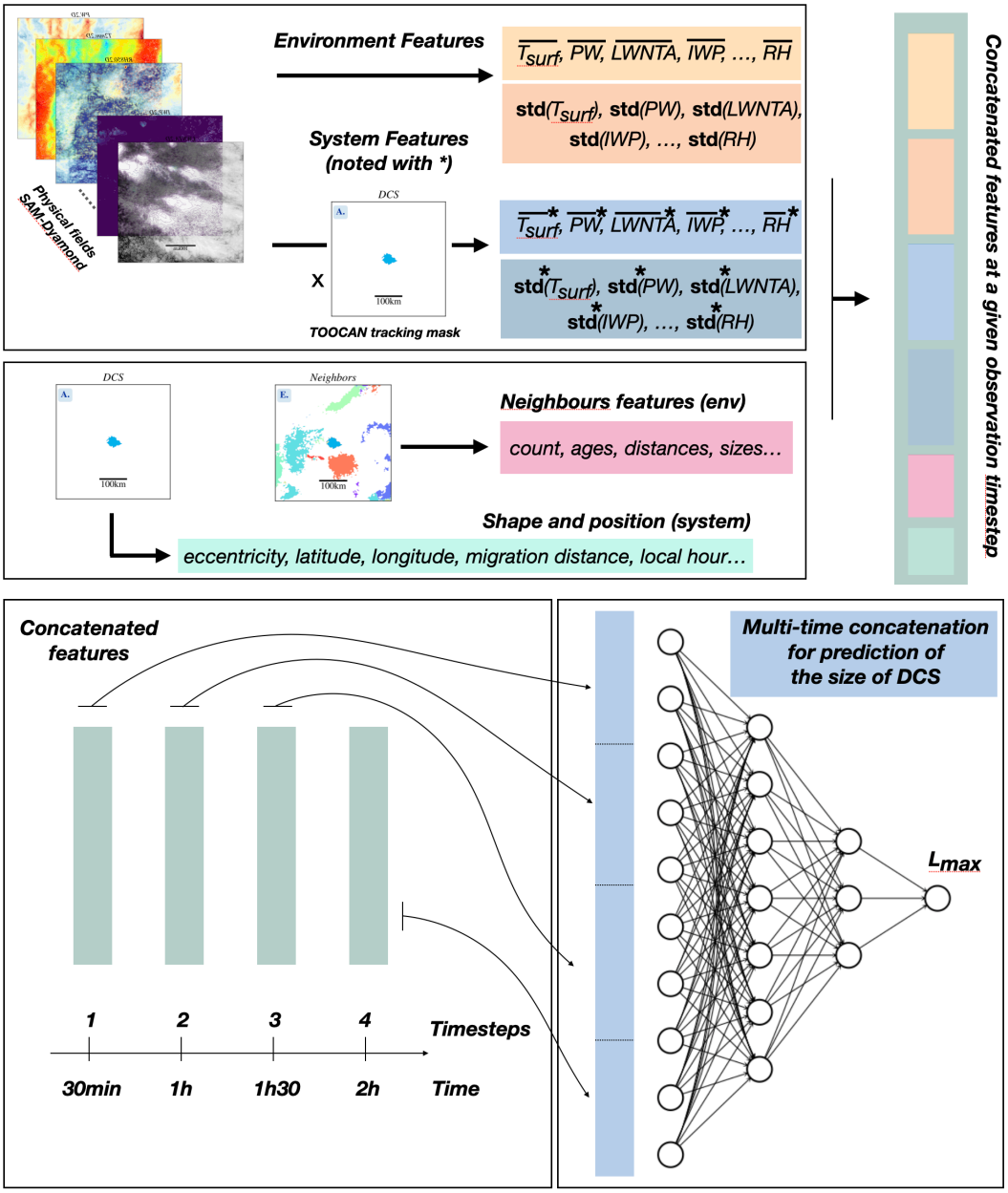

Figure 2. Method to derive from physical fields in SAM-Dyamond the interpretable features used as input of the models. Based on the tracking algorithm, physical fields in a window of 5deg ×5deg centerd on the barycenter of a DCS are extracted. We then perform the mean and the standard deviation either on the total window (environment features) or only within the system by applying the mask of the considered DCS (system features). Additionnaly we compute features based on the system shape and position, as well as features related to the neighboring systems. All these features are concatenated for a given timestep and a given DCS. We can repeat this procedure for each timestep and concatenate over the full period of observation. This is eventually the input of our models, here only the neural network is represented but it is the same for the Lasso and Random Forest.

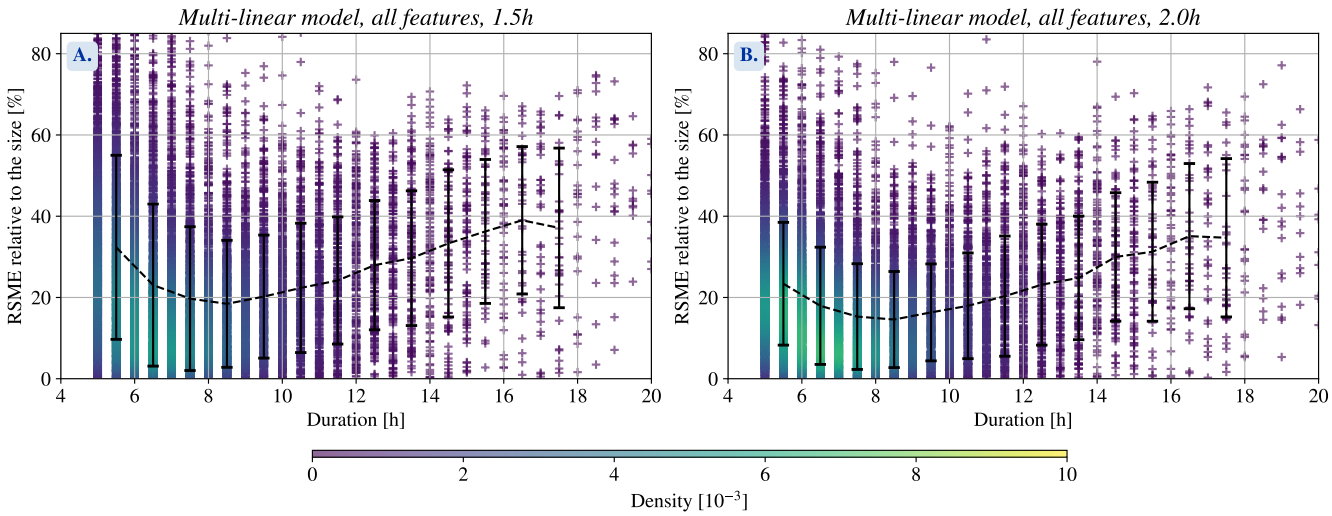

Figure 3. The figure shows the evolution of the root mean-square error (RMSE) as a function of the duration of the systems when the multilinear model is trained for either 1 hr (panel A) or 1.5 hrs (panel B). The graphical representation illustrates only a weak dependence of model precision on duration. Notably, the mean RMSE remains below 15% for systems lasting 7.5 hours and approximately 20% for those with a duration of 12.5 hours. Thus, we conclude that the model can predict the maximum area even for long duration.

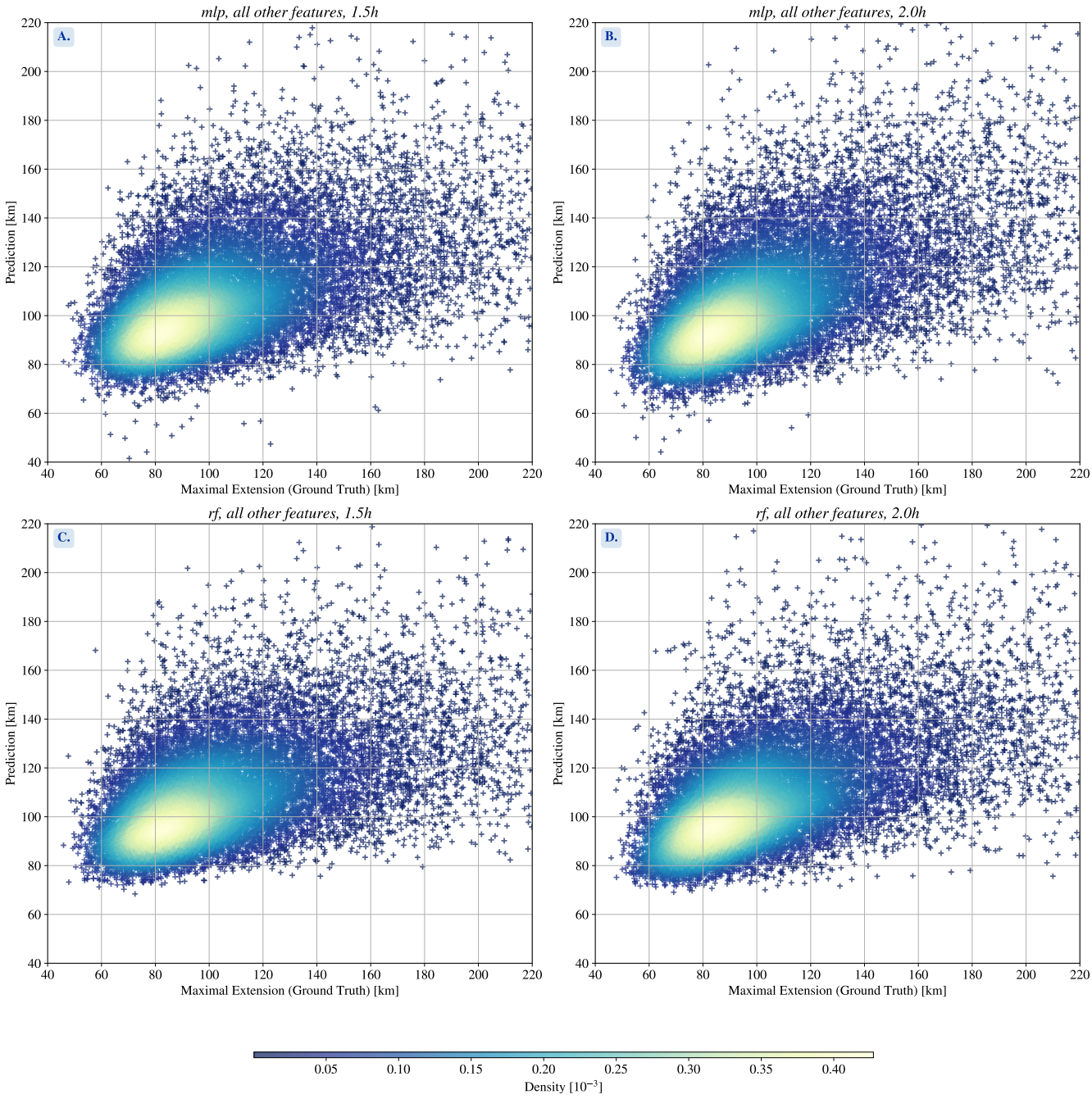

Figure 4. One-to-one diagrams for the prediction and the target for the multi-linear model (top) and random forest (bottom). The results are displayed for models trained with all features, and observation period of 1 hr (left) and 1.5 hrs (right).

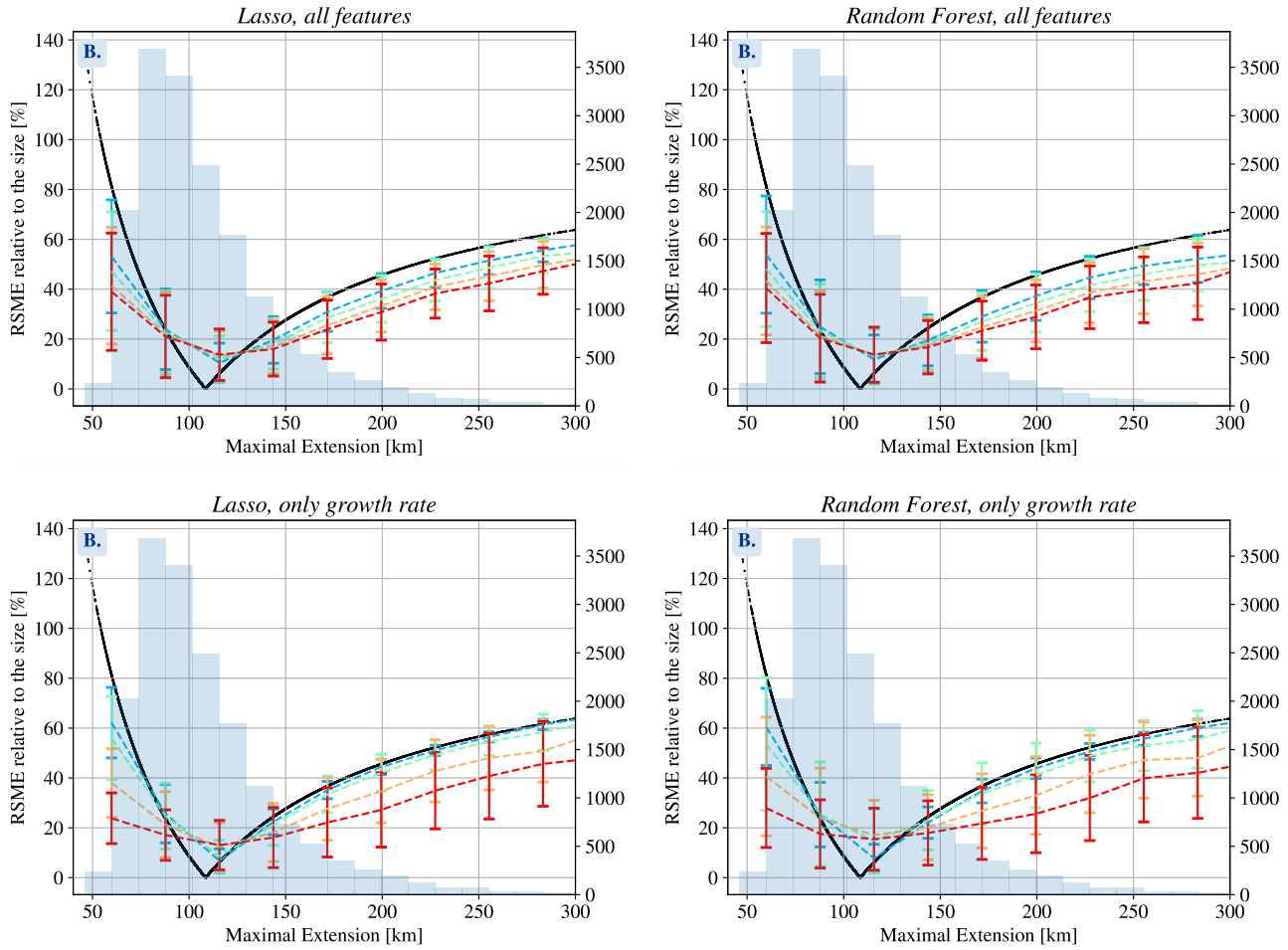

Figure 5. Mean (dashed line) and standard deviation (error bars) of the mean-squared error from the multi-linear model (lasso, top panels) and random forest model (bottom panels) prediction as a function of the effective maximum size of DCSs (x-axis). Blue curves correspond to 30 min of observation and red curves to 2 hrs of observation (different colors from blue to red correspond to 30 min increments in the observation period). The black line represents the mean-squared error for a model that predicts constantly the mean of the set. On the left, only growth rate of area is considered to predict the final size, whereas on the right, models use additional features, including shape, physical fields, migration distance and neighboring systems influence.

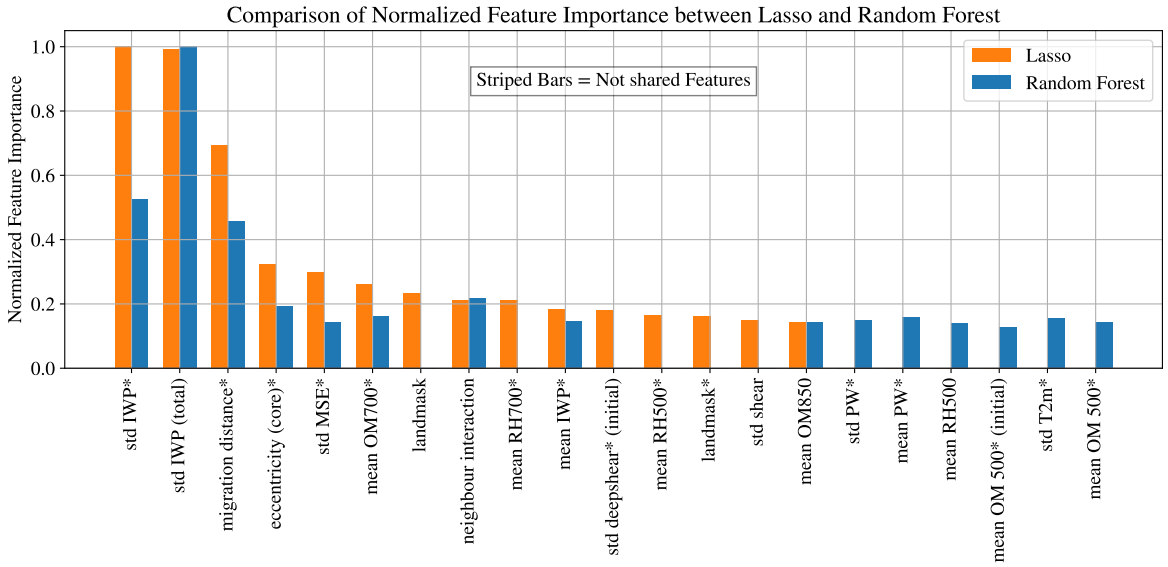

Figure 6. In orange, the normalized Coefficients optimized for the multi-linear model to solve the supervised task of predicting the maximal extension, based on the first 1.5 hour evolution of the DCS growth rate and additional features (ice water path IWP, migration distcance, etc). In blue, similar but for the Random Forest model (averaged Gini index scores). There are 15 features displayed, other features are considered negligible (coefficients are below 0.1). Starred variables are the ones characterizing the system, while unstarred variables characterize its environment. For instance std IWP* refers to the standard deviation of ice water path computed over the system cloud shield; similarly std MSE denotes the standard deviation of the moist static energy computed over the system cloud shield. Although they do not appear in exactly the same order, nor without exactly the same coefficients, 9 over the top 15 features are selected by both of these models.

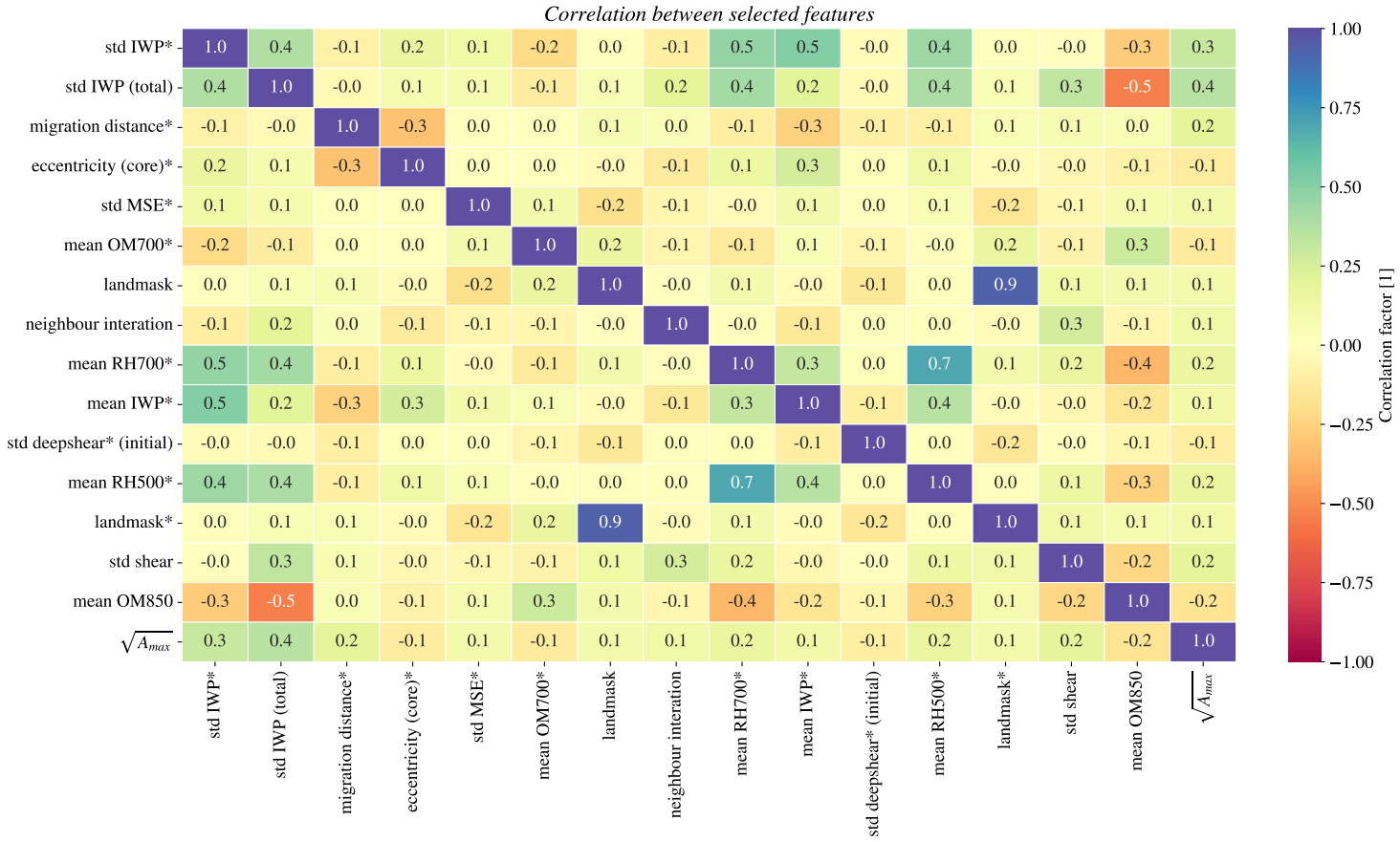

Figure 7. Correlation matrix between the 14 selected features from the value of their coefficient. It shows that features are not independant and there might be redundant information. Nevertheless, all correlations between different variables are smaller than 0.5 (the only larger correlations are between the same variable at different heights or locations, such as RH at 700hPa and 500hPa, as expected).

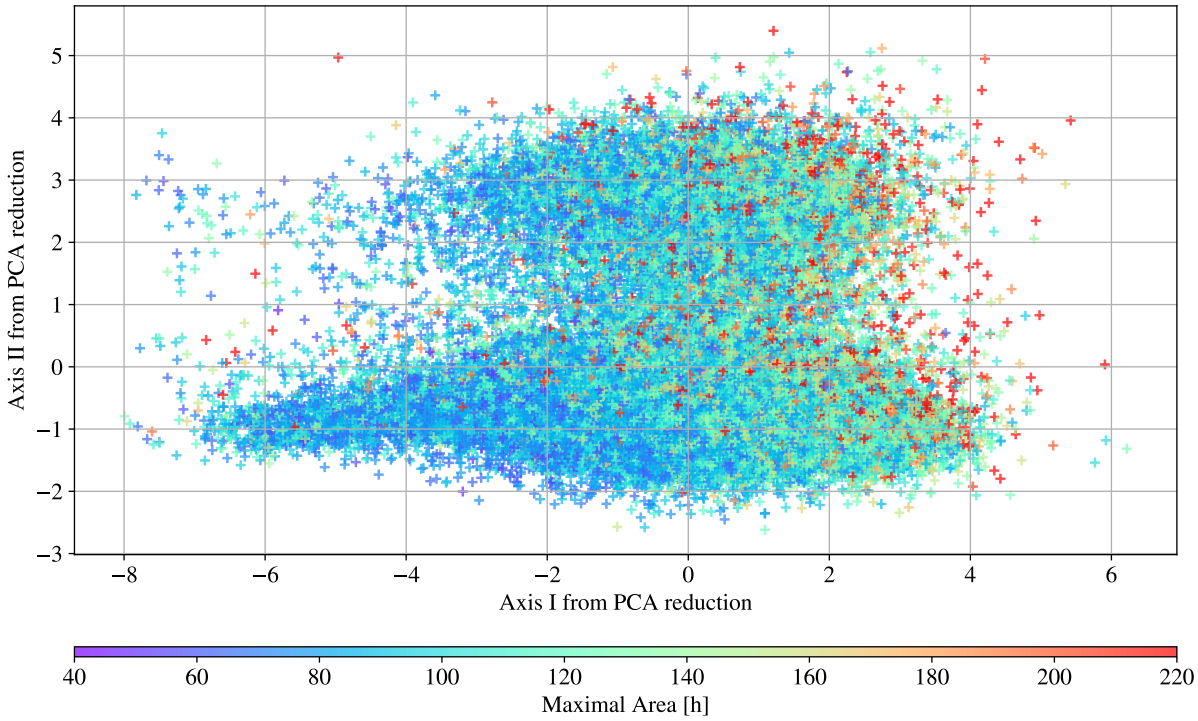

Figure 8. Principal component analysis applied on the 15 features selected from the multi-linear model (lasso). The axes are a linear combination of these features and it is not straightforward to attribute them a physical meaning. However this analysis shows an overall clustering of small (blue) and large (red) systems.

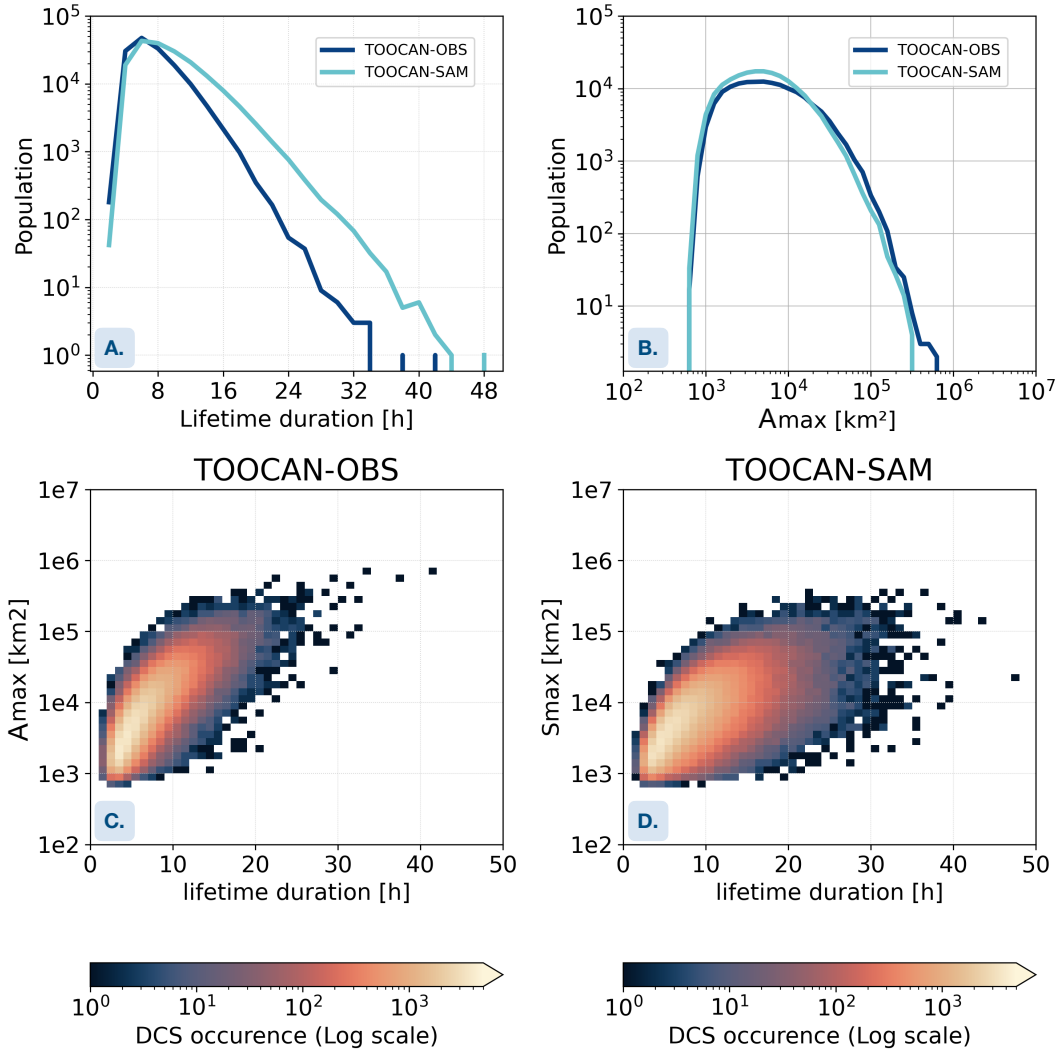

Figure 9. A. distribution of the DCS lifetime duration simulated by SAM and observed by a fleet of geostationary satellites for the period 10 August-10 September 2016 over the entire tropical belt [30°S;30°N]. B. distribution of the DCS maximum extent simulated by SAM and observed over similar period and region. C. occurrence of the observed deep convective systems according to maximum extent and lifetime duration. D. same but for simulated DCS.

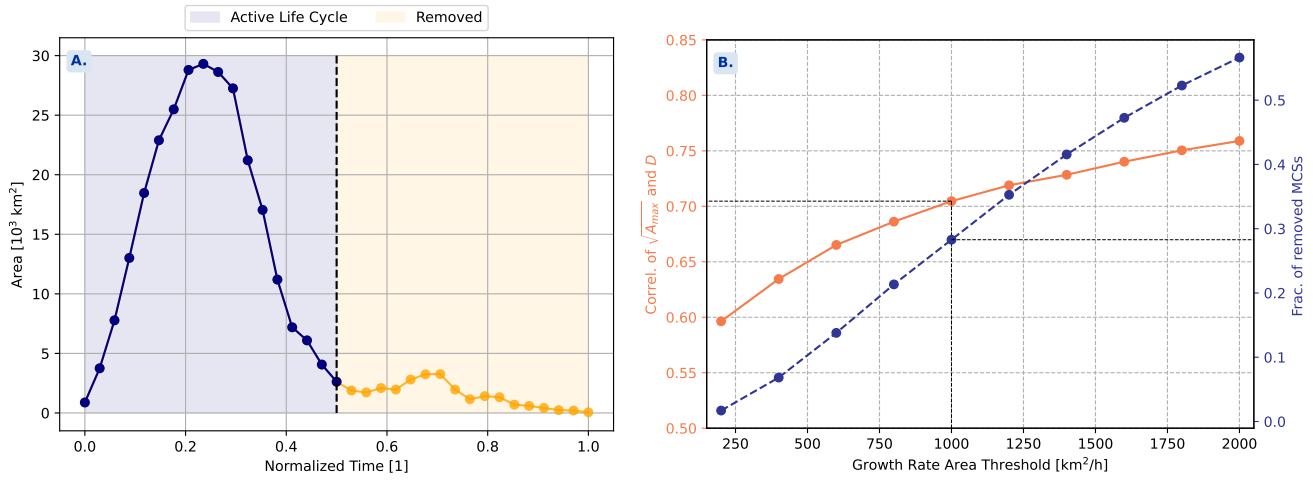

Figure 10. A. Example of a system evolution truncation of the life cycle to consider only the active life cycle. B. Determination of threshold in area growth rate as a trade off between a maximized correlation of duration and the square root of maximal area, and minimized number of systems removed. Empirically, the threshold is found at 1000 km^2/hr.