Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture

Diego Gómez
Pablo Salvador
Julia Sanz
Carlos Casanova
Daniel Taratiel
Jose Luis Casanova

Machine learning approach to locate desert locust breeding areas based on ESA CCI soil moisture

Diego Gómez, * **Pablo Salvador, Julia Sanz, Carlos Casanova,**
Daniel Taratiel, and Jose Luis Casanova
University of Valladolid, LATUV-Remote Sensing Laboratory, Valladolid, Spain

Abstract. Desert locusts have attacked crops since antiquity. To prevent or mitigate its effects on local communities, it is necessary to precisely locate its breeding areas. Previous works have relied on precipitation and vegetation index datasets obtained by satellite remote sensing. However, these products present some limitations in arid or semiarid environments. We have explored a parameter: soil moisture (SM); and examined its influence on the desert locust wingless juveniles. We have used two machine learning algorithms (generalized linear model and random forest) to evaluate the link between hopper presences and SM conditions under different time scenarios. RF obtained the best model performance with very good validation results according to the true skill statistic and receiver operating characteristic curve statistics. It was found that an area becomes suitable for breeding when the minimum SM values are over 0.07 m^3/m^3 during 6 days or more. These results demonstrate the possibility to identify breeding areas in Mauritania by means of SM, and the suitability of ESA CCI SM product to complement or substitute current monitoring techniques based on precipitation datasets. © *The Authors. Published by SPIE under a Creative Commons Attribution 3.0 Unported License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.* [DOI: 10.1117/1.JRS.12.036011]

1 Introduction

Desert locust outbreaks have been a problem since antiquity and periodically have caused devastation over local communities in Northern Africa and Middle East countries. It is well documented by ancient literature: in the Har-ra list (Assyria—the Ashurbanipal Royal Library, 669 to 626 B.C.), in decorations found in Egyptian tombs (sixth Dynasty, 2420 to 2270 B.C.), as well as in Biblical, Rabbinical, Greek and Roman literature, while control measures are also reported during Biblical, Grecian, Roman, Mishnaic, Talmudic, Byzantine, and modern times.[1] They affect local economies and living conditions, decreasing yield production in areas with water scarcity and extreme weather conditions. Desert locusts are the earliest diverging species among the genus Schistocerca and the unique one settled in Africa, indicating its high adaptability to the local conditions. Unlike other species of the same genus, it has kept some of its original traits, such as the ability to change their behavior.[2] In spite of its long pest occurrence, efforts to control its population have been in vain, at least until the late 20th century.

Schistocerca gregaria (Forskål, 1775) or desert locust is an insect that belongs to the Acrididae family, having three main stages throughout its life cycle: egg, hopper, and adult. With breeding purposes, females lay their eggs when certain moist soil conditions are met from 5 to 10 cm deep.[3] Depending on some environmental variables such as soil moisture (SM), temperature, or wind, the egg development may last between 10 and 65 days.[4,5] The newborn nymph moults five to six times as its body grows to prepare the individual for flying and reproduction purposes. After the last moult, the new adults known as fledglings, already have wings although too soft to fly yet. The next stage is the immature adult, with fully capabilities to fly. Afterward, those immature adults become sexually mature and capable to copulate and lay eggs

to complete their life cycle.[5] During this final stage, the locusts are very mobile and can travel great distances.[6] Alike to other species in the animal kingdom, desert locusts have a phase polyphenism that implies drastic changes when population density increases, either in adult or nymph stage.[2,7,8] Even though behavioral gregarization may occur within hours,[9] it takes several generations to fully display gregarious characters.[10] The phase transition induces physiological changes in lifespan, metabolism, immune responses, and reproductive physiology.[11,12] In their solitarious phase, locusts are generally bigger[10] and they present higher fecundity and smaller eggs.[13]

Solitarious desert locust populations are usually constraint into the recession areas, where annual rainfall is <200 mm.[14] However, they are able to increase rapidly their numbers when suitable conditions are met.[4] These insects are very well adapted to arid environments with erratic but sometimes high intensity precipitation episodes.[15] Some environmental events such as green vegetation blooms or rainfall are closely linked to the desert locust development, having triggering effects and enhancing outbreaks.[16,17] Temperature variability has also been demonstrated to have effects on some Schistocerca species as described by Ref. 18. This work indicated that the frequency of locust outbreaks may be altered by changes in climatic patterns. Among many environmental factors that may affect locusts, SM is the variable that mostly influences egg-laying location, egg-survival, and egg-hatching rate,[19] in addition to temperature.[20] In general, female locusts prefer open and warm sites of dry, soft, and sandy soils in which over 6 cm of depth have enough moist soil conditions.[3,21] Successful breeding conditions are usually triggered by rainfall, which provides enough moisture to the soil enhancing egg laying, development, and hatching,[16] as well as an adequate vegetation for their hoppers to feed on.[6,14] The success of preventive measures is subjected to the inaccessibility of some important breeding areas.[5] Within the recession area, there are some seasonal breeding areas in which the lack of rain may cause that some are not infested for a particular year. So that, even though breeding areas are constraint to the recession area, they may vary in accordance to suitable ecological conditions.[5]

Some authors have proposed the use of remote sensing platforms to monitor large and inaccessible locust breeding areas,[16,22–27] which usually occur away from crops.[28] Remote-sensed vegetation and precipitation are being used to derive potential grasshopper and locust habitats[22] by means of satellite platforms as LANDSAT, NOAA, Meteosat, SPOT, TERRA, or AQUA.[29] International organizations such as the Desert Locust Information Service (DLIS) from FAO have been using earth observation methods since the 1980's to assess favorable environmental conditions to the desert locust.[29]

However, monitoring arid environments can present some limitations. The vegetation is usually sparse and geomorphological features are not always well identified.[30,31] The normalized difference vegetation index (NDVI) is a proxy for vegetation presence[32] and it has been widely used to assess suitable environmental conditions for desert locust.[31] Nevertheless, this index is highly sensitive to the noise of the soil background.[33] NDVI values cannot be distinguished from sparse vegetation because bare soils have often spectral characteristics in the red and near-infrared.[34] Furthermore, the vegetation is drought tolerant due to adaptive mechanisms such as canopy architecture, leaf structure, and leaf angle. Another common proxy to identify suitable conditions for desert locust is precipitation.[35] Rainfall detection probabilities may range from 70% to 20% in arid and semiarid regions by means of remote sensing, with a high overestimation of rainfall occurrences.[36]

Currently, there is an ongoing initiative "dEsert Locust earLy Survey (SMELLS)" from the European Space Agency (ESA) to derive SM with forecasting purposes. They propose to divide the month into three decades in order to provide averaged surface SM, which comes from daily estimates. According to this initiative, relevant ranges for locust monitoring are settled between 0.10 and 0.20 m^3/m^3. Satellite SM estimations stand out as a very useful tool to overcome the high uncertainty of precipitation in arid and semiarid areas, improving the probability of locust prediction.[37] In spite of being very promising, very few studies have addressed the link between SM remote sensing and desert locusts.[19] Traditional SM measures are ground based so that survey areas are usually limited for being an expensive and time consuming activity.[38,39] Laboratory and ground-based experiments have demonstrated that SM intervenes in egg development and interruption under particular conditions of humidity.[40] According to the same authors, eggs may

remain viable in arrested state as long as 1.5 months, and then hatch after return to wet sand. In addition, locust densities are associated with relative high moisture availability.[41] These studies indicate that SM is a good proxy to identify desert locust, and it can substitute rainfall products.[42]

Species distribution models (SDM) are numerical tools to analyze the link between species occurrences and environmental factors. They provide an ecological insight to predict species distribution over space or time given certain environmental characteristics.[43] Their machine learning methods increase traditional predictive performance and their capacity to incorporate complex interaction among variables,[44] being eligible to work with large ecological datasets.[45] The random forest (RF)[46] and generalized linear model (GLM)[47] are two commonly used machine learning algorithms to generalize species distributions. RF has been available for almost 20 years, and it performs very well in ecological predictions.[48] GLMs are mathematical extensions of linear models that do not force data into unnatural scales, and thereby allow for nonlinearity and nonconstant variance structures in the data.[49] They have also been used to analyze ecological relationships given their flexibility in comparison to classical Gaussian distributions.[50]

The aim of this study is to identify suitable SM conditions for desert locust eggs as well as to hopper desert locusts in solitarious phase. It is based on SM estimations from satellite remote sensing imagery and ground-based observations of hopper desert locusts. We have used SDMs to better understand the link between SM and desert locusts to predict their likely distribution across landscapes and breeding areas. The study area is Mauritania and the survey period goes from 1985 to 2015.

2 Materials and Methods

2.1 Study Area

The study site is Mauritania, which is located in the Maghreb region of Western Africa (Fig. 1). We have chosen this study area to be one of the major breeding and recession regions for desert
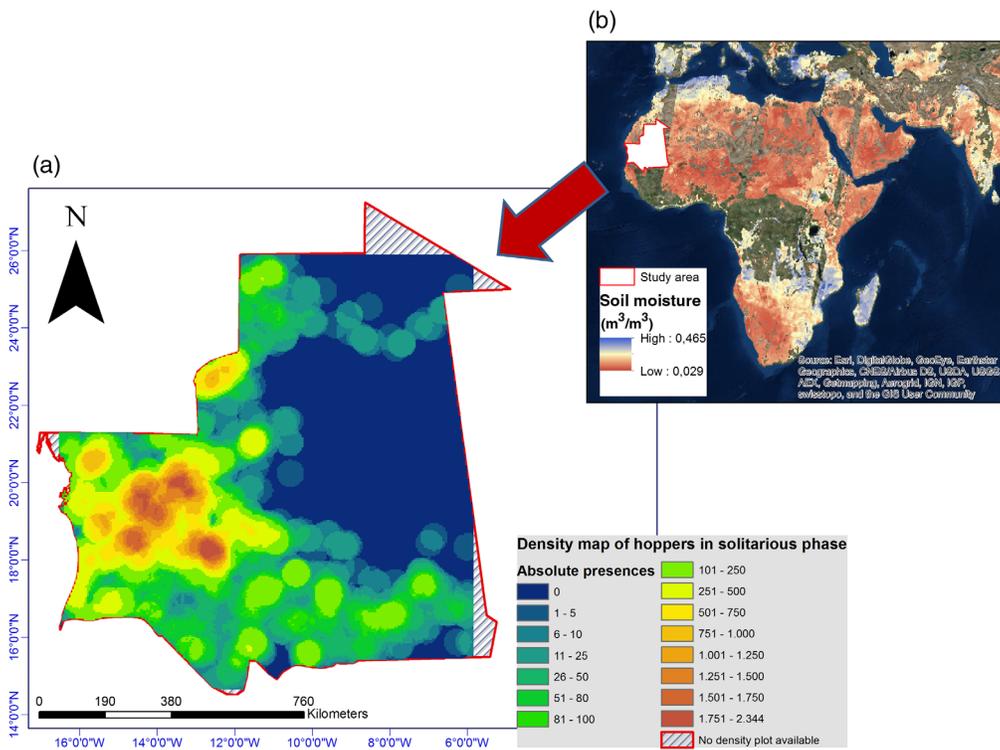

Fig. 1 (a) Study area location within the African continent with an ESA CCI SM image (January 5, 2015). (b) The density plot of solitarious hoppers between 1985 and 2015 in the study area. Data presences come from SWARMS database from FAO.

locust.[51] Mauritania is a vast country of 1,030,700 km^2 with large arid plains and only one continuous water flow, the Senegal River.

According to Koppen classification,[52] two climate types are present: hot desert climate "BWh" and hot semiarid climate "BSh." BWh is predominant in most of the country, which spatially coincides with part of the Sahara Desert (north) and the Sahelian belt (south). Rainfall is scarce and intense, being generally <150 mm/year in average (Fig. 2). BSh accounts for the Southernmost strip, where the rainfall average is higher than 200 mm/year, in addition to cooler and less fluctuating "day-night" temperatures.

2.2 Survey Data

Schistocerca WARning and Management System (SWARMS) is a database used by the Desert Locust Information Service (DLIS) at FAO for desert locust global monitoring and early warning. It compiles desert locust data since 1985 that have been collected by national survey and control teams of affected countries. It geo-locates field observations on a daily basis although some uncertainties may be expected.[26,53] For this study, we selected hoppers on a solitarious phase as the target population for two reasons: solitary phase accounts for nonrestricting conditions and hopper stage (wingless nymph) may have lower mobility than adults due to the lack
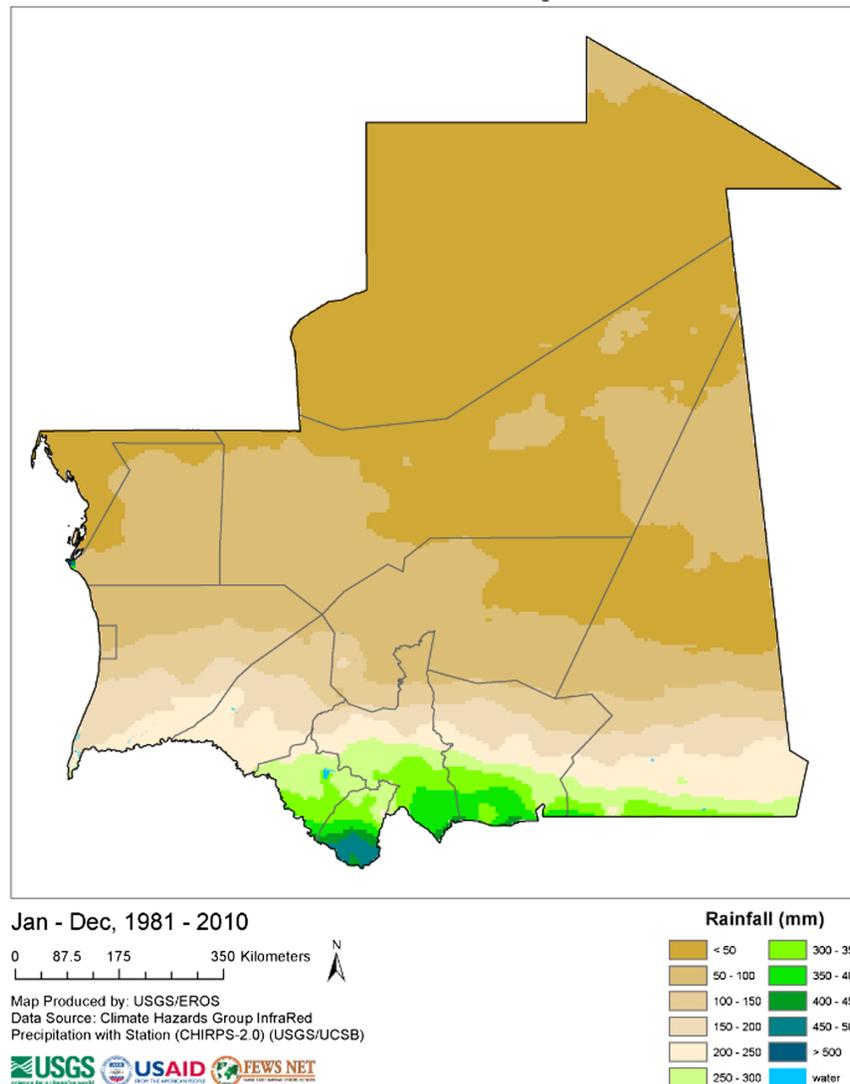

Fig. 2 Mauritania historical average rainfall (1981 to 2010). Source: USGS/EROS.

of wings. There were 12,027 solitarious hopper sightings for the time span 1985 to 2015, spatially distributed as seen in Fig. 1. Even though the database contemplates the absence records, we have not considered them for two reasons. First, during the recession periods, individuals are mostly solitary (solitarious phase) and many times go unnoticed for survey teams.[54] Second, the number of absence records is very low, which causes unbalance between samples of presences and absences.

2.3 Satellite Data

The ESA CCI SM v03.2 is a multidecadal and global satellite-observed SM dataset generated via the climate change initiative (CCI) of the ESA. It is a product that combines various single active and passive sensors into three harmonized products: a merged active, a merged passive, and a merged from active and passive sensors. Based on the existing literature, these merged products generally outperform the single-sensor input products.[55]

For the purpose of this study, we have used the merged active and passive product to be more complete. It uses the pixel from either the active or passive source, or the average value of both depending on the performance of the vegetation optical depth from the Advanced Microwave Scanning Radiometer for EOS (AMSR-E) C-band observations.[56] The combination of images from radar (active) and radiometer sensors (passive) provides information about the volumetric surface SM (up to 5 cm depth), and it is expressed in m^3/m^3 units. Its spatial resolution is 0.25 deg and offers daily coverage worldwide from 1978 up to 2015.[55,57,58] This product comprises active data retrieved from C-band scatterometers on board of ERS-1, ERS-2, MetOp-A, and MetOp-B satellites (generated by the "TU Wien") and passive data obtained from microwave observations by the following sensors: Nimbus 7 SMMR, DMSP SSM/I, TRMM TMI, Aqua AMSR-E, Coriolis WindSat, GCOM-W1 AMSR2, and SMOS (generated by VU University Amsterdam in collaboration with NASA) (Table 1). This product has been validated against ground-based reference measures or alternate estimates from other projects and sensors.[55,57] In general, ESA CCI SM dataset provides good estimations of SM with respect to land surface models and *in situ* observations. Nevertheless, it presents some uncertainties with particular surface conditions such as dense vegetation or organic soils,[55] which are not the case of our study area.

2.4 Methods

The ESA CCI SM v03.2 product was used to geographically compare the seasonal presence of solitarious hoppers of desert locust by months, with SM values from 1985 to 2015. Breeding areas in Mauritania vary widely throughout the year according to the National Centre for Prevention and Control of Desert Locusts in Mauritania (CNLA). During summer months, desert locusts usually breed in southern parts of the country. Whereas breeding occurs in the center and the northwestern part from September to December, and from December to May in the northern areas of Mauritania.[59] It is widely accepted that these insects have regional migrations following certain environmental conditions.[60]

We have extracted the coordinates of each hopper in solitarious phase and its corresponding date from SWARMS database. Even though the database does have some absence records, we did not use them for being very unbalanced in comparison with presences. In addition to that, those records can be also considered as "pseudoabsences" owing to hoppers in solitarious phase may go unnoticed at low densities.[26] Thus, we found it convenient to randomly generate a grid of "pseudoabsences" as reported in other studies using SDMs.[61,62]

Pseudoabsence samples were computed based on two principles. First, they were located within a maximum of 50-km radius mask created of ever desert locust presence (1985 to 2015), aiming to select areas with environmental and geophysical potentialities and to reduce geographical bias. We chose this distance for matching visually with the density map (Fig. 1), where most of the areas with no presences are masked out. Otherwise, it could misguide SDM predictions.[63]

Second, date allocation was done using a uniform random arrangement with R-software. Each pseudoabsence location was assigned a date within the first and the last hopper presence

Table 1 List of satellite platforms, onboard sensors to measure SM at specific frequency, producer of the product, and time availability of each single product.[55]

Platform sensor	Frequency used for SM retrieval (GHz)	Product name/producer	Dataset availability
Radiometers			
Nimbus7 SMMR	6.6	VU University Amsterdam (VUA)/National Aeronautics and Space Administration (NASA) [Land Parameter Retrieval Model (LPRM)]	October 1978 to August 1987
DMSP SSM/I	19.4	VUA/NASA (LPRM)	June 1987 onwards
TRMM TMI	10.7	VUA/NASA (LPRM)	November 1997 to April 2015
		Princeton University (LSMEM)	January 1998 to December 2004
AQUA AMSR-E	6.9, 10.7	VUA/NASA (LPRM)	June 2002 to October 2011
		University of Montana/ Numerical Terradynamic Simulation Group	June 2002 to October 2011
		US National Snow and Ice Data Center (NSIDC)	June 2002 to October2011
		Japanese Aerospace Exploration Agency (JAXA)	June 2002 to October 2011
		Princeton University (LSMEM)	June 2002 to September 2011
Coriolis WindSat	6.8, 10.7	VUA/NASA (LPRM)	January 2003 to August 2012
		U.S. Naval Research Laboratory	January 2003 onward
SMOS MIRAS	1.4	ESA/Centre Aval de Traitement des Données SMOS (CATDS)	November 2009 onward
		ESA/EUMETCAST (for L2-SM-NRT-NN product)	November 2009 onward
		VUA/VanderSat (LPRM)	November 2009 onward
Aquarius	1.4	NSIDC	August 2011 to June 2015
FengYun-3B MWRI	10.7	VUA/NASA (LPRM)	July 2011 onward
GCOM W1 AMSR2	6.9, 7.3, 10.7	VUA/NASA (LPRM)	July 2012 onward
		JAXA	July 2012 onward
SMAP	1.4	NASA	February 2015 onward
		VUA/NASA (LPRM)	February 2015 onward
Scatterometers			
ERS-1/2 AMI WS	5.3	Vienna University of Technology (TU Wien/WARP), ESA	August 1991 to July 2011
MetOp-A/B ASCAT	5.3	EUMETSAT H-SAF, (TU Wien/WARP)	January 2007 onward

date of the SWARMS database (1985 to 2015). These pseudoabsence points were generated randomly and equally weighted to the presences (pseudoabsence and presence weighted sums are equal) for predicting species occurrences or distribution.[64] It may occur that some presences and pseudoabsences coincide geographically within the same pixel; however, it is very unlikely that they have the same assigned date. Each pseudoabsence date has been randomly allocated from 1985 to 2015, which implies that they will likely not have the same SM values.

The duration of locust life cycles is variable, depending on the environmental conditions of the habitat,[65] nevertheless we rely on the following premises to create the variables in our study. Eggs are laid at 5 to 10 cm depth, and the egg incubation period may range from 10 to 65 days.[4] After hatching, nymph phase may last between 24 and 95 days since the egg was laid. Thus, under the most severe environmental circumstances, the maximum expected egg-hopper development time would be 95 days.[5] SWARMS database registers the sighting date and phase but not the age of each individual so that we have established up to 95 days prior the sighting record as the time analysis. Figure 3 shows the sequence of the proposed method as a flow chart.

Given the coordinates of each presence and pseudoabsence record, the corresponding daily SM value was extracted based upon the sighting or assigned date, up to 95 days backward. Based on these antecedent SM conditions, we generated variables dividing the analysis time into different time intervals (16, 12, 8, and 6 days) and assess the performance of the model with each of them. By this method, we aim to cover and differentiate critical events in the locust lifecycle such as egg-laying, egg-hatching, and early stages of the nymph phase individuals as well as to deal with punctual missing data (Fig. 3). Some areas of SM imagery had missing data due to the satellite revisit times used to generate ESA CCI SM v03.2. We have computed the minimum, mean, and maximum SM values within each time interval to obtain a representative value of such period. Then, we assess which descriptive statistic provides better information to the model in terms of performance. If no value was found for a particular time interval, the presence or absence record is not included in the model. In this way, we mitigate the effect that the missing information could provoke on the model results. Even though SM may vary greatly on a daily basis,[66] the biological evolution for egg and hopper development needs some days to be altered,[5] so that we found convenient this approach to generate the model variables.

Therefore, we have studied four different scenarios: A, B, C, and D. As previously mentioned, we have first extracted SM values, on a daily basis, up to 95 days before the presence or pseudoabsence date record. Each of the proposed scenarios contemplates a different division in terms of days: A = 16 days, B = 12 days, C = 8 days, and D = 6 days. Hence, we aimed to obtain one representative SM value per each subdivision of time, within each scenario. In order
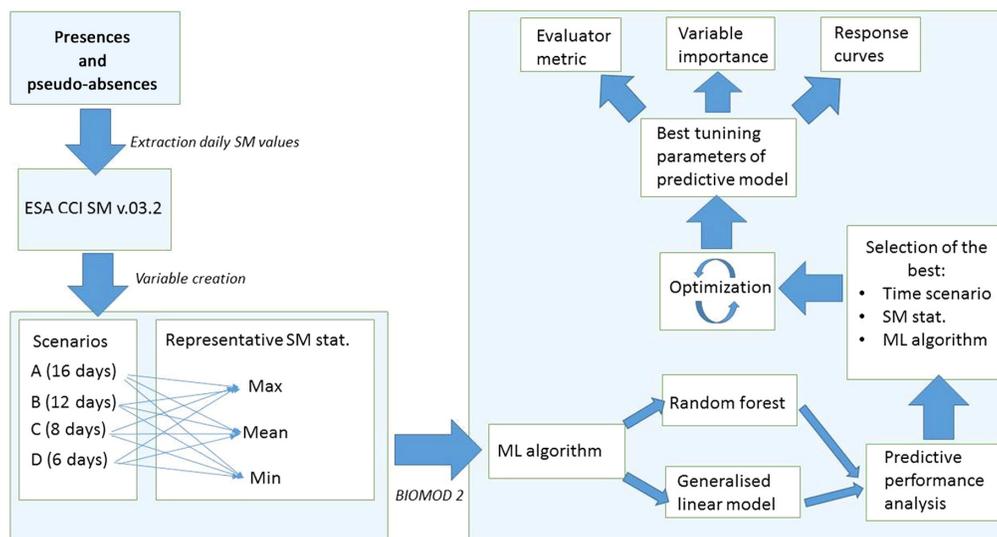

Fig. 3 Flowchart of the proposed methodology to study the link of ESA CCI SM with desert locusts using machine learning approach.

to acquire this representative SM value, we have computed the minimum, mean, and maximum out of the daily SM values contained in every time interval.

Thus, Fig. 4 shows variable creation for each scenario (A, B, C, and D) based on SM and presence and pseudoabsence dates. For instance, scenario (A) contemplates equal time intervals of 16 days so that (SM1) indicates the SM value on the local pixel between -95 and -80 days (both included) prior the presence or pseudoabsence date. (SM2) SM values on the local pixel between -79 and -64 days prior the presence or pseudoabsence date and the rest accordingly as detailed in Fig. 4. Time interval for scenario (A) is 16 days, which generates 6 variables; 12 days for (B) with 8 variables; 8 days for (C) with 12 variables; and 6 days for (D) with 16 variables. Time equals to 0 ($t = 0$) corresponds to the presence or pseudoabsence sighting date. Within each scenario, three different alternatives are independently tested (minimum, mean, and maximum SM value within the given time interval).

Some publications suggest the suitability of machine-learning (ML) approaches to model species distributions, since they may perform better than the traditional regression-based algorithms.[44] In this study, we have used BIOMOD2 tool[67] implemented for R software.[68] We tested two different ML modeling techniques to describe and model the link between desert locust and SM: GLM[47] and RF.[46] GLM is a very popular modeling approach that has been widely used to model and predict habitats and species distribution.[69,70] The formula object was set to be "quadratic" (default) and the information criteria for the stepwise selection procedure was the Akaike information criteria. GLM approach implemented in BIOMOD2 only runs on presence-absence data, so binomial distribution family was used. RF algorithm is a flexible and easy to use ML approach that has been demonstrated to have good predictive performances in ecology and species distribution.[48] It can be used both for classification and regression problems. The most important tuning parameters are the "mtry" (number of variables randomly selected at each split of the tree as it grows) and "ntree" (number of trees). We have set these two parameters with their default values: "ntree" $= 500$[71,72] and "mtry" (in classification) $=$ the squared of the number of variables.[73] The minimum size of terminal nodes "NodeSize" and the maximum number of terminal nodes "MaxNodes" were also left with their defaults values, which are five and null, respectively.[74]

In spite of the generalized use of some statistics to assess model performances, there is still an ongoing debate about their use.[75,76] We decided to select three broadly used evaluation methods for cross-comparisons: relative operating characteristics "ROC,"[77] Cohen's Kappa "KAPPA,"[78] and true skill statistic "TSS."[75]

The ROC evaluation method uses the area under the curve (AUC) to discriminate between events and nonevents. Its score ranges from 0 (worst score) to 1 (perfect score), and values under 0.5 are considered to indicate random chance of the prediction.[79]

KAPPA statistic is one of the most used methods to measure model performance on presence-absence predictions, and it indicates the relative accuracy of the forecast comparing with the random chance. It ranges between -1 (the worst score) to 1 (perfect score), where values
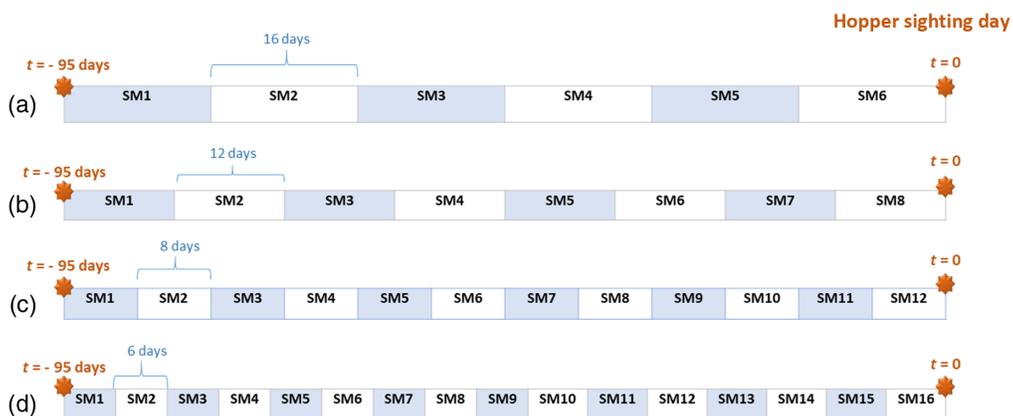

Fig. 4 Variable names and their distribution back in time for four different scenarios: A, B, C, and D.

Table 2 Index for classifying model prediction accuracy.[67]

Accuracy	AUC	KAPPA/TSS
Excellent or high	0.9 to 1	0.8 to 1
Good	0.8 to 0.9	0.6 to 0.8
Fair	0.7 to 0.8	0.4 to 0.6
Poor	0.6 to 0.7	0.2 to 0.4
Fail or null	0.5 to 0.6	0 to 0.2

under 0 indicates no predictive skill. Although these evaluation procedures could be used independently, it is recommended to use several so as to assess the accuracy of the statistical models. This is an index for classifying model prediction accuracy (Table 2).

The Biomod2 package allows the user to randomly subset the original dataset into two subsets, 70% of the data to calibrate the models and 30% to validate the predictions. When found the best scenario and variables to choose, we repeated the process five times to the best performing algorithm to obtain a robust test of the model, where each replicate uses a unique random split 70% to 30% of the data.[67] Presence and pseudoabsences were set to have the same importance in the calibration process, with a prevalence value of 0.5. The most effective SDM require data on both species presence and the available environmental conditions at random where no presences were reported (known as pseudo-absence data) in the area.[64]

Based on model results, the best performing algorithm with the best scenario and representative statistic of SM values is selected. Then, we applied an optimization process to ensure that the algorithm we have settled on is presenting the best possible performance.[80] We tuned the algorithm hyperparameters to find their best combination in terms of predictive performance, and finally an objective comparison of the results. The best tuning parameters were chosen to run the final model.

We used the response curves to assess the prediction of the model, which are independent of the SDM algorithm used. The response curves allow comparing the probability of presence based on ROC, TSS, and Kappa metrics with the variables used in the model. It facilitates the interpretation of relationships between environmental variables and predicted responses of species, even though they may not be apparent from the outputs of the model.[81] The contribution of each variable to the final model is analyzed. The higher the value is, the more influential the variable is in the model. A 0 value means no influence at all.

The aim is to evaluate desert locust presence probabilities to locate potential breeding areas, based on remotely sensed SM conditions.

3 Results

SM monthly averages (Figs. 5 and 6) suggest a spatial correlation with usual breeding areas, indicating high SM values in the south for the months: July, August, September, and October; whereas higher values are found in the north and northeastern parts of Mauritania during December, January, and February. In general, autumn breeding sites (blue dots in Fig. 6) do not show visual correlation with the monthly mean SM values. Nevertheless, a statistical analysis was not done on a monthly basis but as detailed in Fig. 4.

GLM and RF algorithms were used with SM variables that relied upon various time intervals (16, 12, 8, and 6 days) and their maximum, minimum, or mean (Tables 3 and 4) SM values. Based on ROC, TSS, and KAPPA statistics, we obtained performance scores from an independent test dataset. The results showed that RF obtained the best performance for our study, whereas GLM performed far behind. The highest scores were obtained when the time interval was 6 days (scenario D) and the representative SM value was the minimum acquired within the time interval. According to Table 2, the RF algorithm obtained a high or very good performance with respect to ROC-AUC with 0.95 and good performance for Kappa and TSS statistics with
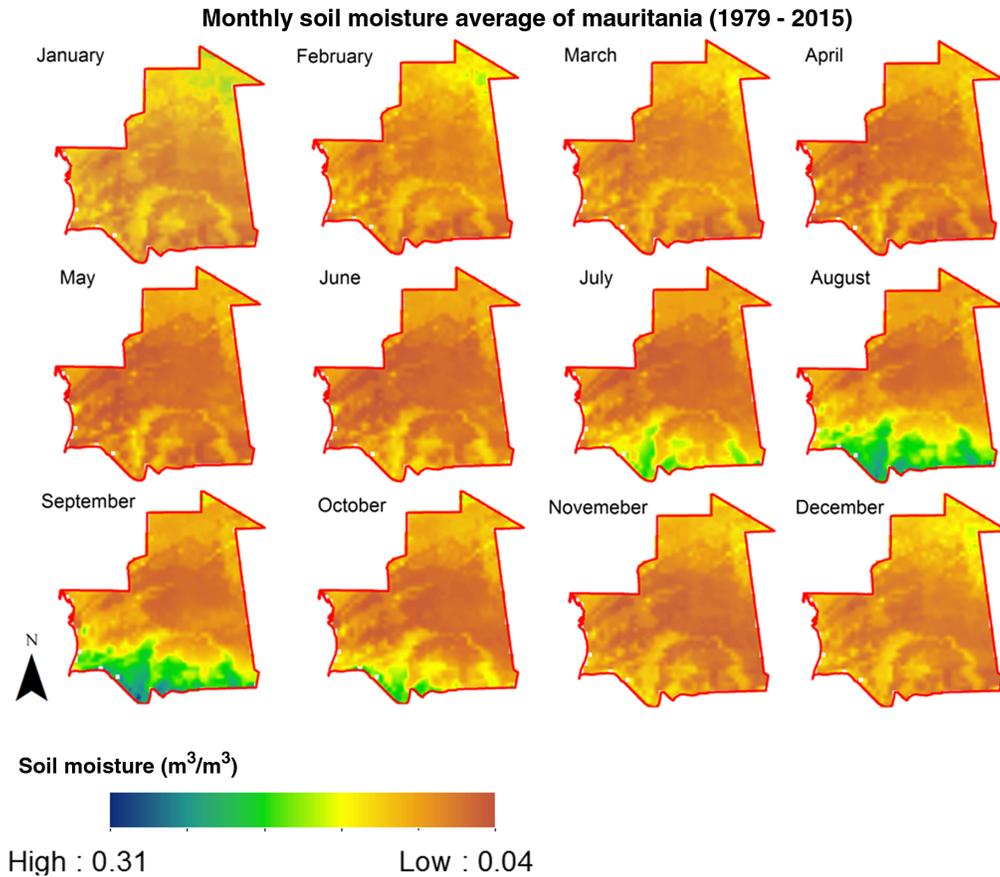
Monthly soil moisture average of mauritania (1979 - 2015)

Fig. 5 SM average per month for the time span 1985 to 2015, units is in m³/m³.
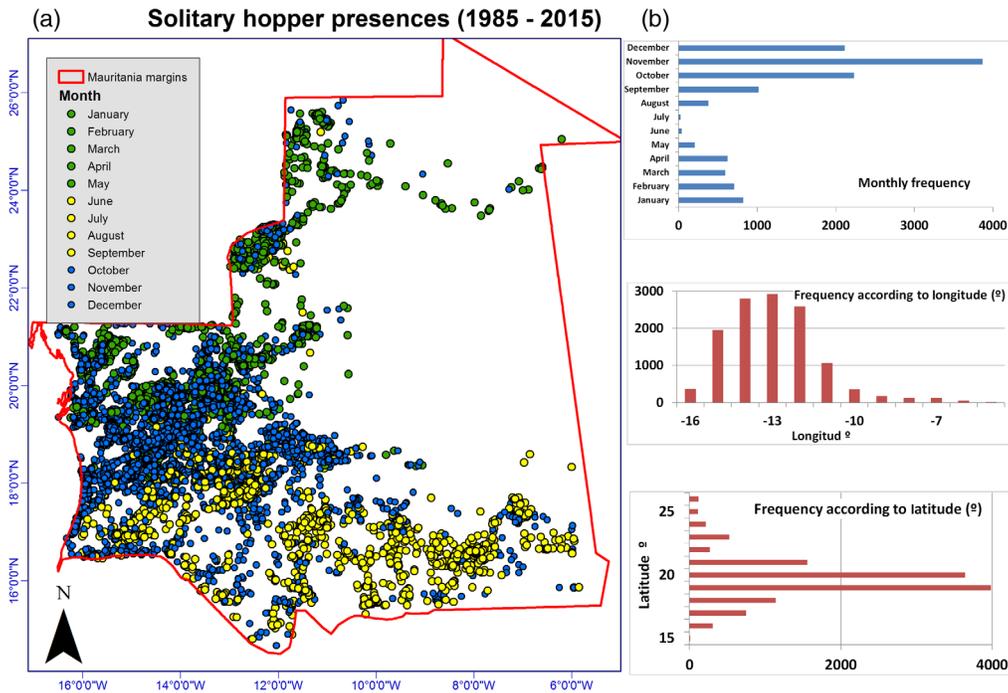

Fig. 6 (a) Location map of solitarious hopper presences reported from 1985 to 2015, grouped per months. (b) Frequency histograms of presences based on months, latitude, and longitude.

Table 3 Random forest results per time-scenario, representative statistic to generate the SM variables (maximum, mean, or minimum per each interval) and the model performance per statistical metric. Sensitivity and specificity are expressed in %.

	16 days (scenario A)			12 days (scenario B)			8 days (scenario C)			6 days (scenario D)		
	Test	Sensitivity	Specificity	Test	Sensitivity	Specificity	Test	Sensitivity	Specificity	Test	Sensitivity	Specificity
						Max.						
ROC-AUC	0.873	75.207	84.323	0.906	79.129	86.56	0.923	81.226	87.695	0.931	80.798	90.096
TSS	0.594	75.635	83.773	0.656	79.129	86.56	0.688	81.856	86.916	0.708	81.147	89.494
KAPPA	0.594	75.635	83.773	0.654	79.129	86.56	0.686	81.856	86.916	0.704	82.129	88.645
						Mean						
ROC-AUC	0.854	70.385	82.672	0.888	74.892	84.873	0.915	77.771	88.279	0.929	78.359	90.98
TSS	0.53	71.498	81.397	0.596	74.892	84.873	0.66	76.509	89.351	0.693	78.485	90.803
KAPPA	0.529	71.498	81.397	0.594	74.892	84.873	0.658	82.457	83.247	0.688	78.485	90.803
						Min.						
ROC-AUC	0.908	79.23	86.902	0.937	85.097	86.468	0.937	86.362	85.422	0.95	87.611	87.619
TSS	0.661	79.315	86.786	0.715	85.097	86.468	0.714	86.002	85.552	0.752	87.421	87.796
KAPPA	0.661	79.315	86.786	0.715	86.855	84.627	0.715	86.753	84.675	0.751	87.421	87.796

Table 4 GLM results per time-scenario, representative statistic to generate the SM variables (maximum, mean, or minimum per each interval) and the model performance per statistical metric. Sensitivity and specificity are expressed in %.

	16 days (scenario A)			12 days (scenario B)			8 days (scenario C)			6 days (scenario D)		
	Test	Sensitivity	Specificity	Test	Sensitivity	Specificity	Test	Sensitivity	Specificity	Test	Sensitivity	Specificity
						MAX						
ROC-AUC	0.610	53.609	63.054	0.703	69.732	60.724	0.262	57.855	68.084	0.29	56.876	72.02
TSS	0.163	52.896	63.75	0.304	59.96	70.359	0.679	56.894	69.448	0.696	65.463	63.601
KAPPA	0.163	52.896	63.75	0.304	70.107	60.08	0.26	57.855	68.084	0.288	62.801	65.83
						Mean						
ROC-AUC	0.640	62.853	57.838	0.246	56.299	68.334	0.221	55.242	66.591	0.233	56.274	66.855
TSS	0.206	61.94	58.679	0.676	57.106	67.874	0.651	54.521	67.727	0.657	55.925	67.351
KAPPA	0.206	61.94	58.679	0.245	56.299	68.334	0.22	55.242	66.591	0.231	56.274	66.855
						Min						
ROC-AUC	0.676	77.261	51.058	0.419	82.012	59.834	0.34	72.424	61.526	0.326	72.085	60.807
TSS	0.283	77.575	50.623	0.755	81.724	60.172	0.699	73.085	61.006	0.704	73.162	60.099
KAPPA	0.284	77.575	50.623	0.421	82.012	59.834	0.341	72.424	61.526	0.327	79.436	52.741

0.75. The sensitivity and specificity was over 87%. Slightly lower values are found when using the maximum or mean SM values across the scenario D, demonstrating the suitability of 6 days coverage time to build the SM variables of the model. Scenario A (16 days) obtained the worst model performance when using mean SM values as representative of the given interval. Nevertheless, this scenario still obtained a fair performance of 0.6 for TSS and kappa statistics, and ROC-AUC = 0.90 when using the minimum SM value across their time length.

Model performance increases when the time interval of the variables gets smaller and the representative SM value is the minimum for such period. Therefore, we suggest regarding minimum SM values over 6 days period to link solitarious hopper presences and SM values of the ground.

RF was the best performing algorithm, using scenario D and the minimum SM values obtained in each time interval. We have tuned RF algorithm for the two most important hyperparameters: the number of trees "ntree" (50, 500, 1000, 2000, and 4000) and the number of variables randomly sampled as candidates at each split "mtry" (2, 4, 6, 8, and 10). First, we optimized the number of trees and second the mtry. As shown in Fig. 7, the default parameters established by Biomod2 for RF (ntree = 500 and mtry = 4) obtained the best model performance, whose evaluator metrics did not greatly differ from other tuning options. The poorest performance was obtained with ntrees = 50 and mtry = 2 (lower value parameters than the default proposed by BIOMOD2). The increase of ntrees or mtry has not improved model results, with relatively very small changes in model performance. It is also noticeable how the ROC-AUC evaluator remains more or less constant across the different attempts, whereas the changes of TSS and KAPPA are slightly larger.

Therefore, the best algorithm (RF) was optimized after the tuning phase with ntree = 500 and mtry = 4. The best model results were obtained using the variables created with scenario D and the minimum SM reached at each time interval. Finally, we ran RF for five iterations to aim for robust results. Model performance scores are compiled in Table 5.

The metric scores are in accordance with the ones obtained in Table 3 for the same scenario (D) and chosen variables (minimum SM). In general, testing values and sensitivity are slightly lower, whereas ROC-AUC and TSS specificity are somewhat higher. In essence, score values do not differ considerably when running more iterations and averaging their metrics. The impact of SM variables in the final model results (RF, scenario D, and minimum SM) is summarized in Fig. 8.

The most relevant variables for the outcome model were SM1, SM2, SM3, and SM4, which stand for the minimum SM values obtained between 95 and 90, 89 and 84, 83 and 78, 77 and 72 days before the sighting record, respectively. Figure 8 indicates the greater impact of these mentioned variables (mostly over 10%) in comparison with the rest, which do not overcome the 5% per each. Figure 9 shows the response curves of these four more relevant variables that are over
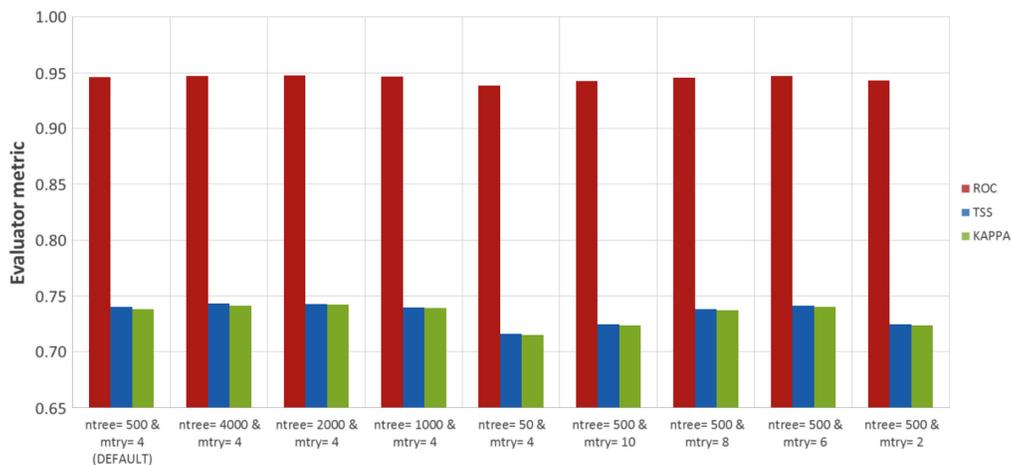

Fig. 7 Comparison of different RF results using different tuning parameters, with scenario D and the minimum SM value per interval (best performances in the previous step). X-axis represents the parameter changes and Y-axis the model performance of each tuning combination according to ROC, KAPPA, and TSS statistics.

Table 5 RF results after five iterations using the best scenario (6 days) with the minimum SM values obtained in each interval. Sensitivity and specificity are expressed in %.

RF			
Five iterations	Test	Sensitivity	Specificity
ROC-AUC	0.946	84.911	89.105
TSS	0.740	85.468	88.461
KAPPA	0.738	87.325	86.508
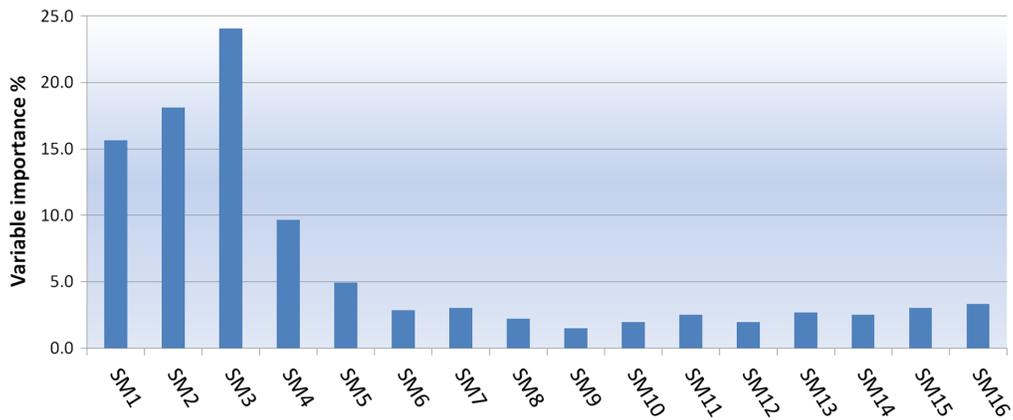
Fig. 8 Variable importance in % of each variable from scenario D (6 days), using the minimum SM value obtained in each time interval for RF.
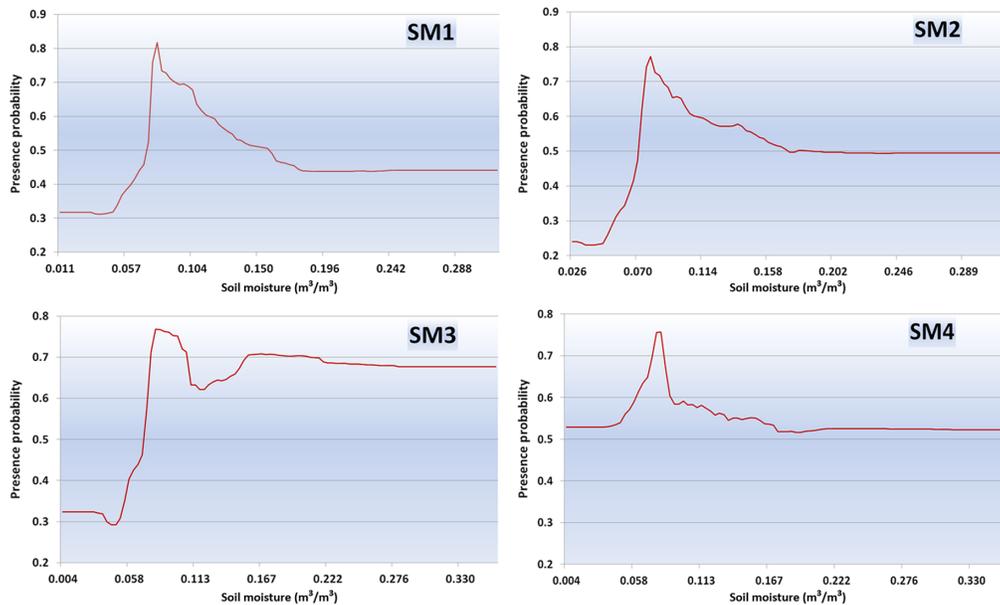

Fig. 9 Response curves for hopper's desert locust for SM1, SM2, SM3, and SM4 variables for RF. The *Y*-axis represents the presence probability of the prediction, while *X*-axis stands for SM values.

5% of importance. The plots suggest some potential thresholds of SM content to increase the probability of presence. The minimum SM values acquired during SM1, SM2, SM3, and SM4 denote a positive influence in hopper occurrences. It is observed that the range of SM values in which the probability of presence is over 0.5 varies. Presence probabilities tend to keep steady by 0.5 when SM values reaches 0.15 for SM1, SM2, and SM4. SM3 keeps a high probability over such figure. Nevertheless, there is a common trend by the 0.07 (m^3/m^3) to increase the probability of presence within 72 and 95 days afterward.

4 Discussion

It is widely assumed that rainfall over 25 mm in two consecutive months is generally enough for locust breeding and development.[82] Nevertheless, remotely sensed precipitation in arid environments has some limitations such as high rainfall overestimation due to subcloud evaporation.[83] Aiming to solve the problems associated with remote sensing precipitation, we have analyzed the link from ESA CCI SM remote sensing product with field surveys of hopper desert locust from SWARMS—FAO. In addition, we assess the suitability of this SM product to derive desert locust breeding sites.

The importance of SM in egg laying and development has been long known, as well as the role of fresh vegetation, which is greatly determined by water availability in the soil.[4] SM monthly averages suggest a spatial correlation with summer and winter breeding areas. It coincides with the regional climatic conditions of Mauritania as reported in other works.[59,60] Winter rainfall is usual in the north while summer rain in the south of the country. Nevertheless, typical autumn breeding areas do not seem to be accounted for the monthly SM patterns. In arid environments, there is a direct relationship between rainfall and SM[84,85] so that problems such as subcloud evaporation[83] may be avoided with the applied methodology. Despite ESA CCI SM only senses the first 5 cm of the top soil, and desert locusts lay eggs usually at a depth down to 10 cm; this system seems appropriate due to the strong relationship of the top SM with deeper layers.[86]

Our analysis reveals the importance of variable creation as a previous step to modeling. We have tested different time intervals for the variable creation. In addition, we have chosen different representative SM values for the given time-span (maximum, mean, and minimum) and presence and pseudoabsence sites. Perhaps, the use of pseudoabsences may be controversial in certain fields because bring some sort of uncertainty into the results.[87] However, their use is generally justified for providing a set of conditions available in the region that need to be included in the SDM.[88]

The highest performance was acquired by the RF algorithm when dividing the whole survey time into ranges of 6 days, and selecting the minimum SM as the variable value. Even though previous literature[70] have used the GLM model with a binomial distribution to identify potential factors that determine species presences or absences, GLM approach did not perform well in our study. It was observed that RF performance did not greatly change using hyperparameter values larger than ntree = 500 and mtry = 4 (default values in BIOMOD 2 for RF). Whereas, lower ntree and mtry values performed slightly worse in terms of TSS and KAPPA metrics. According to Ref. 67, our RF model has had an excellent performance based on ROC-AUC metric with 0.946, and a good performance for TSS and Kappa statistics with 0.740 and 0.738, respectively. The probability of hopper detection (sensitivity) is over 85%, being able to correctly identify (specificity) over 86% of the pseudoabsence records. The variables with more weight in the model results were SM1, SM2, SM3, and SM4, whose cover time range from 95 to 72 days before the sighting record. Locust eggs develop and hatch successfully when there is enough moisture in the soil,[40] whereas insufficient moisture may stop egg development or dry them out.[4] Our results indicate that the minimum SM conditions over at least 6 days should remain higher than 0.07 m^3/m^3. This value is in accordance, although slightly lower, with the SM range proposed by Ref. 89, which is between 0.10 and 0.20 m^3/m^3. Hopper mortality is closely linked to food shortage,[4] which in arid environments is closely linked with inadequate precipitation.[6,41] Thus, remotely sensed SM may also be a good indicator of suitable conditions to infer hopper presences and locate breeding areas. A good understanding of the geographical relationship between desert locust populations and their potential breeding habitats can improve desert locust survey and control operations.[41]

The applied methodology offers very promising results to correctly identify breeding areas based on 30 years of SM values. The ESA CCI SM dataset is the most complete and consistent global SM data record available.[58] To the best knowledge of the authors, there has not been any previous desert locust analysis using this SM dataset. Given the acknowledged importance of SM for desert locust and the length of ESA CCI SM dataset, our results may signify a breakthrough to complement the ongoing locust monitoring techniques used until today.

5 Conclusions

This paper aimed to assess the significant importance of satellite SM products to locate breeding areas for desert locusts in solitarious phase. Despite remote sensing techniques greatly evolving to date, very few works have addressed the SM relationship to identify desert locusts by earth observation methods. This survey is based on the ESA CCI SM product, the most complete and consistent available SM dataset. We have used a machine learning approach to assess the relationship between desert locust presences and antecedent SM conditions and estimate the accuracy of our model. This study confirmed the robustness of the applied methodology, where 30 years of locust records and SM values were used to feed the model, but note that some uncertainty is expected due to the use of pseudoabsence data.

The monthly SM values suggest a spatial correlation with usual breeding areas in Mauritania. So far, desert locust suitable sites have been mainly delimited based on rainfall estimates from satellite remote sensing. However, some literature marks the high overestimation of these products over dry regions. Therefore, we suggest the use of ESA CCI SM product to overcome that problem either to complement other rainfall products or to substitute them in certain instances of high uncertainty.

Furthermore, we have modeled quantitatively the relationship between hopper presences and SM under different scenarios and variables. The best model performance was obtained by RF, when using the minimum SM value within 6 days interval, for a maximum survey time of 95 days before the sighting date. The validation phase acknowledged the suitability of this methodology to identify hopper presences with an ROC-AUC of 0.94 and TSS and Kappa of 0.74. The importance of SM thresholds and survey time has also been addressed: when the minimum SM value of a certain location overcomes 0.07 m^3/m^3 during 6 days or more, the area becomes favorable as a breeding zone. However, these figures should be taken carefully. Variable importance showed that the most relevant variables of the model would cover between 95 and 72 days before the sighting record. It implies, as highlighted in other works, that certain SM levels need to be maintained over time not just for egg laying but egg development and hatching. So that, monitoring periods should be longer than 6 days to those favorable areas for a successful egg development and hatching.

This paper proposes a machine learning approach based on SM time series to predict breeding areas, by means of remote sensing. According to these results, the observed SM during certain periods stands as a very reliable contributor to accurately predict hopper presences in Mauritania; consequently, its monitoring may reduce the locust impact on local communities. Future researches may aim to ensemble other studied environmental variables along with SM datasets to implement more developed warning systems. This increasing amount of information that remote sensing platforms are providing will require the use of artificial intelligence approaches. For instance, the correct use of ensemble SDM may sometimes improve the performance of individual models, which might contribute to solve problems like the exposed in this work.

References

1. D. Nevo, "The desert locust, *Schistocerca gregaria*, and its control in the land of Israel and the Near East in antiquity, with some reflections on its appearance in Israel in modern times," *Phytoparasitica* **24**(1), 7–32 (1996).

2. H. Song et al., "Phylogeny of locusts and grasshoppers reveals complex evolution of density-dependent phenotypic plasticity," *Sci. Rep.* **7**(1), 6606 (2017).

3. B. Uvarov, *Grasshoppers and Locusts. A Handbook of General Acridology, Vol. 2. Behaviour, Ecology, Biogeography, Population Dynamics*, Centre for Overseas Pest Research, London (1977).

4. D. Pedgley, *Desert Locust Forecasting Manual (Volume 1 of 2)*, Centre for Overseas Pest Research, London (1981).

5. P. M. Symmons and K. Cressman, *Desert Locust Guidelines: Biology and Behaviour*, FAO, Rome (2001).

6. L. V. Bennett, "The development and termination of the 1968 plague of the Desert locust, *Schistocerca gregaria* (Forskål)(Orthoptera, Acrididae)," *Bull. Entomol. Res.* **66**(3), 511–552 (1976).

7. M. P. Pener and S. J. Simpson, "Locust phase polyphenism: an update," *Adv. Insect Physiol.* **36**, 1–272 (2009).

8. S. J. Simpson, G. A. Sword, and N. Lo, "Polyphenism in insects," *Curr. Biol.* **21**(18), R738–R749 (2011).

9. P. E. Ellis, "The behaviour of locusts in relation to phases and species," Colloques Internationaux du Centre National de la Recherche Scientifique, No. 114. Physiologie, Comportement et Écologie des Acridiens en Rapport avec la Phase, CNRS, Paris (1962).

10. U. R. Ernst et al., "Epigenetics and locust life phase transitions," *J. Exp. Biol.* **218**(1), 88–99 (2015).

11. M. P. Pener and Y. Yerushalmi, "The physiology of locust phase polymorphism: an update," *J. Insect Physiol.* **44**(5–6), 365–377 (1998).

12. D. A. Cullen et al., "From molecules to management: mechanisms and consequences of locust phase polyphenism," in *Advances in Insect Physiology*, Vol. **53**, pp. 167–285, Academic Press, Oxford (2017).

13. K. Maeno and S. Tanaka, "Is juvenile hormone involved in the maternal regulation of egg size and progeny characteristics in the desert locust?" *J. Insect Physiol.* **55**(11), 1021–1028 (2009).

14. J. A. Tratalos and R. A. Cheke, "Can NDVI GAC imagery be used to monitor desert locust breeding areas?" *J. Arid. Environ.* **64**(2), 342–356 (2006).

15. B. Uvarov, *Grasshoppers and Locusts: A Handbook of General Acridology, Vol. 1, Anatomy, Physiology, Development, Phase Polymorphism, Introduction to Taxonomy*, Anti-Locust Research Centre at the University Press, London (1966).

16. C. J. Tucker, J. U. Hielkema, and J. Roffey, "The potential of satellite remote sensing of ecological conditions for survey and forecasting desert-locust activity," *Int. J. Remote Sens.* **6**(1), 127–138 (1985).

17. J. U. Hielkema, J. Roffey, and C. J. Tucker, "Assessment of ecological conditions associated with the 1980/81 desert locust plague upsurge in West Africa using environmental satellite data," *Int. J. Remote Sens.* **7**(11), 1609–1622 (1986).

18. G. Yu, H. Shen, and J. Liu, "Impacts of climate change on historical locust outbreaks in China," *J. Geophys. Res.* **114**, D18 (2009).

19. Z. Liu et al., "Relationship between oriental migratory locust plague and soil moisture extracted from MODIS data," *Int. J. Appl. Earth Obs. Geoinf.* **10**(1), 84–91 (2008).

20. Y. Nishide and S. Tanaka, "Desert locust, *Schistocerca gregaria*, eggs hatch in synchrony in a mass but not when separated," *Behav. Ecol. Sociobiol.* **70**(9), 1507–1515 (2016).

21. G. Popov, "Ecological studies on oviposition by swarms of the desert locust (*Schistocerca gregaria* Forskal) in Eastern Africa," *Anti-Locust Bull.* **31**, 1–70 (1958).

22. G. Tappan, D. G. Moore, and W. I. Knausenberger, "Monitoring grasshopper and locust habitats in Sahelian Africa using GIS and remote sensing technology," *Int. J. Geogr. Inf. Syst.* **5**(1), 123–135 (1991).

23. P. Ceccato et al., "The desert locust upsurge in West Africa (2003–2005): information on the desert locust early warning system and the prospects for seasonal climate forecasting," *Int. J. Pest Manage.* **53**(1), 7–13 (2007).

24. J. Pekel et al., "Development and application of multi-temporal colorimetric transformation to monitor vegetation in the desert locust habitat," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **4**(2), 318–326 (2011).

25. F. Waldner et al., "Operational monitoring of the desert locust habitat with earth observation: an assessment," *ISPRS Int. J. Geo-Inf.* **4**(4), 2379–2400 (2015).

26. C. Renier et al., "A dynamic vegetation senescence indicator for near-real-time desert locust habitat monitoring with MODIS," *Remote Sens.* **7**(6), 7545–7570 (2015).

27. C. Piou et al., "Mapping the spatiotemporal distributions of the desert locust in Mauritania and Morocco to improve preventive management," *Basic Appl. Ecol.* **25**, 37–47 (2017).

28. P. Symmons, "Strategies to combat the desert locust," *Crop Prot.* **11**(3), 206–212 (1992).

29. A. Latchininsky et al., "Applications of remote sensing to locust management," in *Land Surface Remote Sensing*, N. Baghdadi and M. Zribi, Eds., pp. 263–293, Elsevier, San Diego (2017).

30. C. Piou et al., "Coupling historical prospection data and a remotely-sensed vegetation index for the preventative control of desert locusts," *Basic Appl. Ecol.* **14**(7), 593–604 (2013).

31. M. Lazar et al., "Location and characterization of breeding sites of solitary desert locust using satellite images Landsat 7 ETM+ and Terra MODIS," *Adv. Entomol.* **3**(1), 6–15 (2015).

32. H. Santin-Janin et al., "Assessing the performance of NDVI as a proxy for plant biomass using non-linear models: a case study on the Kerguelen archipelago," *Polar Biol.* **32**(6), 861–871 (2009).

33. A. R. Huete, "A soil-adjusted vegetation index (SAVI)," *Remote Sens. Environ.* **25**(3), 295–309 (1988).

34. E. Despland, J. Rosenberg, and S. J. Simpson, "Landscape structure and locust swarming: a satellite's eye view," *Ecography* **27**(3), 381–391 (2004).

35. J. U. Hielkema and F. L. Snijders, "Operational use of environmental satellite remote sensing and satellite communications technology for global food security and locust control by FAO: the ARTEMIS and DIANA systems," *Acta Astronaut.* **32**(9), 603–616 (1994).

36. T. Dinku et al., "Evaluating detection skills of satellite rainfall estimates over desert locust recession regions," *J. Appl. Meteorol. Climatol.* **49**(6), 1322–1332 (2010).

37. J. Bolton, M. Brown, and P. Ceccato, "Improving desert locust decision support in Africa and Asia using SMAP soil moisture estimates," in *NASA Soil Moisture Active Passive (SMAP) Applications Workshop*, NOAA, Silver Spring, Maryland (2009)

38. B. Q. Sun et al., "Evolution feature on the moisture of soil for Loess Highland in Gansu," *Adv. Earth Sci.* **20**(9), 1041–1046 (2005).

39. G. Huang et al., "Effects of conservation tillage on soil moisture and crop yield in a phased rotation system with spring wheat and field pea in dryland," *Acta Ecol. Sin.* **26**, 1176–1185 (2006).

40. A. Shulov and M. P. Pener, "Studies on the development of eggs of the desert locust (*Schistocerca gregaria* Forskål) and its interruption under particular conditions of humidity," *Anti-Locust Bull.* **41** (1963).

41. G. W. Teklu, "Habitats and spatial pattern of solitarious desert locusts (*Schistocerca gregaria* Forsk.) on the coastal plain of Sudan," PhD Thesis, Wageningen University (2003).

42. M. Cherlet et al., "Spot vegetation contribution to desert locust habitat monitoring," in *Proc. of the Vegetation Workshop*, Lake Maggiore, Italy (2000).

43. J. Elith and J. R. Leathwick, "Species distribution models: ecological explanation and prediction across space and time," *Annu. Rev. Ecol. Evol. Syst.* **40**, 677–697 (2009).

44. J. Elith et al., "Novel methods improve prediction of species' distributions from occurrence data," *Ecography* **29**(2), 129–151 (2006).

45. P. T. Robinson et al., "Mapping the global distribution of livestock," *PLoS One* **9**(5), e96084 (2014).

46. L. Breiman, "Random forests," *Mach. Learn.* **45**(1), 5–32 (2001).

47. P. McCullagh, "Generalized linear models," *Eur. J. Oper. Res.* **16**(3), 285–292 (1984).

48. C. Mi et al., "Why choose random forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence," *PeerJ* **5**, e2849 (2017).

49. J. T. Hastie and R. J. Tibshirani, *Generalized Additive Models, Volume 43 of Monographs on Statistics and Applied Probability*, Chapman and Hall, London (1990).

50. M. P. Austin, "Models for the analysis of species' response to environmental gradients," *Vegetatio* **69**, 35–45 (1987).

51. H. Culmsee, "The habitat functions of vegetation in relation to the behaviour of the desert locust *Schistocerca gregaria* (Forskål)(Acrididae: Orthoptera)-a study in Mauritania (West Africa)," *Phytocoenologia* **32**(4), 645–664 (2002).

52. M. Kottek et al., "World map of the Köppen-Geiger climate classification updated," *Meteorol. Z.* **15**(3), 259–263 (2006).

53. M. A. B. Ebbe, "Biogéographie du criquet pèlerin en Mauritanie: Fonctionnement d'une aire grégarigène et conséquences sur l'organisation de la surveillance et de la lutte anti-acridienne (No. AGP/DL/TS/31), Stations de recherche acridienne sur le terrain, séries techniques," FAO, Rome, Italy (2003).

54. C. Meynard et al., "Climate-driven geographic distribution of the desert locust during recession periods: subspecies' niche differentiation and relative risks under scenarios of climate change," *Global Change Biol.* **23**, 4739–4749 (2017).

55. W. Dorigo et al., "ESA CCI soil moisture for improved earth system understanding: state-of-the art and future directions," *Remote Sens. Environ.* **203**, 185–215 (2017).

56. Y. Liu et al., "Trend-preserving blending of passive and active microwave soil moisture retrievals," *Remote Sens. Environ.* **123**, 280–297 (2012).

57. A. Gruber et al., "Triple collocation-based merging of satellite soil moisture retrievals," *IEEE Trans. Geosci. Remote Sens.* **55**(12), 6780–6792 (2017).

58. W. Wagner et al., "Fusion of active and passive microwave observations to create an essential climate variable data record on soil moisture," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **7**, 315–321 (2012).

59. M. O. B. Ebbe, "Preventative control for desert locust pest in Africa: experiences of mauritania" (2012), https://www.jircas.go.jp/sites/default/files/publication/proceedings/2012-session-41_0.pdf (21 November 2017).

60. A. Van Huis, K. Cressman, and J. I. Magor, "Preventing desert locust plagues: optimizing management interventions," *Entomol. Exp. Appl.* **122**(3), 191–214 (2007).

61. A. E. Zaniewski, A. Lehmann, and J. McC Overton, "Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns," *Ecol. Modell.* **157**(2–3), 261–280 (2002).

62. R. Engler, A. Guisan, and L. Rechsteiner, "An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data," *J. Appl. Ecol.* **41**(2), 263–274 (2004).

63. A. M. Barnes et al. "Geographic selection bias of occurrence data influences transferability of invasive Hydrilla verticillata distribution models," *Ecol. Evol.* **4**(12), 2584–2593. (2014)

64. M. Barbet-Massin et al., "Selecting pseudo-absences for species distribution models: how, where and how many?" *Meth. Ecol. Evol.* **3**(2), 327–338 (2012).

65. A. T. Showler, "The desert locust in Africa and western Asia: complexities of war, politics, perilous terrain, and development," in *Radcliffe's IPM worldtextbook*, (2009), E. B. Radcliffe and W. D. Hutchison, Eds., University of Minnesota, St. Paul, Minnesota https://ipmworld.umn.edu/showler-desert-locust (16 August 2018).

66. T. Wang et al., "Effect of vegetation on the temporal stability of soil moisture in grass-stabilized semi-arid sand dunes," *J. Hydrol.* **521**, 447–459 (2015).

67. W. Thuiller, B. Lafourcade, and M. Araujo, "ModOperating manual for BIOMOD," in *BIOMOD: Species/Climate Modelling Functions*, W. Thuiller and B. Lafourcade, Eds., (2010) R package version: 1-1 (2009), Université Joseph Fourier, Grenoble, http://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/inst/doc/Biomod%20Manual.pdf?revision=67&root=biomod&pathrev=218 (16 August 2018).

68. R Development Core Team, "R: a language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria (2012).

69. A. Guisan and N. E. Zimmermann, "Predictive habitat distribution models in ecology," *Ecol. Modell.* **135**(2–3), 147–186 (2000).

70. J. A. Sanchez-Zapata et al., "Desert locust outbreaks in the Sahel: resource competition, predation and ecological effects of pest control," *J. Appl. Ecol.* **44**(2), 323–329 (2007).

71. J. Elith and C. H. Graham, "Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models," *Ecography* **32**(1), 66–77 (2009).

72. M. B. Garzón et al., "Intra-specific variability and plasticity influence potential tree species distributions under climate change," *Global Ecol. Biogeogr.* **20**(5), 766–778. (2011)

73. R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern Recognit. Lett.* **31**(14), 2225–2236 (2010).

74. W. Thuiller, D. Georges, and R. Engler, "biomod2: Ensemble platform for species distribution modeling. R package version 3.1-64," http://CRAN.R-project.org/package=biomod2, (2016), ftp://ftp2.de.freebsd.org/pub/misc/cran/web/packages/biomod2/biomod2.pdf (16 August 2018).

75. O. Allouche, A. Tsoar, and R. Kadmon, "Assessing the accuracy of species distribution models: prevalence, Kappa and the true skill statistic (TSS)," *J. Appl. Ecol.* **43**(6), 1223–1232 (2006).

76. A. Ruete and G. C. Leynaud, "Goal-oriented evaluation of species distribution models' accuracy and precision: true skill statistic profile and uncertainty maps," *PeerJ* **3**, e1208v1 (2015).

77. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**(1), 29–36 (1982).

78. R. A. Monserud and R. Leemans, "Comparing global vegetation maps with the Kappa statistic," *Ecol. Modell.* **62**(4), 275–293 (1992).

79. T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.* **27**(8), 861–874 (2006).

80. J. Brownlee "Machine learning mastery" (2014), http://machinelearningmastery.com/discover-feature-engineering-howtoengineer-features-and-how-to-getgood-at-it.

81. J. Elith et al., "The evaluation strip: a new and robust method for plotting predicted responses from species distribution models," *Ecol. Modell.* **186**(3), 280–289 (2005).

82. FAO and WMO, "Weather and desert locusts" (2016), http://www.fao.org/ag/locusts/common/ecg/2350/en/2016_WMOFAO_WeatherDLe.pdf (19 April 2018).

83. T. Dinku, P. Ceccato, and S. J. Connor, "Challenges of satellite rainfall estimation over mountainous and arid parts of east Africa," *Int. J. Remote Sens.* **32**(21), 5965–5979 (2011).

84. S. E. Nicholson and T. J. Farrar, "The influence of soil type on the relationships between NDVI, rainfall, and soil moisture in semiarid Botswana. I. NDVI response to rainfall," *Remote Sens. Environ.* **50**(2), 107–120 (1994).

85. L. Brocca et al., "A new method for rainfall estimation through soil moisture observations," *Geophys. Res. Lett.* **40**(5), 853–858 (2013).

86. C. Albergel et al., "From near-surface to root-zone soil moisture using an exponential filter: an assessment of the method based on in-situ observations and model simulations," *Hydrol. Earth Syst. Sci. Discuss.* **12**, 1323–1337 (2008).

87. T. Hastie and W. Fithian, "Inference from presence-only data: the ongoing controversy," *Ecography* **36**(8), 864–867 (2013).

88. S. J. Phillips et al., "Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data," *Ecol. Appl.* **19**(1), 181–197 (2009).

89. M. J. Escorihuela et al., "SMOS based high resolution soil moisture estimates for desert locust preventive management," *Remote Sens. Appl.* **11**, 140–150 (2018).

Diego Gómez is a PhD candidate at University of Valladolid (LATUV). He graduated in environmental sciences and received his master's degree in earth sciences and environmental geology. His areas of interest are natural hazards, environmental and agricultural monitoring. The sustainability journal has recently published his master's thesis about the rise of the Menor sea level. Currently, he researches the problem of desert locusts in Mauritania by means of Earth observation methods and artificial intelligence.

Biographies for the other authors are not available.