

A POINT PROCESS APPROACH FOR SPATIAL STOCHASTIC MODELING OF THUNDERSTORM CELLS

BY

BJOERN KRIESCHE, REINHOLD HESS AND VOLKER SCHMIDT

Abstract. In this paper we consider two different approaches for spatial stochastic modeling of thunderstorms. Thunderstorm cells are represented using germ-grain models from stochastic geometry, which are based on Cox or doubly-stochastic cluster processes. We present methods for the operational fitting of model parameters based on available point probabilities and thunderstorm records of past periods. Furthermore, we derive formulas for the computation of point and area probabilities according to the proposed germ-grain models. We also introduce a conditional simulation algorithm in order to increase the model's ability to precisely predict thunderstorm events. A systematic comparison of area probabilities, which are estimated from the proposed models, and thunderstorm records concludes the paper.

2000 AMS Mathematics Subject Classification: 60D05, 60G55, 86A10

Key words and phrases: Stochastic modeling, Cox process, Cluster process, Germ-grain model, Monte Carlo simulation, Thunderstorm cell

1. INTRODUCTION

1.1. Motivation. One of the major responsibilities of meteorological services such as Deutscher Wetterdienst (DWD) is the issuing of reliable weather forecasts for potentially harmful weather events. Classically, deterministic weather forecasts are derived from numerical models describing the atmosphere, which is referred to as numerical weather prediction (NWP) in literature, see e.g. [10]. Although numerical models still provide the basis of almost all operational weather forecasts, the application of a wide range of probabilistic postprocessing methods has become of growing interest during the last decades to improve forecast quality, see e.g. [22]. Calibrated probabilistic forecasts are related to single geographical locations (i.e., points, in mathematical terms). This includes, e.g., the probability for the occurrence of a weather event at some point (which is denoted as a point probability) or the expected value of a random variable describing the weather at some fixed location. Another approach is to consider probabilistic weather forecasts that are related to geographical regions (or areas, in mathematical terms). An example

is given using so-called area probabilities, which describe the chance of a weather event occurring somewhere in an area (rather than at a fixed point). Area probabilities can be of particular interest for the generation of automated warning guidances. While the computation of precise and reliable point forecasts is very well developed, there are hardly any methods available for the analytical derivation of further operationally applicable probabilistic forecasts (as, e.g., area probabilities). In [15] approaches for the computation of area probabilities based on point probabilities are proposed. However, the formulas derived in [15] are only valid under some restrictive assumptions, which make them inappropriate for the use in operational weather predictions on a non-local scale (e.g., the territory of Germany). A promising alternative is given by the development of spatial (or spatio-temporal) stochastic models for the considered weather events, which allow us to derive probabilistic forecasts based on repeated Monte-Carlo simulation.

1.2. Previous development. In the present paper we focus on the weather event ‘occurrence of thunderstorms’. Unfortunately, stochastic models for thunderstorms can rarely be found in literature. In [17] a model for wind loads in thunderstorms is proposed at single sites but no approaches for spatially continuous modeling of thunderstorm cells seem to be available. However, there exist a variety of spatial, temporal and spatio-temporal stochastic models for precipitation that have been used for applications in various fields of hydrology, see, e.g., [3]. A frequently used approach is to represent precipitation cells as circular or elliptical discs whose centers are modeled using spatial (or spatio-temporal) Poisson or cluster processes. Brief overviews of existing point-process-based models can be found, e.g., in [18] or [19]. Unfortunately, certain limitations prevent the use of these models in operational weather prediction. Most approaches assume temporal and spatial stationarity and model fitting procedures (mostly for large parameter sets) always rely on observed precipitation patterns derived from radar data. In operational weather forecasting, however, it is crucial to account for the (permanently changing) weather conditions in the current period and future forecast periods rather than for weekly, monthly or yearly averages, which contradicts the assumption of temporal stationarity. Model fitting based on radar observations is not suitable, either. On the one hand, radar observations from periods prior to the forecast period become inaccurate already after one hour. On the other hand, radar data for the (future) forecast period are not available at the time the forecast is made. Finally, forecasters are typically interested in stochastic models to be applied on a non-local scale, which does not allow for spatial stationarity assumptions due to different meteorological (areas of low or high pressure, changing over time) and geographical (plains or mountainous regions, not changing over time) conditions. Although not directly applicable in operational weather prediction, the mentioned papers still provide some valuable ideas for the approaches presented in this paper.

1.3. Outline. To overcome the limitations mentioned above, a new spatial stochastic model for precipitation cells has been developed recently, see [13] and

[14], which better fulfills the requirements of operational weather prediction. In that approach, precipitation cells occurring in a one-hour forecast period are represented by a germ-grain model with circular grains, which is based on a non-stationary Cox point process. In Sect. 2 we first want to find out if this kind of germ-grain model can also be applied for the modeling of thunderstorm cells. We introduce model construction, describe model fitting and give a validation using thunderstorm observations, where it turns out that the Cox germ-grain model is not suitable for thunderstorm cells. Therefore, in Sect. 3 we propose a more general and sophisticated germ-grain model for thunderstorm cells based on clustered point processes. We again introduce model construction, provide procedures for model fitting and describe a conditional simulation algorithm that provides a seamless transition from nowcasting based on observations to forecasting. A validation shows that this kind of approach is able to provide point and area probabilities that are much more reliable than those from previous methods.

2. MODELING THUNDERSTORM CELLS BASED ON COX PROCESSES

2.1. Underlying probability space. Before we describe a first approach for the spatial modeling of thunderstorm cells, a suitable mathematical framework needs to be specified. In the following, we always consider a fixed one-hour forecast period T that can be interpreted as some subinterval of the real line with a length of 60 minutes. With $T - d$ for some $d \geq 0$ we denote the one-hour time period that starts (and ends) d minutes earlier than T . In order to adequately describe an underlying probability space, we first need to briefly sketch the mechanics of commonly used weather forecast models. At DWD, e.g., a first probabilistic forecast is provided by the ensemble COSMO-DE-EPS¹. Such forecasts, however, are subject to both random and systematic errors, which result from uncertainties and inaccuracies in model specifications and boundary values due to discretization and parametrization. In order to eliminate systematic errors, a post-processing method, e.g., Model Output Statistics (MOS), can be applied, which provides statistically unbiased probabilistic forecasts, see [12] and [22]. These forecasts, however, are still subject to random errors and can thus be interpreted as estimators of the unknown future weather activity. To account for this, we introduce the probability space (Ω, \mathcal{F}, P) , where Ω is an abstract space, which describes all possible weather scenarios and the corresponding forecasts provided by the weather forecast models of DWD, \mathcal{F} denotes a suitable σ -algebra of subsets of Ω and P is some probability measure on (Ω, \mathcal{F}) . As mentioned above, probabilistic forecasts are subject to

¹Probabilistic forecasts are provided using an ensemble of 20 numerical forecasts of the COSMO-DE, see [2]. For the ensemble, the numerical model COSMO-DE is computed with a variation of model parameters and initial and boundary values, see [8]. From the distribution of the resulting forecasts various probabilistic characteristics as, e.g., event probabilities can be estimated.

a random error, which is modeled by the random variable $E : \Omega \rightarrow \mathbb{S}$ in the following, with \mathbb{S} being the (abstract) measurable space of all such possible errors. Heuristically speaking, conditioning on $\{E = e\}$ for some realization $e \in \mathbb{S}$ of E indicates that a specific forecast provided by the models of DWD (with error e) is given, which is always the case in applications.

2.2. Modeling of point probabilities. As indicated in Sect. 1.1 and 2.1, reliable, unbiased point forecasts including point probabilities for the occurrence of thunderstorms are derived by DWD using numerical models and statistical postprocessing. Since these forecasts are interpreted as estimators, we model point probabilities as random variables in the following. Let $\{P_t, t \in W\}$ be a random field in a bounded and convex observation window $W \subset \mathbb{R}^2$, where $P_t : \Omega \rightarrow [0, 1]$ denotes the random point probability for the occurrence of a thunderstorm at location $t \in W$ within the considered forecast period T . For each $t \in W$, we assume the random variable P_t to be $\sigma(E)$ -measurable, where $\sigma(E) \subset \mathcal{F}$ is the sub- σ -algebra of events generated by E . Accordingly, if conditioned on $\{E = e\}$ for any realization e of E , the value of P_t is non-random and only depends on e . In this case any realization of P_t can be identified by the conditional expectation $f(e) = \mathbb{E}(P_t | E = e)$. In general, point probabilities can be estimated by DWD for any location t inside the observation window W . In practice, however, this is only done for a finite number of sites (e.g., a network of weather stations or a regularly spaced lattice). For this purpose, we suppose that there is a sequence $s_1, \dots, s_n \in W$ of geographically distinct locations at which point probabilities $p_{s_1} = \mathbb{E}(P_{s_1} | E = e), \dots, p_{s_n} = \mathbb{E}(P_{s_n} | E = e)$ for a particular realization e of E are available. A fundamental assumption in the modeling of probabilities for the occurrence of thunderstorms is that there is a thunderstorm at some location $t \in W$ if and only if t is covered by at least one thunderstorm cell. Consequently, for a random set $M \subset W$, which can be considered a model for the union of all thunderstorm cells in W , the random point probability P_t is represented as $P_t = P(t \in M | E)$ for each $t \in W$. In order to give a more precise representation of point probabilities (and also further probabilistic characteristics such as, e.g., area probabilities) the random union set M of thunderstorm cells needs to be specified.

2.3. Modeling of thunderstorm cells. In a first attempt, we propose a model for thunderstorm cells based on spatial Cox processes. A similar approach has been applied successfully to the modeling of precipitation cells, see [13] and [14]. One major requirement for the application in operational weather prediction is spatial non-stationarity to account for geographical differences as well as locally varying weather conditions in the considered forecast period T . For this purpose, we suppose that thunderstorm cells, or more precisely their cell centers, occur in W according to a random intensity function $\{\Lambda_t, t \in W\}$, where $\Lambda_t : \Omega \rightarrow [0, \infty)$ is a non-negative random variable modeling the intensity for the occurrence of a thunderstorm at location $t \in W$. Analogous to the random point probabilities, Λ_t

is assumed to be $\sigma(E)$ -measurable for each $t \in W$. In order to find a closed representation of $\{\Lambda_t, t \in W\}$, which is crucial for model fitting and simulation, some simplification is necessary. Since it is assumed that estimated point probabilities are only available at the sites s_1, \dots, s_n , see Sect. 2.2, we suppose that, conditioned on $\{E = e\}$ for each $e \in \mathbb{S}$, the intensity function $\{\Lambda_t, t \in W\}$ is piecewise constant in certain neighborhoods of s_1, \dots, s_n . The most intuitive choice of such a neighborhood is the Voronoi tessellation $\{V(s_1), \dots, V(s_n)\}$ of s_1, \dots, s_n , where $V(s_i)$ is the Voronoi cell of s_i being defined as

$$V(s_i) = \{x \in W : \|x - s_i\| < \|x - s_j\| \text{ for all } j = 1, \dots, n \text{ with } j \neq i\}$$

for $i = 1, \dots, n$. With $\|x - s\|$ we denote the Euclidean distance of two locations $x, s \in W$. This implies the following representation of the intensity function $\{\Lambda_t, t \in W\}$:

$$\Lambda_t = \sum_{j=1}^n A_j \mathbb{1}_{V(s_j)}(t) \quad \text{for all } t \in \bigcup_{i=1}^n V(s_i),$$

where $\mathbb{1}_V : W \rightarrow \{0, 1\}$ denotes the indicator function of the set $V \subset W$ and $A_1, \dots, A_n : \Omega \rightarrow [0, \infty)$ are some non-negative, $\sigma(E)$ -measurable random variables that can be interpreted as random local intensities for the occurrence of a thunderstorm. If $t \in W$ is not located within any of the Voronoi cells, we set Λ_t to the minimum intensity of all adjacent Voronoi cells. Having specified the intensity of thunderstorm occurrence, a model for the centers of thunderstorm cells can be given. For the approach considered in this section we use a spatial Cox point process $\{X_i, i = 1, \dots, Z_X\}$ in W with random intensity function $\{\Lambda_t, t \in W\}$, see [4], where $Z_X : \Omega \rightarrow \{0, 1, \dots\}$ denotes the total number of thunderstorm cells in W . In particular, the random variable Z_X is almost surely finite. Clearly, the Cox process $\{X_i, i = 1, \dots, Z_X\}$ cannot be assumed to be $\sigma(E)$ -measurable because even given a specific realization of the weather forecast, the weather activity in the (future) forecast period T (which includes the occurrence of thunderstorms) is still considered to be random.

While there are well developed methods for the detection and identification of thunderstorm cell centers, see, e.g., [16], the determination of cell shapes and sizes is hardly possible. Thus, it is unclear how the shape of thunderstorm cells should be represented realistically. A reasonable simplifying approach is to model thunderstorm cells by circular discs with a common random radius $R : \Omega \rightarrow (0, \infty)$, which is assumed to be $\sigma(E)$ -measurable. This is equivalent to the assumption that there is a thunderstorm at some location $t \in W$ if t has a distance of not more than R km to at least one thunderstorm cell center. Accordingly, the random set M of thunderstorms, which was introduced in Sect. 2.2, is represented as a germ-grain model (see e.g. [1]), i.e.,

$$M = \bigcup_{i=1}^{Z_X} b(X_i, R),$$

with $b(x, r) = \{y \in \mathbb{R}^2 : \|y - x\| \leq r\}$ denoting the two-dimensional (closed) ball with center $x \in \mathbb{R}^2$ and radius $r > 0$. Note that the approach of modeling thunderstorm cells as circular discs is also used by DWD, see NowCastMIX data described in Sect. 2.5. Furthermore, circular or elliptical discs are often considered for the modeling of precipitation cells, see e.g. [18], [19] or Sect. 1.2. Analogous to the Cox process $\{X_i, i = 1, \dots, Z_X\}$ of cell centers, the germ-grain model M cannot be assumed to be $\sigma(E)$ -measurable.

2.4. Model fitting. Given $\{E = e\}$ for some $e \in \mathbb{S}$, the distribution of the germ-grain model M for the union of thunderstorm cells is completely characterized by the corresponding realizations of the random local intensities A_1, \dots, A_n for the occurrence of thunderstorms and the random cell radius R . We propose fitting methods to determine these model characteristics based on the available point probabilities P_{s_1}, \dots, P_{s_n} . In applications, fitting as well as simulation of the model is always performed in dependence of a particular realization of the weather forecast models of DWD. Thus, in the following we consider a fixed realization $e \in \mathbb{S}$ of the random error E and the corresponding realizations $p_{s_1} = \mathbb{E}(P_{s_1} | E = e), \dots, p_{s_n} = \mathbb{E}(P_{s_n} | E = e), a_1 = \mathbb{E}(A_1 | E = e), \dots, a_n = \mathbb{E}(A_n | E = e)$ and $r = \mathbb{E}(R | E = e)$. Note that conditioned on $\{E = e\}$ the point process $\{X_i, i = 1, \dots, Z_X\}$ is a Poisson process with deterministic intensity function $\{\lambda_t, t \in W\}$, where $\lambda_t = \mathbb{E}(\Lambda_t | E = e)$ for $t \in W$, and the germ-grain model M is a Boolean model, see [1] and [9]. At DWD, observed thunderstorm cells are modeled as circular discs with a constant radius of 10 km. In order to provide consistence with those observations in applications as performed in Sect. 2.5, we also set $r = 10$ km in the following². Thus, only the intensities a_1, \dots, a_n need to be estimated. For that purpose, we first derive a representation formula for area probabilities according to the proposed model. Let $\mathcal{B}(W)$ denote the Borel σ -algebra on W . For any $B \in \mathcal{B}(W)$, the area probability $\pi(B)$ for the occurrence of thunderstorms in B given $\{E = e\}$ is modeled as $\pi(B) = P(B \cap M \neq \emptyset | E = e)$. Using the distributional properties of Poisson processes, see [9], the following representation formula can be derived:

$$\begin{aligned} \pi(B) &= P(B \cap M \neq \emptyset | E = e) \\ &= 1 - P(\#\{i : X_i \in B \oplus b(o, R)\} = 0 | E = e) \\ &= 1 - \exp \left\{ - \int_{B \oplus b(o, r)} \sum_{i=1}^n a_i \mathbb{1}_{V(s_i)}(t) dt \right\} \\ &= 1 - \exp \left\{ - \sum_{i=1}^n a_i \int_{\mathbb{R}^2} \mathbb{1}_{(B \oplus b(o, r)) \cap V(s_i)}(t) dt \right\} \end{aligned}$$

²In [13] a statistical method has been presented to estimate the radius r based on the correlation structure of the available point probabilities p_{s_1}, \dots, p_{s_n} . This method has been applied successfully to the modeling and simulation of precipitation cells.

$$= 1 - \exp \left\{ - \sum_{i=1}^n a_i \nu_2((B \oplus b(o, r)) \cap V(s_i)) \right\},$$

where $o \in \mathbb{R}^2$ denotes the origin, $\nu_2(\cdot)$ is the two-dimensional Lebesgue measure, $\#C$ describes the cardinality of a countable set C and $A \oplus B = \{x + y, x \in A, y \in B\}$ denotes the Minkowski sum of two sets A and B . By setting $B = \{s_j\}$ for $j = 1, \dots, n$ we immediately get the following representation formula of the corresponding point probability p_{s_j} for the occurrence of thunderstorms at site s_j :

$$p_{s_j} = 1 - \exp \left\{ - \sum_{i=1}^n a_i \nu_2(b(s_j, r) \cap V(s_i)) \right\} \quad \text{for all } j = 1, \dots, n.$$

These equations can easily be rearranged to

$$\log \left(\frac{1}{1 - p_{s_j}} \right) = \sum_{i=1}^n a_i \nu_2(b(s_j, r) \cap V(s_i)) \quad \text{for all } j = 1, \dots, n,$$

which describes a system of n linear equations with the unknowns a_1, \dots, a_n . However, under the constraint that $a_1, \dots, a_n \geq 0$, this system of equations, in general, cannot be solved exactly. Therefore, a_1, \dots, a_n are computed by solving this system of equations in a non-negative least squares sense, i.e.,

$$(a_1, \dots, a_n) = \operatorname{argmin}_{a'_1, \dots, a'_n \geq 0} \left\{ \sum_{j=1}^n \left(\log \left(\frac{1}{1 - p_{s_j}} \right) - \sum_{i=1}^n a'_i \nu_2(b(s_j, r) \cap V(s_i)) \right)^2 \right\}.$$

2.5. Application and model validation. To conclude this chapter, a validation of the proposed Cox germ-grain model for thunderstorm cells is performed. We consider a convex observation window W containing the boundaries of Germany, and a regularly spaced lattice consisting of 1,575 locations $s_1, \dots, s_{1575} \in W$ together with point probabilities for the occurrence of thunderstorms at these locations for a sequence of 2,205 one-hour forecast periods covering the months May, June and July 2016 with forecast lead times of 1 to 3 hours ahead. The observation window W , the locations s_1, \dots, s_{1575} and the point probabilities $p_{s_1}, \dots, p_{s_{1575}}$ for the forecast period July 11, 2016, 15-16 UTC (Universal Time, Coordinated) are illustrated in Fig. 1 (top left), where the Voronoi cell $V(s_i)$ of each location $s_i \in \{s_1, \dots, s_{1575}\}$ is colored according to the point probability p_{s_i} . A typical realization of the fitted germ-grain model for the same period is also shown in Fig. 1 (bottom left). In order to analyze the quality of the developed model, we use it for the computation of area probabilities. A reasonable choice for test areas would be some partition of the observation window W . The Voronoi tessellation $\{V(s_1), \dots, V(s_{1575})\}$, however, is not appropriate since most Voronoi cells appear to be of same size and shape. To obtain test areas with varying shapes, sizes

and orientations, we generate a realization of a homogeneous Poisson point process in W , whose intensity is chosen such that the expected number of points in W is equal to 1,000. We obtained a realization with 999 points and use the cells of the corresponding Voronoi tessellation, denoted as $B_1 \dots, B_{999}$ in the following, as test areas. For each available forecast period the area probabilities $\pi(B_1), \dots, \pi(B_{999})$ are computed³ according to the formula stated in Sect. 2.4. In Fig. 1 (top right) the test areas B_1, \dots, B_{999} are shown, where each area B_i is colored according to the corresponding area probability $\pi(B_i)$ for the forecast period July 11, 2016, 15-16 UTC. Computed area probabilities correspond well with the underlying data showing higher values in regions with higher point probabilities and vice versa. Probabilities are not computed for some areas close to the boundary of the observation window to avoid edge effects.

In order to perform a systematic model validation, we use NowCastMIX data, which are based on radar and lightning⁴ observations. In NowCastMIX, centers of thunderstorms are first recorded and thunderstorm cells are then modeled as discs with a fixed radius of 10 km (which is the reason for choosing the same cell radius in Sect. 2.4). Additional thunderstorm characteristics such as movement speed, movement direction and hail flag (an indicator for thunderstorm strength taking the values 0, 1 or 2) are derived from radar reflectivities using several radar processing methods and are included in NowCastMIX data, see [11]. An example is shown in Fig. 1 (bottom right), where cells are colored according to an internal classification of DWD and additionally some warning cones are illustrated that show the possible movements of cells during the following hour. For each test area $B \in \{B_1, \dots, B_{999}\}$, we consider the sequence $\pi_1(B), \dots, \pi_{2205}(B)$ of area probabilities (for the 2,205 available forecast periods) and a sequence $I_1(B), \dots, I_{2205}(B)$ of thunderstorm indicators, which are 1 if there is a thunderstorm within B in the corresponding forecast period with respect to NowCastMIX data and 0 otherwise. In order to compare area probabilities and thunderstorm indicators, we analyze three score functions: the bias, the logarithmic skill score and the empirical correlation coefficient. Another common score for probabilistic weather forecasts is the Brier skill score, which is, however, less suitable for rare events such as the occurrence of thunderstorms. The definition of the bias (difference of mean area probability and mean thunderstorm indicator from NowCastMIX) and the correlation coefficient should be clear but the logarithmic skill score requires some more explanation, see also [22]. For each $j \in \{1, \dots, 2205\}$ the ignorance i_j of $\pi_j(B)$ is put to $-\log(\pi_j(B))$ if $I_j(B) = 1$ and to $-\log(1 - \pi_j(B))$ if $I_j(B) = 0$. Then, the logarithmic score ls is defined as the mean of i_1, \dots, i_{2205} ,

³Area probabilities can also be estimated based on repeated Monte Carlo simulation of the Cox germ-grain model, which turns out to be more efficient than direct computation in most cases.

⁴Lightning data is LINET provided by Nowcast GmbH.

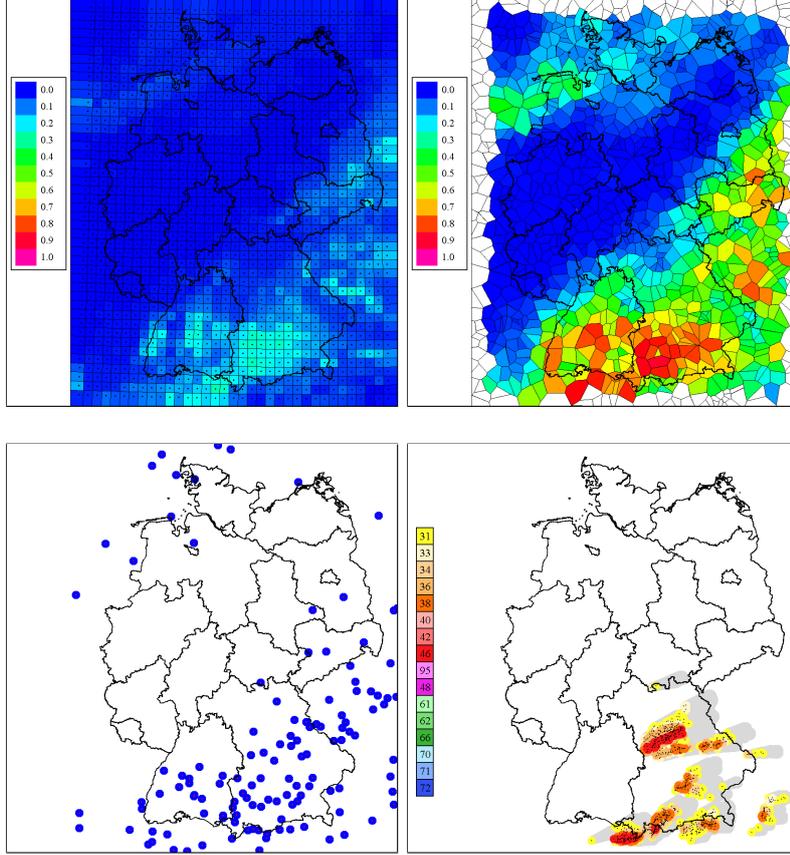


FIGURE 1. Examples of data and simulation results for July 11, 2016, 15-16 UTC: point probabilities at s_1, \dots, s_{1575} (top left), area probabilities for B_1, \dots, B_{999} computed using the Cox germ-grain model (top right), typical realization of the Cox germ-grain model (bottom left), thunderstorm observations from NowCastMIX (bottom right).

i.e., $ls = \frac{1}{2205} \sum_{j=1}^{2205} i_j$. Since the logarithmic score is difficult to compare for different forecast periods, it is related to the logarithmic score \tilde{ls} of a reference prediction, where usually $\tilde{\pi}_i(B) = \frac{1}{2205} \sum_{j=1}^{2205} I_j(B)$ for all $i = 1, \dots, 2205$ is used. Then, the logarithmic skill score lss is defined as $lss = 1 - ls/\tilde{ls}$. Of course, logarithmic scores of analyzed area probabilities should not be bigger than those of the (naive) reference prediction, which is why the logarithmic skill score is requested to be clearly positive. The three scores are computed and illustrated for all test areas that are not too close to the boundaries of W , see Fig. 2-4 (right), where each area is colored according to the value of the corresponding score function. Since the quality of computed area probabilities is expected to strongly depend on the precision of the underlying point probabilities, the same score functions are

also computed for the available data, see Fig. 2-4 (left), where each Voronoi cell $V(s_i)$ is colored according to the value of the score function at the corresponding location s_i . At some locations in the Northwest of the observation window no thunderstorms occurred during the entire period, which is why the corresponding Voronoi cells are left white. We find that available point probabilities do not seem to

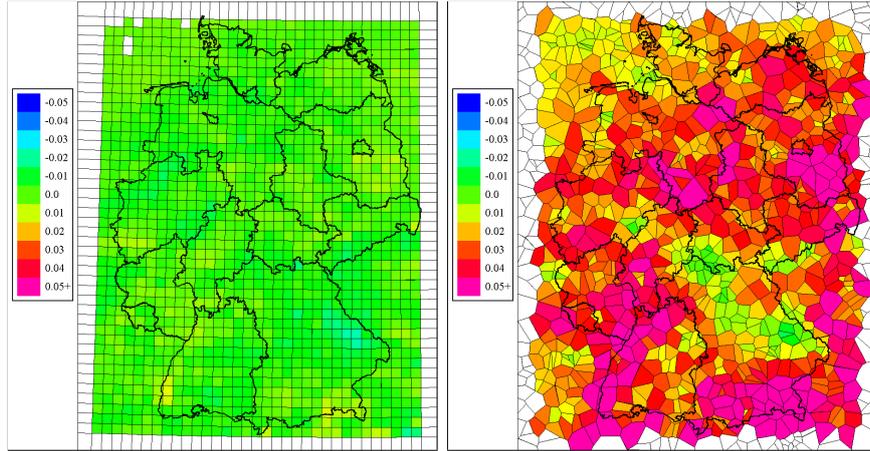


FIGURE 2. Biases of point probabilities from available data (left) and area probabilities computed using the Cox germ-grain model (right) against NowCastMIX data.

contain any systematic bias (mean bias is -0.2% , single values ranging between -2% and 1%) but only moderate logarithmic skill scores (mean value of 0.24 , most single values between 0 and 0.35) and correlation coefficients (mean value of 0.27 , most single values ranging from 0 to 0.4) are obtained. The results for the area probabilities are different. We get reasonably high correlation coefficients (significantly higher than for point probabilities with values between 0.1 and 0.6 for most areas), which shows that the proposed model indeed produces higher probabilities in periods and areas where thunderstorms occur than in those where thunderstorms do not occur. However, the biases show that area probabilities are systematically too high (the mean bias is 3% , single values reaching up to 7%), which makes the Cox germ-grain model inappropriate for applications to spatial stochastic modeling of thunderstorm cells. Logarithmic skill scores show slightly smaller values than for point probabilities (mean value of 0.2), with even having negative values for a few test areas.

3. MODELING THUNDERSTORM CELLS BASED ON CLUSTER PROCESSES

3.1. Model description. As shown in Sect. 2.5, the modeling of thunderstorm cells based on Cox processes is not appropriate. A probable reason for the model's failure to provide reliable area probabilities is that it generates thunderstorm cells

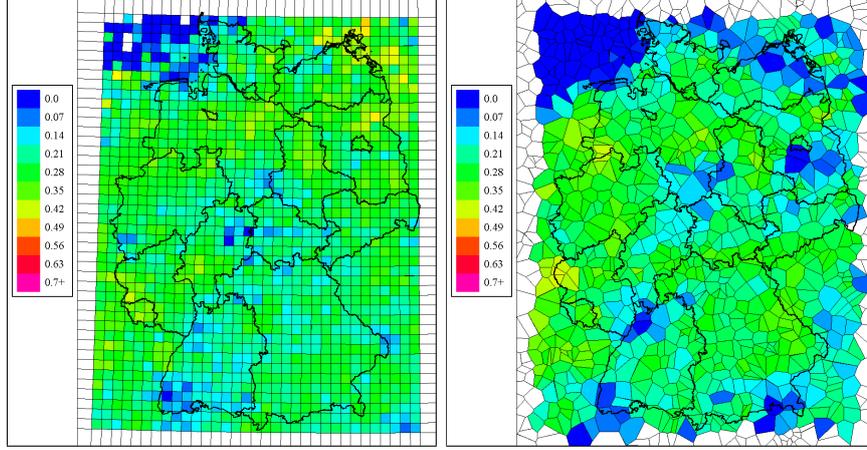


FIGURE 3. Logarithmic skill scores of point probabilities from available data (left) and area probabilities computed using the Cox germ-grain model (right) against NowCastMIX data.

independently of each other. Observed thunderstorms, however, seem to occur in clusters, see Fig. 1 (bottom right). Therefore, a different approach for the spatial modeling of thunderstorm cells is proposed in this section. In the following, we again consider a fixed one-hour forecast period T as well as the probability space (Ω, \mathcal{F}, P) and the random error E of the weather forecast models of DWD introduced in Sect. 2.1. Furthermore, $s_1, \dots, s_n \in W$ describe a sequence of locations in the observation window W at which realizations of the random $\sigma(E)$ -measurable point probabilities P_{s_1}, \dots, P_{s_n} for the occurrence of thunderstorms are available, see Sect. 2.2. Finally, $\{V(s_1), \dots, V(s_n)\}$ again denotes the Voronoi tessellation of s_1, \dots, s_n in W as introduced in Sect. 2.3. We start with the modeling of cluster centers. For that purpose, consider a spatial Cox process $\{Y_i, i = 1, \dots, Z_Y\}$ defined on (Ω, \mathcal{F}, P) with random intensity function $\{\Lambda_t^{(0)}, t \in W\}$ defined by

$$\Lambda_t^{(0)} = \sum_{j=1}^n A_j^{(0)} \mathbb{1}_{V(s_j)}(t) \quad \text{for all } t \in \bigcup_{i=1}^n V(s_i).$$

The $\sigma(E)$ -measurable random variables $A_1^{(0)}, \dots, A_n^{(0)} : \Omega \rightarrow [0, \infty)$ can be interpreted as local random intensities for the formation of a thunderstorm cluster. If $t \in W$ is located on the boundaries of one or more Voronoi cells (i.e., $t \notin \bigcup_{i=1}^n V(s_i)$), then $\Lambda_t^{(0)}$ is set to the minimum intensity of all adjacent Voronoi cells (like in the definition of $\{\Lambda_t, t \in W\}$ in Sect. 2.3). For the modeling of the clusters themselves some simplification is necessary again since the shapes of observed thunderstorm clusters vary significantly across space and time and can therefore hardly be determined. We suggest modeling these clusters as circular

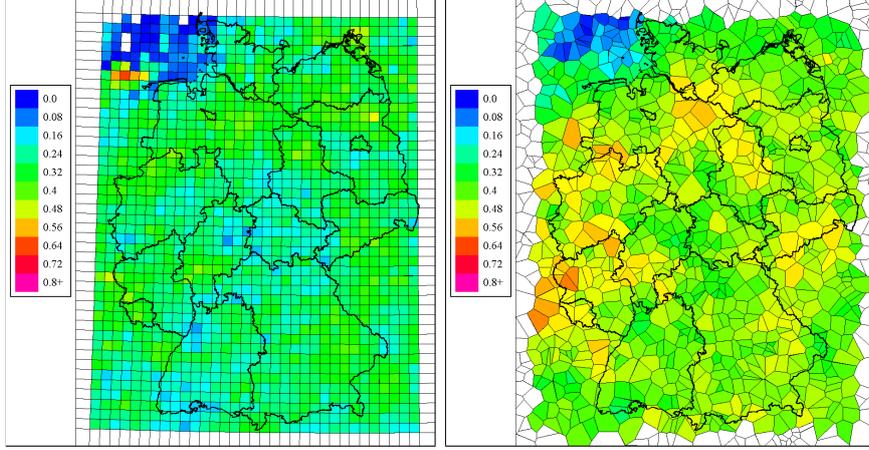


FIGURE 4. Empirical correlation coefficients of point probabilities from available data (left) and area probabilities computed using the Cox germ-grain model (right) against NowCastMIX data.

discs around the points of the Cox process $\{Y_i, i = 1, \dots, Z_Y\}$ as follows. The centers of thunderstorms occurring in each cluster are also described using Cox processes, where we assume that in different clusters they are (conditionally) independent of each other, given $\{E = e\}$. In order to specify this mathematically, we consider a sequence $\{X_i^{(1)}, i = 1, \dots, Z_{X^{(1)}}\}, \{X_i^{(2)}, i = 1, \dots, Z_{X^{(2)}}\}, \dots$ of identically distributed spatial Cox processes in W , which are (conditionally) independent of each other and of the point process $\{Y_i, i = 1, \dots, Z_Y\}$ of cluster centers given that $\{E = e\}$, and have a random intensity function $\{\Lambda_t^{(1)}, t \in W\}$ defined as

$$\Lambda_t^{(1)} = A^{(1)} \mathbb{1}_{b(o, R^{(1)})(t)} \quad \text{for all } t \in W.$$

Here, the $\sigma(E)$ -measurable random variable $A^{(1)} : \Omega \rightarrow [0, \infty)$ can be interpreted as random cluster intensity and the $\sigma(E)$ -measurable random variable $R^{(1)} : \Omega \rightarrow (0, \infty)$ describes the cluster radius. In order to give a proper representation of the process of all thunderstorm centers, we also consider the random counting measures $\{N_B^{(1)}, B \in \mathcal{B}(W)\}, \{N_B^{(2)}, B \in \mathcal{B}(W)\}, \dots$ that correspond to the Cox processes $\{X_i^{(1)}, i = 1, \dots, Z_{X^{(1)}}\}, \{X_i^{(2)}, i = 1, \dots, Z_{X^{(2)}}\}, \dots$, i.e., $N_B^{(j)} : \Omega \rightarrow \{0, 1, \dots\}$ with $N_B^{(j)} = \#\{i : X_i^{(j)} \in B\}$ for all $B \in \mathcal{B}(W)$ and $j \in \{1, 2, \dots\}$. Based on this, we introduce the random counting measure $\{N_B, B \in \mathcal{B}(W)\}$ defined as

$$N_B = \sum_{i=1}^{Z_Y} N_{(B-Y_i) \cap W}^{(i)} \quad \text{for all } B \in \mathcal{B}(W).$$

Then, there is a uniquely defined point process $\{X_i, i = 1, \dots, Z_X\}$ that satisfies $N_B = \#\{i : X_i \in B\}$ for all $B \in \mathcal{B}(W)$, see Lemma 9.1.XIII in [6], which is used

as a model for the centers of thunderstorm cells in W within the forecast period T . In literature, this kind of process is referred to as a (doubly-stochastic) cluster process, see, e.g., Chap. 6.3 in [5]. Due to the Cox process $\{Y_i, i = 1, \dots, Z_Y\}$ of cluster centers being non-stationary, the process $\{X_i, i = 1, \dots, Z_X\}$ of thunderstorm cell centers also has this property. Finally, the random closed set M that models the union set of thunderstorm cells is represented as a germ-grain model based on the cluster process $\{X_i, i = 1, \dots, Z_X\}$, i.e.,

$$M = \bigcup_{i=1}^{Z_X} b(X_i, R),$$

where the $\sigma(E)$ -measurable random variable $R : \Omega \rightarrow (0, \infty)$ describes the random radius of thunderstorm cells.

3.2. Model fitting: intensities of cluster centers. Since the germ-grain model M for the union of thunderstorm cells specified in Sect. 3.1 is based on a more complex point process model than that one considered in Sect. 2, there are also more model parameters that need to be determined. In our approach proposed in Sect. 3.1 for spatial stochastic modeling of thunderstorm cells, the conditional distribution of M given $\{E = e\}$ is completely characterized by the corresponding realizations of the random local intensities $A_1^{(0)}, \dots, A_n^{(0)}$ for the formation of thunderstorm clusters, the random cluster intensity $A^{(1)}$, the random cluster radius $R^{(1)}$ and the random cell radius R . Analogous to the model fitting described in Sect. 2.4, we consider in the following a fixed realization $e \in \mathbb{S}$ of the random error E occurring in the weather forecast models of DWD and the point probabilities $p_{s_1} = \mathbb{E}(P_{s_1} | E = e), \dots, p_{s_n} = \mathbb{E}(P_{s_n} | E = e)$ for the occurrence of thunderstorms at sites s_1, \dots, s_n . Furthermore, with $a_1^{(0)} = \mathbb{E}(A_1^{(0)} | E = e), \dots, a_n^{(0)} = \mathbb{E}(A_n^{(0)} | E = e)$, $a^{(1)} = \mathbb{E}(A^{(1)} | E = e)$, $r^{(1)} = \mathbb{E}(R^{(1)} | E = e)$ and $r = \mathbb{E}(R | E = e)$ the corresponding realizations of the model characteristics describing M are denoted. Conditioned on $\{E = e\}$, the point process $\{X_i, i = 1, \dots, Z_X\}$ of thunderstorm cell centers is a Neyman-Scott process, see, e.g., [4] and [20], with random intensity function $\{\Lambda_t, t \in W\}$ defined by

$$\Lambda_t = a^{(1)} \sum_{i=1}^{Z_Y} \mathbb{1}_{b(Y_i, r^{(1)})}(t) \quad \text{for all } t \in W.$$

Furthermore, given $\{E = e\}$, the point process $\{Y_i, i = 1, \dots, Z_Y\}$ is a Poisson process with intensity function $\{\lambda_t^{(0)}, t \in W\}$ defined by $\lambda_t^{(0)} = \mathbb{E}(\Lambda_t^{(0)} | E = e)$. In order to provide comparability of simulated thunderstorm cells with observed ones in applications as performed in Sect. 3.5, we set $r = 10$ km in the following. For the estimation of the intensity parameters $a_1^{(0)}, \dots, a_n^{(0)}$ we use a similar approach as in Sect. 2.4. At first, we derive a representation formula for area probabilities according to the proposed cluster model. Due to the properties of

$\{X_i, i = 1, \dots, Z_X\}$ being a Cox process and $\{Y_i, i = 1, \dots, Z_Y\}$ being a Poisson process conditioned on $\{E = e\}$, the area probability $\pi(B)$ for the occurrence of thunderstorms in $B \in \mathcal{B}(W)$ can be computed by

$$\begin{aligned}
\pi(B) &= P(B \cap M \neq \emptyset \mid E = e) \\
&= 1 - P(\#\{i : X_i \in B \oplus b(o, R)\} = 0 \mid E = e) \\
&= 1 - \mathbb{E} \left(\exp \left\{ - \int_{B \oplus b(o, r)} a^{(1)} \sum_{i=1}^{Z_Y} \mathbb{1}_{b(Y_i, r^{(1)})}(t) dt \right\} \right) \\
&= 1 - \mathbb{E} \left(\prod_{i=1}^{Z_Y} \exp \left\{ -a^{(1)} \nu_2 \left((B \oplus b(o, r)) \cap b(Y_i, r^{(1)}) \right) \right\} \right) \\
&= 1 - \exp \left\{ \int_{\mathbb{R}^2} \left(\exp \left\{ -a^{(1)} \nu_2 \left((B \oplus b(o, r)) \cap b(t, r^{(1)}) \right) \right\} - 1 \right) \lambda^{(0)}(t) dt \right\} \\
&= 1 - \exp \left\{ - \sum_{i=1}^n a_i^{(0)} \int_{V(s_i)} \left(1 - \exp \left\{ -a^{(1)} \nu_2 \left((B \oplus b(o, r)) \cap b(t, r^{(1)}) \right) \right\} \right) dt \right\},
\end{aligned}$$

where in the third equality we use the distributional properties of Cox processes, see [5], Sect. 6.2, and in the fifth equality a representation formula for the generating functional of Poisson processes is applied, which can be found in [20], Theorem 3.2.4. In particular, this implies a representation formula for point probabilities by setting $B = \{t\}$ for any $t \in W$. Accordingly, the intensities $a_1^{(0)}, \dots, a_n^{(0)}$ should satisfy

$$p_{s_j} = 1 - \exp \left\{ - \sum_{i=1}^n a_i^{(0)} I(s_i, s_j) \right\} \quad \text{for } j = 1, \dots, n$$

and for fixed $a^{(1)}, r^{(1)}$ and r , where

$$I(s_i, s_j) = \int_{V(s_i)} \left(1 - \exp \left\{ -a^{(1)} \nu_2 \left(b(s_j, r) \cap b(t, r^{(1)}) \right) \right\} \right) dt \quad \text{for } i, j = 1, \dots, n.$$

By rearranging the latter equation appropriately we obtain that

$$\log \left(\frac{1}{1 - p_{s_j}} \right) = \sum_{i=1}^n a_i^{(0)} I(s_i, s_j) \quad \text{for } j = 1, \dots, n,$$

which describes a system of n linear equations with unknowns $a_1^{(0)}, \dots, a_n^{(0)} \geq 0$. Due to this constraint, there is no exact solution of this system of equations in general, which is why we compute $a_1^{(0)}, \dots, a_n^{(0)}$ in a non-negative least squares sense according to

$$(a_1^{(0)}, \dots, a_n^{(0)}) = \operatorname{argmin}_{a'_1, \dots, a'_n \geq 0} \left\{ \sum_{j=1}^n \left(\log \left(\frac{1}{1 - p_{s_j}} \right) - \sum_{i=1}^n a'_i I(s_i, s_j) \right)^2 \right\}.$$

3.3. Model fitting: cluster intensity and cluster radius. It remains to determine the cluster intensity $a^{(1)}$ and the cluster radius $r^{(1)}$. Unfortunately, it does not seem that these parameters can also be estimated based on the available point probabilities. Thus, we suggest using the thunderstorm observations from NowCastMIX, see Sect. 2.5, for this purpose. In the presented fitting approach, clusters of thunderstorm cells are first identified using an established cluster algorithm and are then used to determine $a^{(1)}$ and $r^{(1)}$ according to the sizes and intensities of obtained clusters. Of course, when making a forecast for the one-hour forecast period T , the thunderstorm records for that period are not yet available. Thus, we consider the latest one-hour period prior to T for which records of NowCastMIX are available and which can be represented as $T - d$ with some $d \geq 60$ minutes. The best case ($d = 60$) is given when the forecast is made directly at the beginning of period T . While the total number and the positions of thunderstorm cells can change quickly over time, we observe that typical sizes of thunderstorm clusters and the number of storms per cluster only change gradually. Thus, we suppose that both $a^{(1)}$ and $r^{(1)}$ (for period T) can be estimated based on NowCastMIX records of period $T - d$. At first, a cluster analysis is performed to identify clusters of thunderstorm cell centers. For this purpose, we implemented the density-based spatial clustering of applications with noise (DBSCAN) algorithm, see [7]. This algorithm seems to be particularly suitable since it can recognize clusters of arbitrary shapes, it is possible to account for outliers (which are interpreted as noise) and the number of clusters to be found does not need to be known a priori (as required, e.g., in the k -means clustering algorithm). DBSCAN has two parameters: the maximum neighborhood radius ε and the minimum number $minPts$ that is required to form a cluster. Comparisons of results for different parameter configurations have shown that when applied to thunderstorm observations, $\varepsilon = 20$ km and $minPts = 3$ seem to be reasonable choices. An example for the NowCastMIX observations from July 11, 2016, 15-16 UTC is illustrated in Fig. 5, where all thunderstorm cells of a given cluster have the same color, compare also with Fig. 1 (bottom right). Let c_1, \dots, c_m denote the clusters detected by the DBSCAN algorithm. Next, in order to find cluster sizes, we determine for each cluster c_i the radius $r_i^{max} > 0$ of the smallest circle that contains all corresponding thunderstorm cell centers. For modeling clusters with an approximately circular shape, e.g., the cluster colored in light blue in Fig. 5, a disc with radius r_i^{max} seems suitable. For more elongated clusters, as, e.g., the green one in southern Germany, however, it is unlikely that the cluster can be represented by a disc with radius r_i^{max} . It seems more realistic that this cluster can be modeled by several circular discs with smaller radii that are located directly next to each other. To account for this, we determine for each cluster c_i the convex hull h_i of the set of all cell centers that belong to c_i and determine the radius $r_i^{min} > 0$ of the biggest circle that is completely contained in h_i as also illustrated in Fig. 5. Then, to each cluster c_i a radius $r_i > 0$ is assigned in dependence of the ratio of r_i^{min} and r_i^{max} according to the following algorithm.

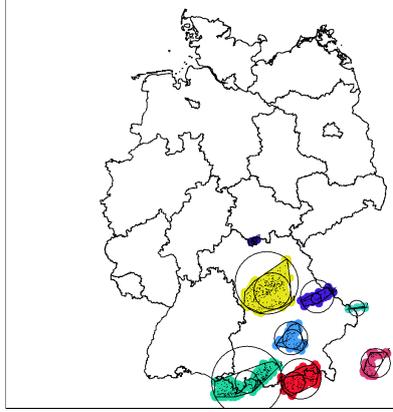


FIGURE 5. Results of the clustering algorithm for July 11, 2016, 15-16 UTC showing 8 identified clusters.

1. The minimum cluster radius is supposed to be equal to 10 km. This implies that if $r_i^{max} \leq 10$ km, then $r_i = 10$ km.
2. If $10 \text{ km} < r_i^{max} \leq 20$ km, then we always put $r_i = r_i^{max}$.
3. If $20 \text{ km} < r_i^{max} \leq 35$ km, then the following applies. If $r_i^{min} < 0.4 r_i^{max}$, we put $r_i = r_i^{min}$, otherwise $r_i = r_i^{max}$.
4. If $35 \text{ km} < r_i^{max} \leq 50$ km, then the following applies. If $r_i^{min} < 0.55 r_i^{max}$, we put $r_i = r_i^{min}$, otherwise $r_i = r_i^{max}$.
5. If $50 \text{ km} < r_i^{max}$, then the following applies. If $r_i^{min} < 0.65 r_i^{max}$, we put $r_i = r_i^{min}$, otherwise $r_i = r_i^{max}$.
6. The maximum cluster radius is assumed to be equal to 70 km. This implies that if $r_i > 70$ km, then r_i is reduced to 70 km.

Finally, in order to find the typical cluster radius $r^{(1)}$, we compute the mean value of the individual cluster radii r_1, \dots, r_m , i.e., $r^{(1)} = \frac{1}{m} \sum_{i=1}^m r_i$.

To conclude model fitting, the cluster intensity $a^{(1)}$ has to be estimated. Let k denote the total number of thunderstorms contained in all clusters c_1, \dots, c_m . For all $i = 1, \dots, m$ we determine the minimal number l_i of discs with radius $r^{(1)}$ that is needed to cover all thunderstorm cell centers in cluster c_i . The sum $l = l_1 + \dots + l_m$ can then be interpreted as the total number of circular clusters with radius $r^{(1)}$ and the ratio k/l denotes the mean number of thunderstorm cells per cluster. Accordingly, the intensity $a^{(1)}$ can be computed as $a^{(1)} = k/(l \pi (r^{(1)})^2)$. In periods with very weak or no thunderstorm activity it may happen that no thunderstorms are observed or that no clusters are detected by the DBSCAN algorithm (i.e., all thunderstorm cells are considered to be noise) in period $T - d$. In these cases, we recommend to put $a^{(1)} = 4/(\pi (r^{(1)})^2)$ and $r^{(1)} = 11$ km. These are the mean values of $a^{(1)}$ and $r^{(1)}$ from all one-hour periods with no detected thunderstorm cluster in the previous period according to the dataset considered in Sect. 3.5.

3.4. Conditional simulation of thunderstorm cells. In this section we suggest realizing the cluster-based germ-grain model of thunderstorm cells conditionally with respect to previously detected thunderstorm cells from the NowCastMIX data. Using this additional data can further increase forecast quality. In order to do so, area probabilities of thunderstorms are estimated based on repeated (conditional) Monte Carlo simulation of the underlying germ-grain model M instead of being computed according to the formula derived in Sect. 3.2. Let T again denote the one-hour time period for which the forecast is made and $T - d$ with $d \geq 60$ minutes is the latest one-hour period for which thunderstorm observations from NowCastMIX are available. With $x_1, \dots, x_p \in W$ we denote the centers of all thunderstorm cells recorded by NowCastMIX for period $T - d$. If d is small, then there is a chance that some of these cells still exist (with changed positions) during period T . For that purpose, the distribution of lifetimes of thunderstorm cells has been estimated, which depends on the thunderstorm's hail flag given in the NowCastMIX data, see Fig. 6 and [21]. At first, the total lifetime of each cell

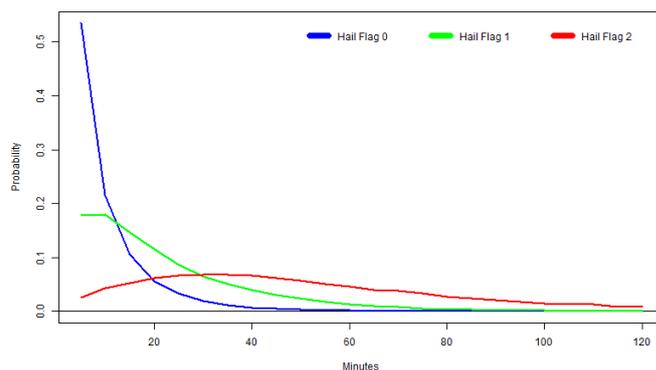


FIGURE 6. Distribution of the lifetimes of thunderstorm cells for different hail flags.

x_i is generated based on the estimated distributions (in dependence of the storms hail flag). Then, the remaining lifetime of x_i (from the time it was observed until its death) is simulated by multiplying the total lifetime with a realization of a standard uniformly distributed random variable. Knowing the exact time when x_i was observed, we can now easily determine whether x_i still exists during period T or not. Let $\{\tilde{x}_1, \dots, \tilde{x}_q\} \subset \{x_1, \dots, x_p\}$ with $q \leq p$ denote all thunderstorm cell centers from period $T - d$ that still exist in period T . As $\tilde{x}_1, \dots, \tilde{x}_q$ represent their positions in the interval $T - d$, the random movements of these cells have to be simulated next. Since NowCastMIX also provides the movement speed and the movement direction at the time of observation, we can determine the area of all possible positions \tilde{x}_i can have between the beginning of period T and its death (similar to the warning cones in Fig. 1, bottom right). This area is computed using a propagation angle of 7.5° in NowCastMIX, making it a triangle or a trapezoid,

within which we uniformly generate the new location $y_i \in W$ of the i -th surviving cell. Thus, $\{y_1, \dots, y_q\}$ can be interpreted as a possible set of thunderstorm cell centers that were observed in period $T - d$ (at different locations) and still exist in the forecast period T .

We propose the following algorithm to generate a realization of the germ-grain model M given the surviving cell centers y_1, \dots, y_q .

1. Let $U = \bigcup_{i=1}^q b(y_i, r^{(1)})$. Compute the expected numbers $\lambda_{in} = \int_U \lambda_t^{(0)} dt$ and $\lambda_{out} = \int_{W \setminus U} \lambda_t^{(0)} dt$ of cluster centers inside and outside of U , respectively.

2. Generate two realizations x_{in} and x_{out} of two Poisson distributed random variables with parameters λ_{in} and λ_{out} , respectively.

3. If possible, simulate x_{in} cluster centers inside U independently according to the intensity function $\{\lambda_t^{(0)}, t \in U\}$ under the condition that each of the discs $b(y_1, r^{(1)}), \dots, b(y_q, r^{(1)})$ contains at least one cluster center. If this is not possible (i.e., if x_{in} is too small), generate more than x_{in} cluster centers (as many as necessary) but reduce x_{out} accordingly by the additional number of simulated centers⁵.

4. Simulate x_{out} cluster centers outside U independently according to the intensity function $\{\lambda_t^{(0)}, t \in W \setminus U\}$. If $x_{out} \leq 0$ due to a possible reduction in step 3, then skip this step and go to step 5.

5. Put a disc with radius $r^{(1)}$ around all cluster centers generated in steps 3 and 4 in order to generate the cluster discs.

6. Generate a realization x' of a Poisson distributed random variable with parameter $(x_{in} + x_{out})a^{(1)}\pi(r^{(1)})^2$. Put $x = x' - q$, which can be interpreted as the number of thunderstorm cell centers to be simulated.

7. Repeat x times the following. Choose one cluster generated in step 3 or 4 at random and generate a uniformly distributed thunderstorm cell center in the corresponding cluster disc.

8. Put a disc with radius r around each point generated in step 7 and around each remaining cell center y_i for $i = 1, \dots, q$. The union of all these discs can be interpreted as a realization of M under the conditions that

- (i) the realization contains the thunderstorm cells with centers y_1, \dots, y_q and
- (ii) the expected number of generated thunderstorm cells is not changed compared to unconditional simulation.

The entire procedure described in this section can also be applied using not only observed thunderstorm cells from period $T - d$ but also from earlier periods (depending on how big d is chosen) since thunderstorms, in particular those with hail flag 2, have a good chance to exist two hours or even longer.

⁵According to the properties of Poisson point processes, the number of cluster centers $\#\{i : Y_i \in U\}$ in U and the number of cluster centers $\#\{i : Y_i \in W \setminus U\}$ outside U should be independent random variables. However, if x_{out} is not reduced accordingly if more than x_{in} cluster centers are needed to get one of them in each disc $b(y_i, r^{(1)})$ for $i = 1, \dots, q$, we will generate too many clusters on average. This will introduce a model bias that leads to significantly too high area probabilities in applications.

3.5. Application and model validation. In this section we perform a validation of the proposed cluster-based germ-grain model for thunderstorm cells. For this purpose, we again consider the forecast periods, point probabilities, test areas and thunderstorm records from NowCastMIX that were introduced in Sect. 2.5. When estimating the cluster parameters as described in Sect. 3.3 and using the conditional simulation algorithm stated in Sect. 3.4, we assume that forecasts are always made directly at the beginning of the corresponding forecast period and that NowCastMIX data are available for the one-hour period before the forecasting period, which means that $d = 60$ minutes. When applying conditional simulation, NowCastMIX data is used for the condition and for validation. However, for the condition only data prior to the start of the forecast period is used, whereas for validation data is considered that is not available when the forecast is made. In Fig. 7 two typical realizations of the germ-grain model M are shown. The image on the left-hand side shows a realization of M as described in Sect. 3.1, whereas for the image on the right-hand side the conditional simulation algorithm stated in Sect. 3.4 is used. Both realizations resemble the corresponding thunderstorm observations shown in Fig. 1 (bottom right) much better than the realization of the Cox germ-grain model in Fig. 1 (bottom left), with the conditionally simulated realization being even more similar to the NowCastMIX data. Also when comparing area probabil-

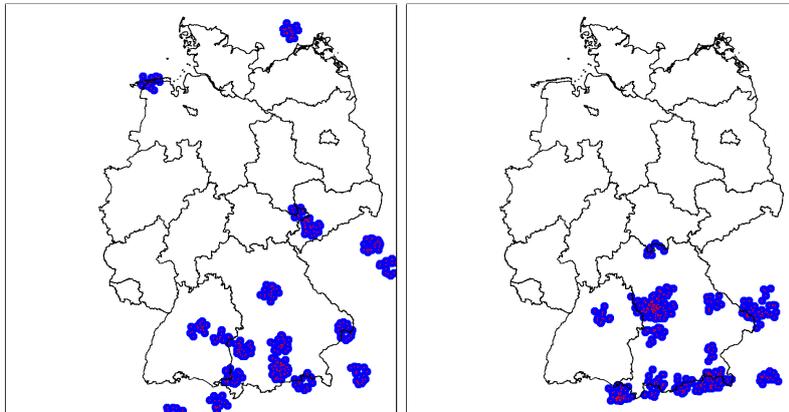


FIGURE 7. Typical realizations of the cluster-based germ-grain model for July 11, 2016, 15-16 UTC generated by unconditional simulation (left) and conditional simulation using NowCastMIX data (right).

ities, we observe significant differences among the various approaches considered in this paper. Fig. 8 illustrates the test areas B_1, \dots, B_{999} colored according to the corresponding area probabilities $\pi(B_1), \dots, \pi(B_{999})$ for the forecast period July 11, 2016, 15-16 UTC, where all area probabilities are computed by means of the formula derived in Sect. 3.2. We observe that the area probabilities in Fig. 8 correspond well to the underlying point probabilities, see Fig. 1 (top left), and also

resemble the area probabilities in Fig. 1 (top right) with the only difference that they are considerably lower now. Estimation of point and area probabilities based on conditional Monte Carlo simulation of the proposed model leads to clearly different results, compare Fig. 9 (left) with Fig. 1 (top left) and Fig. 9 (right) with Fig. 8. Both point and area probabilities look much sharper now, forecasting a high probability for a few points/areas, instead of low or medium probabilities in the entire South and East of Germany. We also observe that the estimated probabilities correspond particularly well with the thunderstorm observations for this period, compared to Fig. 1 (bottom right).

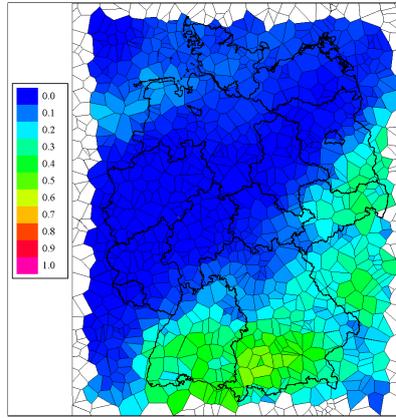


FIGURE 8. Area probabilities of B_1, \dots, B_{999} computed (unconditionally) according to the cluster-based germ-grain model for July 11, 2016, 15-16 UTC.

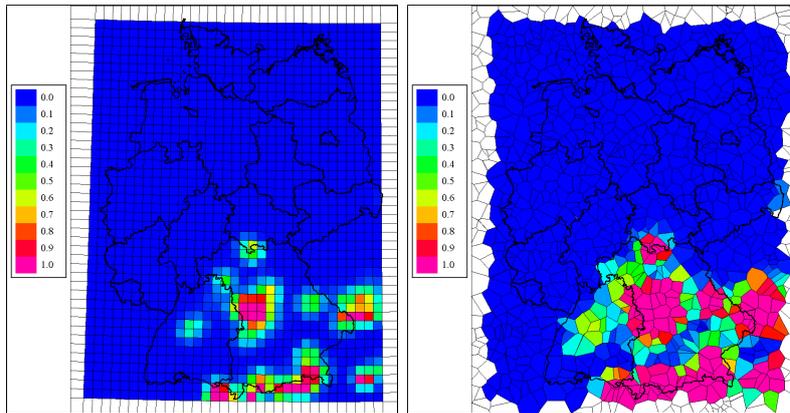


FIGURE 9. Thunderstorm probabilities for July 11, 2016, 15-16 UTC estimated by conditional simulation of the cluster-based germ-grain model using NowCastMIX data: point probabilities of s_1, \dots, s_{1575} (left), area probabilities of B_1, \dots, B_{999} (right).

To conclude this example of our application and to formally show the perfor-

mance of the model, we compute the biases, logarithmic skill scores and empirical correlation coefficients of estimated probabilities and observed thunderstorm indicators. In Fig. 10 the test areas B_1, \dots, B_{999} are colored according to the corresponding scores for area probabilities computed by means of the formula derived in Sect. 3.2. The results reveal that the cluster-based germ-grain model is much more suitable for the modeling of thunderstorm cells than the Cox germ-grain model considered in Sect. 2. Compared to Fig. 2 (right), biases were reduced drastically (the mean bias is -0.4 %, single values are between -3 % and 2 %). The lowest and highest biases occur in those regions, where similar biases are also present in the underlying point probabilities, see Fig. 2 (left). Logarithmic skill scores are positive for all test areas over land (mean value of 0.28, single values ranging between 0.15 and 0.4) and thus clearly higher than for the Cox germ-grain model of Sect. 2, see Fig. 3 (right). Correlation coefficients have similar values as for the Cox germ-grain model (mean value of 0.4, most single values between 0.25 and 0.6), compare to Fig. 4 (right). Finally, we observe that forecast quality can be improved significantly when point and area probabilities are estimated based on conditional simulation of the cluster-based germ-grain model. In Fig. 11 and 12 the considered score functions for point and area probabilities are shown. When comparing estimated point probabilities with those probabilities from the available data, see Fig. 2-4 (left), we find that logarithmic skill scores (mean value 0.51, single values reaching up to 0.7) and correlation coefficients (mean value 0.63, single values reaching up to 0.8) are considerably higher when using conditional simulation. Furthermore, no model bias is introduced. The results for area probabilities are even more convincing. A comparison of scores with those in Fig. 10 shows drastically increased logarithmic skill scores (mean value of 0.49, single values up to 0.65) and correlation coefficients (mean value of 0.71, almost all single values between 0.55 and 0.8) together with a mean bias of less than 1%. This shows impressively that estimation of both point and area probabilities (conditioned on past thunderstorm observations) according to the method presented in this section has a very high precision and thus is a valuable tool in the forecasting of thunderstorm events for short lead times.

4. CONCLUSION

In the present paper we considered two different approaches for spatial stochastic modeling of thunderstorm cells with the purpose of estimating point and area probabilities for the occurrence of thunderstorms. In both approaches thunderstorm cells were represented by germ-grain models that are based on different types of spatial point processes. For the developed models, formulas for the computation of point and area probabilities were derived. The first approach, where cell centers were modeled using Cox processes, turned out to be inappropriate, due to introducing systematic biases to computed area probabilities. The second approach in-

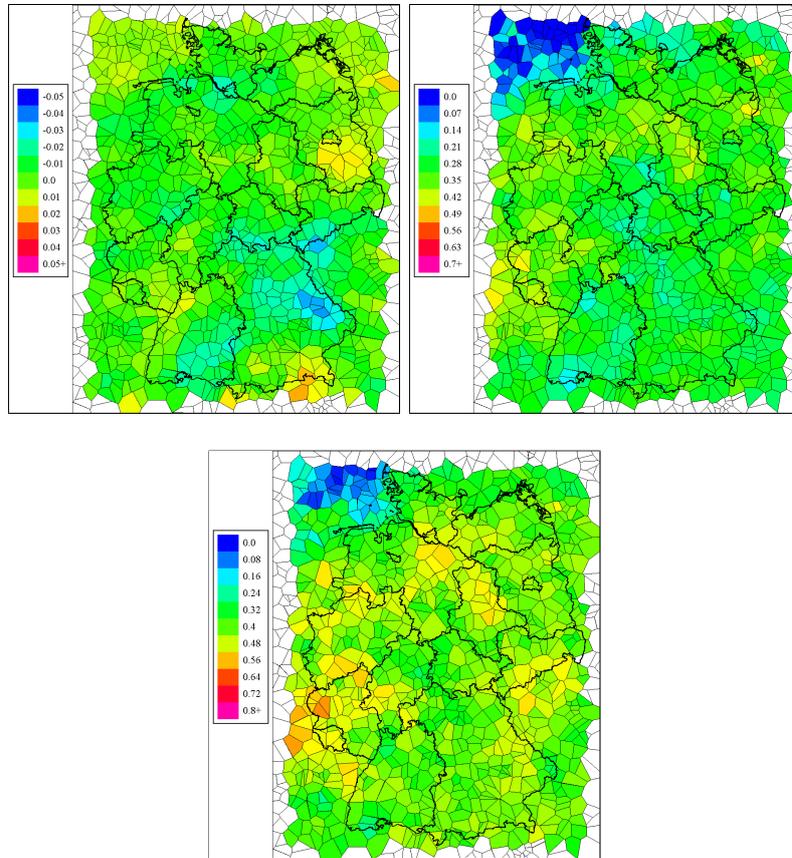


FIGURE 10. Comparison of area probabilities computed according to the cluster-based germ-grain model with thunderstorm indicators from NowCastMIX: biases (top left), logarithmic skill scores (top right) and empirical correlation coefficients (bottom).

volved the utilization of spatial doubly-stochastic cluster processes for modeling of thunderstorm cell centers. In contrast to the Cox germ-grain model, which can be fitted relying solely on available point probabilities, the cluster-based approach also requires thunderstorm records of past periods to statistically determine all model parameters. The benefit of using this more complex model is its ability to provide reliable area probabilities, which perform even better (according to computed validation scores) than the underlying point probabilities. The forecast quality can be considerably increased even more if realizations of the model are generated conditionally on thunderstorm records from past periods leading to a very high correspondence of simulated model realizations and thunderstorm observations for short forecast lead times. In this way, nowcasting data based on radar and lightning observations is seamlessly combined with forecast probabilities from numerical

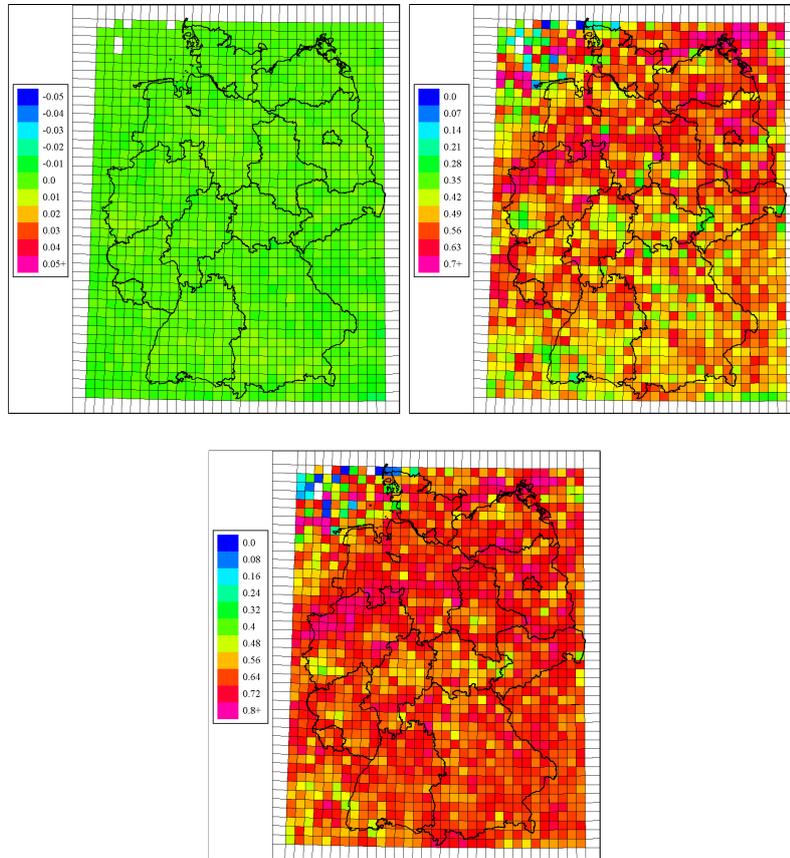


FIGURE 11. Comparison of point probabilities estimated using conditional simulation of the cluster-based germ-grain model with thunderstorm indicators from NowCastMIX: biases (top left), logarithmic skill scores (top right) and empirical correlation coefficients (bottom).

models and statistical postprocessing. The model does neither assume spatial nor temporal stationarity and model fitting does not require any observations from the period the forecasts are made for. This, together with the reasonable computation time, makes the model suitable for applications in operational weather prediction. A certain limitation is that the model is designed for short lead times only (up to six hours maximum), whereas the conditional simulation algorithm only makes sense for lead times up to two hours. A possible topic of future work could be, e.g., the incorporation of elliptic clusters, spatially varying cluster parameters or even the integration of precipitation produced by single thunderstorm cells in order to provide meteorologically more realistic thunderstorm realizations.

Acknowledgments. The research presented in this paper has been supported by the DWD project 2015EMF-10. The authors would like to thank Paul James,

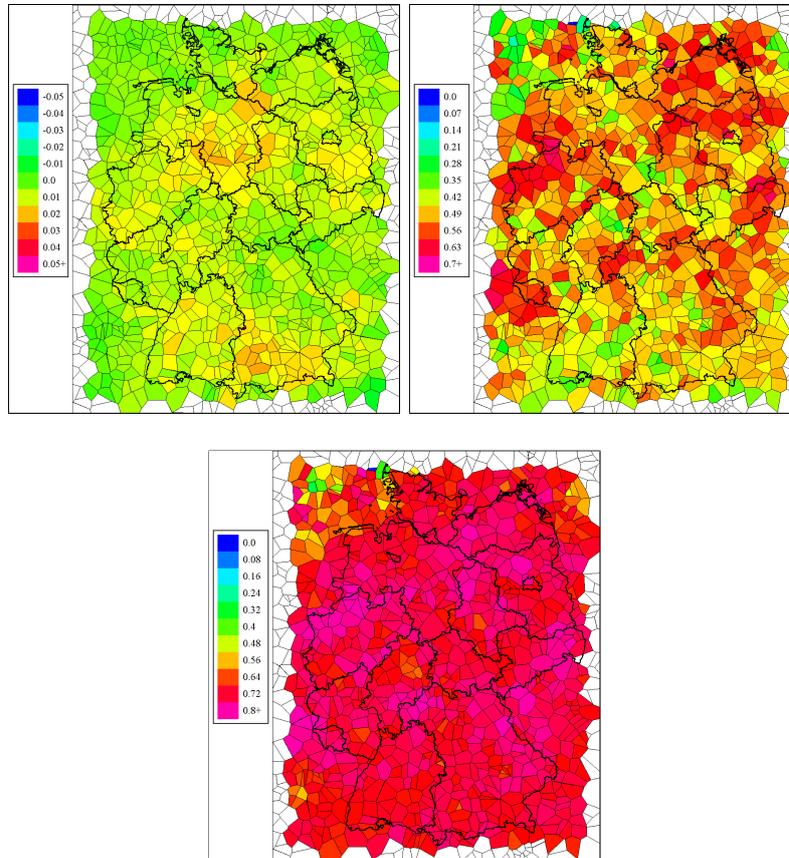


FIGURE 12. Comparison of area probabilities estimated using conditional simulation of the cluster-based germ-grain model with thunderstorm indicators from NowCastMIX: biases (top left), logarithmic skill scores (top right) and empirical correlation coefficients (bottom).

Kathrin Wapler, Bernhard Reichert, Andreas Veira, Jan Becker and Sebastian Trepte from DWD for providing expertise that greatly assisted the performed research and for participating in fruitful discussions.

REFERENCES

- [1] A. Baddeley, I. Bárány, R. Schneider and W. Weil *Stochastic Geometry*, Springer, Berlin, 2007.
- [2] M. Baldauf, A. Seifert, J. Förstner, D. Majewski, M. Raschendorfer and T. Reinhardt, *Operational convective-scale numerical weather prediction with the COSMO model: description and sensitivities*, Mon. Weather Rev. 139-12 (2011), pp. 3887–3905.
- [3] A. Burton, C.G. Kilsby, H.J. Fowler, P.S.P. Cowpertwait and P.E. O’Connell, *RainSim: A spatiotemporal stochastic rainfall modelling system*, Environ. Modell. Softw. 23-12 (2008), pp. 1356–1369.

-
- [4] S.N. Chiu, D. Stoyan, W.S. Kendall and J. Mecke *Stochastic Geometry and its Applications, 3rd ed.*, J. Wiley & Sons, Chichester, 2013.
- [5] D.J. Daley and D. Vere-Jones *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods, 2nd ed.*, Springer, New York, 2003.
- [6] D.J. Daley and D. Vere-Jones *An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure, 2nd ed.*, Springer, New York, 2008.
- [7] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, Proc. Second Int. Conf. Knowl. Discov. Data Min. (1996), Portland, pp. 226–231.
- [8] C. Gebhardt, S.E. Theis, M. Paulat and Z.B. Bouallègue, *Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries*, Atmos. Res. 100 (2011), pp. 168–177.
- [9] J. Illian, A. Penttinen, D. Stoyan and H. Stoyan *Statistical Analysis and Modelling of Spatial Point Patterns*, J. Wiley & Sons, Chichester, 2008.
- [10] P.M. Inness and S. Dorling *Operational Weather Forecasting*, J. Wiley & Sons, Chichester, 2013.
- [11] P. James, B.K. Reichert and D. Heizenreder *NowCastMIX Optimized automatic warnings from continuously monitored nowcasting systems based on fuzzy-logic evaluations of storm attributes*, 8th Eur. Conf. Severe Storms (2015), Vienna.
- [12] K. Knüpfper, *Methodical and predictability aspects of MOS systems*, Proc. 13th Conf. Probab. Stat. Atmos. Sci. (1996), San Francisco, pp. 190–197.
- [13] B. Kriesche, R. Hess, B.K. Reichert and V. Schmidt, *A probabilistic approach to the prediction of area weather events, applied to precipitation*, Spat. Stat. 12 (2015), pp. 15–30.
- [14] B. Kriesche, A. Koubek, Z. Pawlas, V. Beneš, R. Hess and V. Schmidt, *On the computation of area probabilities based on a spatial stochastic model for precipitation cells and precipitation amounts*, Stoch. Environ. Res. Risk Assess. (2016) doi:10.1007/s00477-016-1321-8.
- [15] R. Krzysztofowicz, *Point-to-area rescaling of probabilistic quantitative precipitation forecasts*, J. Appl. Meteorol. 38 (1998), pp. 786–796.
- [16] P. Lang, *Cell tracking and warning indicators derived from operational radar products*, Proc. 30th Int. Conf. Radar Meteorol. (2001), Munich, pp. 207–211.
- [17] C.Q. Li, *A stochastic model of severe thunderstorms for transmission line design*, Probabilist. Eng. Mech. 15-4 (2000), pp. 359–364.
- [18] C.J. Onof, R.E. Chandler, A. Kakou, P.J. Northrop, H.S. Wheeler and V.S. Isham, *Rainfall modelling using Poisson-cluster processes: a review of developments*, Stoch. Environ. Res. Risk Assess. 14-6 (2000), pp. 384–411.
- [19] A. Paschalis, P. Molnar, S. Fatichi and P. Burlando, *A stochastic model for high-resolution space-time precipitation simulation*, Water Resour. Res. 49-12 (2013), pp. 8400–8417.
- [20] R. Schneider and W. Weil *Stochastic and Integral Geometry*, Springer, Berlin, 2008.
- [21] K. Wapler, *The life-cycle of hail storms: lightning, radar reflectivity and rotation characteristics*, Atmos. Res. (submitted).
- [22] D.S. Wilks *Statistical Methods in the Atmospheric Sciences, 3rd ed.*, Academic Press, San Diego, 2011.

Ulm University, Institute of Stochastics
Helmholtzstr. 18, 89069 Ulm, Germany
E-mail: bjoern.kriesche@uni-ulm.de

Deutscher Wetterdienst, Research and
Development
Frankfurter Str. 135, 63067 Offenbach
E-mail: Reinhold.Hess@dwd.de

Ulm University, Institute of Stochastics
Helmholtzstr. 18, 89069 Ulm, Germany
E-mail: volker.schmidt@uni-ulm.de

Received on xx.xx.xxxx;
revised version on xx.xx.xxxx
