UNIVERSITÀ DEGLI STUDI DI BERGAMO
DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE
E METODI MATEMATICI°

QUADERNI DEL DIPARTIMENTO

Department of Information Technology and Mathematical Methods

Working Paper

Series "*Mathematics and Statistics*"

n. 3/MS – 2011

Dynamical Random-Set Modeling of Concentrated Precipitation in North America

by

N. Cressie, R. Assunção, S. H. Holan, M. Levine, O. Nicolis, J. Zhang, J. Zou

° Viale Marconi. 5, I – 24044 Dalmine (BG), ITALY, Tel. +39-035-2052339; Fax. +39-035-562779

[§] L'accesso alle *Series* è approvato dal Comitato di Redazione. I *Working Papers* della Collana dei Quaderni del Dipartimento di Ingegneria dell'Informazione e Metodi Matematici costituiscono un servizio atto a fornire la tempestiva divulgazione dei risultati dell'attività di ricerca, siano essi in forma provvisoria o definitiva.

Dynamical Random-Set Modeling of Concentrated Precipitation in North America

by

Noel Cressie (The Ohio State University),

Renato Assunção (Universidade Federal de Minas Gerais),

Scott H. Holan (University of Missouri),

Michael Levine (Purdue University),

Orietta Nicolis (University of Bergamo),

Jun Zhang (SAMSI, CICS-NC),

Jian Zou (NISS)

March 18, 2011

Abstract

In order to study climate at scales where policy decisions can be made, regional climate models (RCMs) have been developed with much finer resolution (∼50 km) than the ∼500 km resolution of atmosphere-ocean general circulation models (AOGCMs). The North American Regional Climate Change Assessment Program (NARCCAP) is an international program that provides 50-km resolution climate output for the United States, Canada, and northern Mexico.

In Phase I, there are six RCMs, from which we choose one to illustrate our methodology. The RCMs are updated every 3 hours and contain a number of variables, including temperature, precipitation, wind speed, wind direction, and air pressure; output is available from 1968–2000 and from 2038–2070. Precipitation is of particular interest to climate scientists, but it can be difficult to study because of its patchy nature: At hourly-up-to-monthly time scales, there are generally many zeroes over the precipitation field. In this research, we study sets of concentrated precipitation (i.e., the union of RCM pixels whose precipitation is above a given threshold), where we are interested in the way these sets evolve from one 3-hour period to the next. Assuming the sets are a realization of a time series of random sets, we are able to build dynamical models for the passage of rainfall fronts over 1-2 days. The dynamics are characterized by a growth/recession model for a time series of random sets, with several parameters that control how the concentrated precipitation changes over time.

Key words: Boolean model; Kernel density estimation; Laslett's theorem; Method-of-moments estimation; NARCCAP; Regional climate model (RCM); Set-valued autoregression (SVAR)

1 Introduction

Although temperature is a widely studied climate variable, for both paleoclimate reconstruction and climate-model projections, precipitation is equally, if not more, important. Eventually, water is expected to be a limiting factor for communities around the globe; as such, the prospect of a drier world is a cause of great concern. Water storage and conservation involve planning and decision-making that are typically made by governments, but the implementation of those plans involve local geography, farming practices, and the distribution of population centers. For such decisions,

reliable predictions of precipitation occurrence and amount are extremely important.

Recently, climate models have become a major tool for understanding climate change and its potential impact, especially due to their wide spatial and temporal coverage. The atmosphere-ocean general circulation models (AOGCMs) have been developed to simulate climate over the entire globe. That is, an atmospheric and oceanic model are linked to generate outputs, typically, on a course scale of 200 to 500 km.

Unfortunately, climate-model projections from global climate models (GCMs) are not useful for describing local climate effects, since they provide limited information at local scales where natural-resource management and environmental policy decisions are made. In contrast, regional climate models (RCMs) with scales on the order of 20–50 km have been developed that are able to account for fine-scale spatial variability. In fact, over the past several years, the RCM (e.g., [16]) has become a widely used tool in downscaling results to a regional scale. While studies of AOGCMs are mainly focused on the analysis of climate change ([3, 45, 47, 48]), RCMs are mainly aimed at analyzing uncertainties on smaller scales ([37, 40, 42, 48, 52]) or at using their outputs to model meteorological variables (e.g., [20, 40]).

One extremely useful set of RCMs for studying precipitation on small spatial and temporal scales can be found in the North American Regional Climate Change Assessment Program (NARCCAP). This international program provides fine-resolution (50 km) climate-output data using a set of AOGCMs to provide the RCMs' boundary conditions over a domain covering the conterminous United States and most of Canada. For a comprehensive discussion concerning RCMs and NARCCAP, including statistical analyses, see [5, 20, 21, 24, 40].

Many statistical models have been developed to describe rainfall processes in continuous time and space and to estimate the distribution of rainfall data (e.g., [8, 44, 53] give useful reviews). One

interesting approach in continuous time is to use stochastic point processes to analyze rain cells and storms at a single site (e.g., [7, 32, 33, 50]). While these models have been shown to be effective, they do not address the case of multiple sites; in the literature, extensions to the spatio-temporal setting where data come from ground-based networks, have been made (e.g., [6, 8, 29, 31, 51]). For example, in [8], a relatively simple spatio-temporal model in which storm centers arrive according to a Poisson process was considered, while in [6] the arrival times of rain cells were modeled according to a clustered point process. Other approaches to spatio-temporal modeling of rainfall data are described in [1, 38, 39, 54].

In this article, we use NARCCAP data to define a threshold, above which the 50-km NARCCAP pixels in North America have "concentrated precipitation" for a sequence of 3-hourly time intervals. Thus, we focus on the major rainfall fronts and track them over a period of days. In particular, we build a dynamical set-valued autoregressive (SVAR) model, and the richness of the NARCCAP data allows us to incorporate variables such as wind speed and direction in the model. While descriptive in nature, the idea behind the model is to condition on the current state and allow pixels at the next time to "light up" (i.e., they are declared as having concentrated precipitation) according to (i) the wind data, and (ii) a random process that says pixels may lose or gain moisture according to a thinning process (whose thinning probability is estimated from the data). This model was adapted from [11], where the growth of *in vitro* breast-cancer cells was modeled over a 72-hour period. In the case of concentrated precipitation, the rapid movement of precipitation fronts means that a different model and different estimation techniques are needed.

The literature on growth models is extensive and includes theory and applications in various fields (e.g., probability, statistics, mathematical biology, meteorology, etc.), however set-based models are much less common. Definitions of random sets and Boolean models can be found in

[2, 10, 27, 41, 46, 49]. Many authors use convex random sets, Boolean models, and hitting functions for the analysis of tumor growth and biological data; some important references are [11, 14, 17, 28]. Recently, growth models have also been considered in the meteorological field for the analysis of storm cells and rainfall, such as in [25] and [26], where a methodology based on nonoverlapping random disk models is used in a multistage hierarchical Bayesian context. Our model does not impose random disks, nor the nonoverlapping restriction, but it also does not have the generality or uncertainty quantification typically associated with being hierarchical. Also, in [54], a Boolean model of rainfall patches is defined from a regional-scale stochastic spatio-temporal model that describes storms. Parameters are estimated from a ground-based network of eight weather stations. As such, their data, model, and inferences are very different from ours.

In Section 2, we review the theory of random sets and define some random-set models. In Section 3, we describe the SVAR model used to model the concentrated precipitation fields in a NARCCAP RCM. Section 4 establishes method-of-moments estimates of the model's parameters; estimates are obtained from the NARCCAP data, and their evolution over time is illustrated. Section 5 contains discussion and conclusions.

2 Random-set models

Broadly speaking, there are two different mathematical views one could take of the world. One is the so-called *field view*, such as a rainfall field expressed as precipitation amount per unit area. During any short period of time (e.g., 3 hours, 24 hours, 1 week, etc.), an amount of precipitation falls and is measured in cm of water equivalent. For example, individual rain gauges measure this precipitation at what are effectively points in space. But rainfall is notoriously patchy, so the "true"

or "target" precipitation field might be the average of point values over a small area; hence units of the field are in cm.

The other mathematical view of the world is the so-called *object view*. At one end of the precipitation scale, this might involve interest in raindrop/snowflake characteristics; at the other end, this might involve geographical regions (thought of as objects) where precipitation is particularly heavy, say above a pre-specified threshold. It is this latter view we shall take; in this article, we use random-set models to characterize the movement of concentrated precipitation regions across North America during a 1–2 day period. The purpose of this section is to review the theory of random sets, since it is through this that we shall build statistical models of objects.

Dynamical random-set modeling of concentrated precipitation represents a way to analyze precipitation over a large domain; see the literature review in Section 1. When data are in the form of gauge readings and radar images, there is a change-of-support problem to solve, namely combining data of different spatial supports that might be used for thresholding at a third spatial support. Recall that the thresholding defines the objects, which here are regions of concentrated precipitation.

A rigorous definition of random sets was given by [22] and [23]. That theory was summarized in [10, Ch. 9], and a number of random-set models were reviewed. There is one particular model that is central to statistical modeling and the object view of the world, namely the *Boolean model*.

Consider the d-dimensional Euclidean space $\mathbb{R}^d$ (in the application given in this article, $d = 2$; see Section 3). Let Z denote a compact (hence closed and bounded) subset of $\mathbb{R}^d$ that contains $\mathbf{0}$. Define

$$Z(\mathbf{s}) \equiv Z \oplus \{\mathbf{s}\} \equiv \{\mathbf{z} + \mathbf{s} \colon \mathbf{z} \in Z\}; \quad \mathbf{s} \in \mathbb{R}^d, \tag{1}$$

to be the translation of Z to location $\mathbf{s}$. Since Z contains the origin, $Z(\mathbf{s})$ contains $\mathbf{s}$. In that sense, we say that $Z(\mathbf{s})$ is *located at* $\mathbf{s}$. Definition (1) is a specific case of the more general definition of set addition,

$$A \oplus B \equiv \{\mathbf{a}+\mathbf{b}\colon \mathbf{a} \in A, \mathbf{b} \in B\}\,, \tag{2}$$

where A and B are any subsets of $\mathbb{R}^d$.

A Poisson point process of events $\{\mathbf{s}_i\}$ in $\mathbb{R}^d$ is a stochastic spatial process with some very attractive properties: The numbers of events in nonoverlapping subsets are independent; there is no more than one event in a small region $d\mathbf{s}$ located at $\mathbf{s}$; and the intensity of events is defined by

$$\lambda(\mathbf{s}) \equiv \lim_{|d\mathbf{s}|\to 0} E(N(d\mathbf{s}))/|d\mathbf{s}|\,; \quad \mathbf{s} \in \mathbb{R}^d\,,$$

where $N(A) \equiv \#$ events in $A \subset \mathbb{R}^d$; $|B|$ denotes the d-dimensional volume of $B \subset \mathbb{R}^d$; and $\lambda(\cdot)$ is the intensity function in units of $1/(d$-dimensional volume).

Let $Z_1, Z_2, \ldots$ be independent and identically distributed (iid) *random* compact sets. Independently, let $\{\mathbf{s}_i\}$ be a Poisson point process. Then the *Boolean model* X is defined to be

$$X \equiv \bigcup_i Z_i(\mathbf{s}_i)\,, \tag{3}$$

which is a random closed set made up of the union of compact sets located at the Poisson events. The book [18] can be consulted for a comprehensive account of the properties of the Boolean model and its generalizations; [13, Figure 4.19a], gives an illustration in $\mathbb{R}^d$, where $Z_1, Z_2, \ldots$ are iid random disks centered at $\mathbf{0}$ with random radii $R_1, R_2, \ldots$, respectively. It should be noted that the

collection of compact sets $\{Z_i(\mathbf{s}_i)\}$ in (3) can overlap, or even a large disk can completely mask the presence of a smaller disk. Consequently, inference on the Boolean model X is not straightforward, since bigger Z-components have a greater chance of being more visible.

Any random (closed) set X is characterized by its hitting function:

$$T_X(K) \equiv Pr(X \cap K \neq \emptyset), \quad \text{for all compact sets } K. \tag{4}$$

This characterization is akin to the characterization of a random variable by its cumulative distribution function. The calculation of the hitting function for various models is not always easy [12], but for the Boolean model X, [23] shows that when $\lambda(\cdot) \equiv \lambda$, (4) is given by,

$$T_X(K) = 1 - \exp\{-\lambda E(|\check{Z} \oplus K|)\}, \tag{5}$$

where $\oplus$ is set addition given by (2), and $\check{A} \equiv \{-\mathbf{a}\colon \mathbf{a} \in A\}$ is the reflection of the set A, reflected through the origin.

Equation (5) can be used to obtain estimators of Boolean-model parameters. For example, if Z is not only compact but also convex, then in $\mathbb{R}^2$, Steiner's formula ([43, p. 111]) yields:

$$E(|\check{Z} \oplus K|) = E(|Z|) + (2\pi)^{-1} E(Per(Z)) Per(K) + |K|, \tag{6}$$

where $Per(A)$ denotes the perimeter of the compact set A. Estimation is typically based on a method-of-moments estimating equation, where different choices for the "test set" K are randomly located throughout the spatial region of interest, and an empirical version of the left-hand side of (5) is matched to the theoretical expression on the right-hand side of (5) and (6). This yields

method-of-moments estimates, $\widehat{\lambda}_{MOM}$, $\widehat{E}(|Z|)_{MOM}$, and $\widehat{E}(Per(Z))_{MOM}$; for example, see [15] and [12].

A dynamical Boolean model was proposed by [9], where X_{t+1} is related to X_t by allowing the events of a Poisson point process (intensity function λ_{t+1}) that are in X_t to serve as locations for the iid random compact sets $Z_{t+1,1}, Z_{t+1,2}, \ldots$ (identically distributed as Z_{t+1}). That is,

$$X_{t+1} \;=\; \cup\{Z_{t+1,i}(\mathbf{x}_{t+1,i})\colon \mathbf{x}_{t+1,i} \cap X_t \neq \emptyset\}\,, \tag{7}$$

where $\{\mathbf{x}_{t+1,i}\}$ are the events of a Poisson spatial point process in $\mathbb{R}^d$ with intensity λ_{t+1}. By restricting the events to X_t, the spatial point process is equivalently Poisson with intensity λ_{t+1} in X_t. Equation (7) defines a SVAR model that can exhibit both growth and recession; in [11], it was used to characterize growth of breast cancer cells on a glass slide photographed 72 hours apart, and its hitting function was derived. On a transformed space, Z_{t+1} was assumed to be a disk of random radius R_{t+1} and, based on the hitting function, method-of-moments estimators (and standard errors) were found for λ_{t+1}, $E(R_{t+1})$, and $\mathrm{var}(R_{t+1})$; $t = 1, 2, \ldots$.

The hitting function for (7) is given by [11] and [10, p. 777]:

$$T_{X_{t+1}}(K) \;=\; 1 - \exp\{-\lambda_{t+1} E[|(\check{Z}_{t+1} \oplus K) \cap X_t|]\}\,, \tag{8}$$

where recall that K is any compact set in $\mathbb{R}^d$, λ_{t+1} is the rate of the Poisson process of events in X_t, and Z_{t+1} is a generic random set located at a generic Poisson event. There are a number of

equivalent formulations of the hitting function (8), owing to the relations,

$$\begin{aligned} |(\check{Z}_{t+1} \oplus K) \cap X_t| &= |(\bigcup_{\mathbf{z} \in Z_{t+1}} K \oplus \{-\mathbf{z}\}) \cap X_t| = |\bigcup_{\mathbf{z} \in Z_{t+1}} [K \cap (X_t \oplus \mathbf{z})]| \\ &= |K \cap [\bigcup_{\mathbf{z} \in Z_{t+1}} (X_t \oplus \mathbf{z})]| = |K \cap (X_t \oplus Z_{t+1})| \,. \end{aligned} \tag{9}$$

In the next section, we shall use a physically motivated set autoregressive process to build a dynamical model for the regions of concentrated precipitation at 3-hourly intervals. The fact that the data come from an RCM, defined on 50×50 km pixels, results in modifications to the dynamical Boolean model given by (7). Further, the SVAR process can take advantage of physical variables (e.g., wind speed and direction) available from the RCM.

3 SVAR models of concentrated precipitation

In what follows, we build an SVAR process that attempts to capture the dynamics inherent in weather. To model the concentrated precipitation field, we use a dynamical Boolean model and the wind-vector field plays the role of a covariate.

Let D_s denote the spatial domain of interest, here defined as all the 50×50 km pixels in the NARCCAP region from 25.2°–47.3° lattitude and 263°–288° longitude. The precipitation field $\{Y_t(\mathbf{s})\colon \mathbf{s} \in D_s\}$ can be thought of as a time series of spatial processes, such as in [13, Ch. 6]. In this article, we choose one of the RCMs known as CRCM-CGCM3, which is a Canadian RCM with boundary conditions supplied by a Canadian AOGCM. Then a 27-hour time period during June 21–22, 1968, starting with NARCCAP time 1372, showed concentrated-precipitation activity. Consequently, NARCCAP time 1373 is 3 hours later, and NARCCAP time 1381 is 27 hours later.

Sets of concentrated precipitation were obtained by thresholding the precipation field, as follows:

$$X_t \equiv \{\mathbf{s} \in D_s \colon Y_t(\mathbf{s}) > k(\mathbf{s})\}\,; \quad t = 1, 2, \ldots\,, \tag{10}$$

where time point t is defined to be $t =$ (NARCCAP time $-$ 1371). For each $\mathbf{s} \in D_s$, $k(\mathbf{s})$ was calibrated so that

$$\widehat{\Pr}(Y_t(\mathbf{s}) < k(\mathbf{s})) \simeq 99.95\%\,,$$

where $\widehat{\Pr}$ is a long-run frequency for the 3-hourly time points between 1968 and 2000. Figure 1 shows a map of $\{k(\mathbf{s}) \colon \mathbf{s} \in D_s\}$, where the units are in cm of precipitation.

Figure 1 here

In what follows, we treat the sequence $\{X_t \colon t = 1, \ldots, 10\}$ as a time series of random sets evolving according to the SVAR process specified by (15) below. The goal of this article is to fit the parameters of the process to the data $\{X_t \colon t = 1, 2, \ldots\}$, using covariate information from the NARCCAP RCM. The covariate we use here is the wind field,

$$\{\mathbf{W}_t(\mathbf{s}) \colon \mathbf{s} \in D_s\}\,, \tag{11}$$

where $\mathbf{W}_t(\mathbf{s}) \equiv (U_t(\mathbf{s}), V_t(\mathbf{s}))'$ is the wind velocity vector in units of km/hr. By convention, $U_t(\mathbf{s})$ and $V_t(\mathbf{s})$ are the E-W and N-S components, respectively, of the wind velocity vector at pixel $\mathbf{s}$ and at time t.

Since the wind moves parcels of air containing moisture that becomes precipitation, we incorporate a dynamic modification to the set autoregressive process (7). To allow the foci of growth to move from t to $t + 1$, we extend each pixel $\mathbf{x}$ in X_t, along a line segment in the forward and

backward direction of $\mathbf{W}_t(\mathbf{x})$. The line segment originating from $\mathbf{x}$ is defined as follows:

$$L_t(\mathbf{x}) \equiv \{3c\mathbf{W}_t(\mathbf{x}) \oplus \mathbf{x} \colon -1 \leq c \leq 1\}, \tag{12}$$

where $3\mathbf{W}_t(\mathbf{x})$ is the displacement in km that a parcel of air at location $\mathbf{x} \in X_t$ and at time t would travel in 3 hours. Then X_t is "fattened" to yield

$$M_t \equiv \cup\{L_t(\mathbf{x}) \colon \mathbf{x} \in X_t\}, \tag{13}$$

which is then "convexified":

$$C_t(X_t) \equiv co(M_t), \tag{14}$$

where $co(A)$ denotes the convex hull of the set $A \subset \mathbb{R}^d$. It is this set, which contains X_t, where centers of precipitation may occur at the next time point, $t+1$. In practice, $C_t(X_t)$ is obtained by joining center points of boundary pixels to give a convex set in $\mathbb{R}^d$. This is then pixellated by a greedy algorithm that puts a pixel in $C_t(X_t)$ if any part of the boundary lines intersect that pixel; finally, the pixels are filled in to yield a discretized convex set.

Recall the set autoregressive model (7), where Poisson events are chosen in X_t. Our modification for modeling the dynamics of concentrated precipitation is to replace X_t in (7), with $C_t(X_t)$. That is, we propose the SVAR model,

$$X_{t+1} = \cup\{Z_{t+1,i}(\mathbf{x}_{t+1,i}) \colon \mathbf{x}_{t+1,i} \cap C_t(X_t) \neq \emptyset\}; \quad t = 1, 2, \ldots, \tag{15}$$

where $Z_{t+1,i}(\mathbf{x}_{t+1,i})$ is a random convex compact set located at an event $\mathbf{x}_{t+1,i}$ of the spatial Poisson point process with intensity λ_{t+1}.

Consequently, the hitting function (8) is modified to be:

$$T_{X_{t+1}}(K) \;=\; 1 - \exp\{-\lambda_{t+1} E[|(\check{Z}_{t+1} \oplus K) \cap C_t(X_t)|]\}\,, \tag{16}$$

where $C_t(X_t)$ is defined by (10)–(12). A similar derivation to (9) yields,

$$T_{X_{t+1}}(K) = 1 - \exp\{-\lambda_{t+1} E[|K \cap (C_t(X_t) \oplus Z_{t+1})|]\}\,,$$

where K is any compact set.

An important component of the model that we need to specify is the random set Z_{t+1} (and its probability law). Again, the physical movement of parcels of air motivates the choice of $Z_{t+1,1}, Z_{t+1,2}, \ldots$ as iid line segments whose distribution is determined by the direction and strength of the local wind field in the vicinity of X_t. Let $Z_{t+1}(\mathbf{x})$ be a random line (actually a sequence of pixels) drawn from $\mathbf{x}$ to a random displacement, $\mathbf{d}_t(\mathbf{x}) \equiv (d_{t,1}(\mathbf{x}), d_{t,2}(\mathbf{x}))'$ in $\mathbb{R}^2$, where

$$\mathbf{d}_t(\mathbf{x}) \;\equiv 3\mathbf{W}_t(\mathbf{x}) = (3U_t(\mathbf{x}), 3V_t(\mathbf{x}))\,, \tag{17}$$

given by (11). Then the random vector $\mathbf{d}_t(\mathbf{x})$ has a density that we denote as $f_{t+1}(\mathbf{d})$, where we use $t+1$ as the subscript since it refers to the probability law of Z_{t+1}. Then estimation of $f_{t+1}(\cdot)$ can be achieved by a kernel smoothing of $\{3W_t(\mathbf{x})\colon \mathbf{x} \in M_t\}$, the set of all displacements originating from the pixels in M_t given by (12). After carrying out extensive exploratory data analysis where moving windows, scatter diagrams, kernel density estimates, and fitted bivariate normal distributions were

compared, we concluded that (robustly) fitted means $\widehat{\mu}_{t+1,1}$, $\widehat{\mu}_{t+1,2}$, standard deviations $\widehat{\sigma}_{t+1,1}$, $\widehat{\sigma}_{t+1,2}$, and correlation $\widehat{\rho}_{t+1}$, were excellent descriptors of the dynamics that displace the parcels of air.

The remaining parameter is λ_{t+1}, the intensity of events in $C_t(X_t)$ that are used as foci of growth for the set of concentrated precipitation, X_{t+1}. This is related to the "thinning" probabilities:

$$p_{t+1} \equiv \Pr(\text{event is at a pixel of } C_t(X_t)) = \lambda_{t+1} \cdot (\text{area of NARCCAP pixel}). \tag{18}$$

Thus, it is equivalent to estimate p_{t+1}, which we accomplish through Laslett's theorem ([10, p. 766]).

4 Estimation of the dynamics

The data we analyze consist of a 30-hour period of NARCCAP concentrated-precipitation fields during June 21–22, 1968, which are defined by (10) and Figure 1; they are denoted as $\{X_t : t = 1, \ldots, 10\}$. After initializing with set X_1, an evolution of the sets $\{X_{t+1} : t = 1, \ldots, 9\}$ is shown in Figure 6 below.

The parameters of the SVAR model (15) control the dynamics. In Section 4.1, the evolution of the random sets $\{Z_{t+1} : t = 1, \ldots, 9\}$ is characterized and estimated. In Section 4.2, the evolution of the thinning probabilities $\{p_{t+1} : t = 1, \ldots, 9\}$ is estimated.

4.1 Estimation of the displacement densities

Recall, from Section 3, the law of the random set Z_{t+1} is characterized by the displacement density defined on $\mathbb{R}^2$. Extensive exploratory data analysis led us to choose a bivariate normal distribution with parameters changing dynamically for $t = 1, \ldots, 9$. The mean vector $(\mu_{t+1,1}, \mu_{t+1,2})$, the stan-

dard deviations $\sigma_{t+1,1}$ and $\sigma_{t+1,2}$, and the correlation coefficient ρ_{t+1} of this bivariate distribution are estimated from the displacement vectors $\{d_{t+1}(\mathbf{x})\colon \mathbf{x} \in M_{t+1}\}$ and $\{d_t(\mathbf{x})\colon \mathbf{x} \in M_t\}$.

As an illustration, Figure 2 shows the case of $t = 3$, and the pixels in M_t are shown as small circles. The displacement vector $d_t(\mathbf{x})$ is attached as an arrow to each $\mathbf{x} \in M_t$. Clearly, the displacement vectors have a well behaved distribution, with a well defined mean direction and size. Nevertheless, there are several displacement vectors that do not follow the overall direction, and these outliers most likely represent local turbulence.

Figure 2 here

The presence of outliers is the reason why empirical means, variances, and covariances are not immediately useful when fitting a distribution to the displacement field. Instead, we use robust estimators that are based on the Least Trimmed Squares procedure first described in [34]. This procedure is based on selecting the "best" (in some sense) data subset and using it to compute an empirical mean vector and an empirical covariance matrix. In our case, we use the minimum covariance determinant method of [35], which involves selecting the data subset that produces the covariance matrix with the smallest determinant. Recall that each symmetric positive-definite matrix can be associated with a multivariate normal destribution whose contour lines are ellipses. Essentially then, the minimum covariance determinant procedure finds the ellipse with minimum area among all those covering at least $h = (n + 3)/2$ of the data points. One can roughly describe this as using the h "most central" points to calculate the covariance matrix; a recent review of the minimum covariance determinant method can be found in [19]. Finally, we use the R package MASS [30] to implement this algorithm.

The results of this robust fitting of normal densities are shown in Figure 3. To obtain these fits, we combine the displacement data from times $t + 1$ and t. The mean vector of the resulting dis-

tribution is $(\widehat{\mu}_{t+1,1}, \widehat{\mu}_{t+1,2})$; the covariance matrix parameters are standard deviations $\widehat{\sigma}_{t+1,1}$, $\widehat{\sigma}_{t+1,2}$, and the correlation coefficient $\widehat{\rho}_{t+1}$.

In order to provide a graphical summary of the changing wind field, we plot a sequence of nine robustly fitted Gaussian densities, superimposed on the data used to fit them. Figure 3 displays them as a 3×3 matrix of plots with t increasing from left to right.

Figure 3 here

One can see rather easily changes in the ellipse center location as t increases. Figure 4 shows the passage of the center of concentrated precipitation as time increases from $2, \ldots, 10$. The change in ellipses is rather gradual for most of the times except at the beginning. This pattern may be related to rather rapid changes in wind velocity and direction when the storm has more energy. As confirmation of this, the degree of spread (as measured by the focal length of the 95% ellipses) seems to subside by the end of observation period.

Figure 4 here

4.2 Estimation of the thinning probabilities

The parameters of the SVAR model (16) consist of the distribution of successive random sets $\{Z_{t+1} \colon t = 1, 2, \ldots\}$ (Section 4.1) and the Poisson rate $\{\lambda_{t+1} \colon t = 1, 2, \ldots\}$. As was shown in (18), λ_{t+1} is proportional to the thinning probability p_{t+1}, which we estimate in this section.

According to (15), X_{t+1} is a Boolean model whose foci of growth are constrained to belong to $C_t(X_t) = co(M_t)$, given by (12)–(14). The construction of M_t and its "convexification" $C_t(X_t)$, for $t = 1, \ldots, 9$, is shown in Figure 5. It is the interaction of $C_t(X_t)$ with X_{t+1} that determines the evolutionary parameters; Figure 6 shows the two sets superimposed, for $t = 1 \ldots, 9$. Notice that,

mostly, X_{t+1} is contained in $C_t(X_t)$, which is an indication that (15) is successfully describing the temporal evolution of $\{X_{t+1}\colon t = 1, \ldots, 9\}$.

Figure 5 here

Figure 6 here

We now estimate the thinning probabilities $\{p_{t+1}\colon t = 1, \ldots, 9\}$. Each convex compact set $Z_{t+1,i}$ in (15) has a so-called *marker point* (in discrete space, we call this a *marker pixel*). This is a unique way of identifying the set's location; for example, we use the extreme south-west corner. If we simply counted the *exposed marker points (pixels)* in the Boolean model, the count would be biased low, since a small set could be masked partially or completely by a larger one. Laslett's theorem [10, p. 766], gives a way to compensate for this bias. One simply removes all pixels of X_{t+1} that are not exposed marker pixels – they are removed both from X_{t+1} and from $C_t(X_t)$. The result is a transformed $C_t(X_t)$, call it $C_t^o(X_t)$, that contains just the exposed marker pixels, X_{t+1}^o, of X_{t+1}. These are shown for $t = 1, \ldots, 9$ in Figure 7.

Figure 7 here

From Laslett's theorem, an unbiased estimate of p_{t+1} is:

$$\widehat{p}_{t+1} \;=\; \frac{\#\text{ pixels in } X_{t+1}^o}{\#\text{ pixels in } C_t^o(X_t)}\,. \tag{19}$$

A time series plot of $\{\widehat{p}_{t+1}\colon t = 1, \ldots, 9\}$, given by (19), is shown in Figure 8. The variability of these estimates over time periods $2, \ldots, 10$ is not large and, based on this figure, there is no indication of nonstationarity of the thinning probabilities during the passage of the concentrated precipitation across the upper Midwest.

Figure 8 here

Finally, all the estimated parameters of the SVAR model are combined into Table 1, showing the evolution of displacement means, standard deviations, and correlations, along with the evolution of the thinning probabilities. Clearly, the former parameters describe the general movement of the concentrated-precipitation field, and the latter parameter describes its shape, or "patchiness."

Table 1 here

5 Discussion and Conclusions

Comparing the results in Table 1 and Figure 3 reveals that the concentrated precipitation sets change in a smooth way, and no abrupt transitions are observed. Then, as expected, the parameters governing the dynamics vary but not greatly, from one time interval to the next. This points to the predictive potential of our SVAR model.

Except for a few outliers, the wind movement is well characterized by a general direction that changes little as time increases. It is apparent from the fitted parameters that the center of the random set X_{t+1} moves most noticeably from west to east, with much smaller displacements in the north-south direction. This is in line with typical weather patterns in the upper Midwest of the USA.

While the concentrated precipitation moves steadily from west to east, there is less apparent change in the displacements' covariances. The fact that $d_{t+1}(\mathbf{x})$ has coordinates that are mostly highly positively correlated, stresses the prevalence of this general wind direction. Similarly, the thinning probabilities given in (18) do not change greatly over time and provide no evidence of nonstationary behavior.

This research is meant to characterize precipitation fronts defined by RCMs, in order to say what mechanisms might control real fronts. Our estimates have relied on knowing the wind-vector field (11). Suppose the SVAR model for concentrated precipitation is assumed; in the case of real fronts, where data and covariates are limited, we would probably have to use method-of-moments parameter estimates defined by matching the theoretical hitting function given by (16) or (17) with an empirical version.

In an object view of the world, the SVAR process offers a flexible way to model dynamics. It remains to put this into a hierarchical statistical model, where it would play the role of a process model and quantification of uncertainty would be naturally facilitated. From this point of view, our contribution in this article has been to assess the feasibility of the SVAR process in this role. The data model would capture all the imperfections of the real world (in contrast to an RCM), where data (precipitation) and covariate (wind field) are observed incompletely, with measurement error, and at different spatial supports.

Acknowledgment

This article was the result of a collaboration in the SAMSI sub-working group, SG2, during and beyond the SAMSI Program, "Space-Time Analysis for Environmental Mapping, Epidemiology and Climate Change (2009-2010)." Cressie led SG2; the order of co-authors after the first is alphabetical. The research was partially supported by SAMSI and the National Science Foundation (NSF) under agreement DMS-0635449. Cressie's research was partially supported by the Office of Naval Research under grant N00014-08-1-0464. Assunção's research was partially supported by Brazilian research support agencies, CNPq and FAPEMIG; Levine's research was partially supported by the NSF

under grant number DMS-0805748.

References

[1] Bardossy, A. and Plate, E. J. (1992). Space-time model for daily rainfall using atmospheric circulation patterns. *Water Resources Research* **28** 1247–1259.

[2] Barndorff-Nielsen, O. E., Kendall, W. S., and van Lieshout, M. N. M. (1999). *Stochastic Geometry, Likelihood and Computation*. New York, NY: Chapman and Hall.

[3] Berliner, L. M., and Kim, Y. (2008). Bayesian design and analysis for superensemble based climate forecasting. *Journal of Climate* **21** 1891–1910.

[4] Chaplain, M., Singh, G. D. , and McLachlan, J. C. (1999). *On Growth and Form: Spatio-Temporal Pattern Formation in Biology*. Chichester, UK: Wiley.

[5] Christensen, W. F. and Sain, S. R. (2011). Spatial latent variable modeling for integrating output from multiple climate models. *Mathematical Geosciences* doi10.1007/s11004-011-9321-1.

[6] Cowpertwait, P. S. P. (1995). A generalized spatial-temporal model of rainfall based on a clustered point process. *Proceedings of the Royal Society of London A* **450** 163–175.

[7] Cowpertwait, P. S. P., Isham V., and Onof C. (2007). Point process models of rainfall: developments for fine-scale structure. *Proceedings of the Royal Society A* **463** 2569–2587.

[8] Cox, D. R. and Isham, V. (1988). A simple spatial-temporal model of rainfall. *Proceedings of the Royal Society of London A* **415** 317–328.

[9] Cressie, N. (1991). Modeling growth with random sets, in *Spatial Statistics and Imaging (Proceedings of the 1988 AMS-IMS-SIAM Joint Summer Research Conference)*, ed. A. Possolo. Institute of Mathematical Statistics, Hayward, CA, 31–45.

[10] Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Revised edn. New York, NY: Wiley.

[11] Cressie, N. and Hulting, F. L. (1992). A spatial statistical analysis of tumor growth. *Journal of the American Statistical Association* **87** 272–283.

[12] Cressie, N. and Laslett, G. M. (1987). Random set theory and problems of modeling. *SIAM Review*, **29** 557–574.

[13] Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Hoboken, NJ: Wiley.

[14] Deijfen, M. (2003). Asymptotic shape in a continuum growth model. *Advances in Applied Probability* **35**, 303–318.

[15] Diggle, P. J. (1981). Binary mosaics and the spatial pattern of heather. *Biometrics*, **37**, 531–539.

[16] Giorgi, F. and Mearns, L. O. (1999). Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research* **104** 6335–6352.

[17] Grenander, U., Srivastava, A., and Saini, S. (2007). A pattern-theoretic characterization of biological growth. *IEEE Transactions on Medical Imaging* **26** 648–659.

[18] Hall, P. (1988). *Introduction to the Theory of Coverage Processes*. New York, NY: Wiley.

[19] Hubert, M. and Debruyne, M. (2009). Minimum covariance determinant. *Computational Statistics* **2**, 36–43.

[20] Kang, E. L., Cressie, N., and Sain, S. (2010). Combining outputs from the NARCCAP regional climate models using a Bayesian hierarchical model. *Technical Report* No. 837, Department of Statistics, The Ohio State University.

[21] Kaufman, C. and Sain, S. (2010). Bayesian functional ANOVA modeling using Gaussian process prior distributions. *Bayesian Analysis* **5** 123–150.

[22] Kendall, D. G. (1974). Foundations of a theory of random sets. In *Stochastic Geometry*, eds E. F. Harding and D. G. Kendall, New York, NY: Wiley, 322–376.

[23] Matheron, G. (1975). *Random Sets and Integral Geometry*, New York, NY: Wiley.

[24] Mearns, L. O., Gutowski, W., Jones, R., Leung, R., McGinnis, S., Nunes, A., and Qian, Y. (2009). A regional climate change assessment program for North America. *Eos Transactions, American Geogphysical Union* **90** 311.

[25] Micheas, A. C. , Fox, N. I., Lack, S. A., and Wikle, C. K. (2007). Cell identification and verification of QPF ensembles using shape analysis techniques. *Journal of Hydrology* **343** 105–116.

[26] Micheas, A. C. and Wikle, C. K. (2009). A Bayesian hierarchical nonoverlapping random disc growth model. *Journal of the American Statistical Association* **104** (485) 274–283.

[27] Molchanov, I. S. (1997). *Statistics of the Boolean Model for Practitioners and Mathematicians*. Chichester, UK: Wiley.

[28] Newton, M. A. (2006). On estimating the polyclonal fraction in lineage-marker studies of tumor origin. *Biostatistics* **7** 503–514.

[29] Onof, C., Chandler, R.E., Kakou, A., Northrop, P., Wheater, H.S., and Isham, V. (2000). Rainfall modeling using Poisson-cluster processes: A review of developments. *Stochastic Environmental Research Risk Assessment* **14** 384–411.

[30] R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (http://www.R-project.org).

[31] Rodriguez-Iturbe, I. and Eagleson, P. S. (1987). Mathematical models of rainstorm events in space and time. *Water Resourses Research* **23** (1) 181–190.

[32] Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. S. (1987). Some models for rainfall based on stochastic point processes. *Proceedings of the Royal Society of London A* **410** 269–288.

[33] Rodriguez-Iturbe, I., Cox, D. R., and Isham, V. S. (1987). A point process model for rainfall: further developments. *Proceedings of the Royal Society of London A* **417** 283–298.

[34] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association* **79**, 871–880.

[35] Rousseeuw, P. J. and Van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 212–223.

[36] Sain, S. R. and Furrer, R. (2010). Combining climate model output via model corrections. *Stochastic Environmental Research and Risk Assessment* **24**, 821–829.

[37] Sain, S. R., Furrer, R., and Cressie, N. (2011). A spatial analysis of multivariate output from regional climate models. *Annals of Applied Statistics* In Press (http://www.imstat.org/aoaos/next_issue.html)

[38] Sansó, B. and Guenni, L. (1999). Venezuelan rainfall data analysed by using a Bayesian space-time model. *Applied Statistics* **48** 345–362.

[39] Sansó, B. and Guenni, L. (2000). A nonstationary multisite model for rainfall. *Journal of the American Statistical Association* **95** 1089–1100.

[40] Schliep, E. M., Cooley, D., Sain, S. R., and Hoeting, J. A. (2010). A comparison study of extreme precipitation from six different regional climate models via spatial hierarchical modeling. *Extremes* **13** 219–239.

[41] Schneider, R. (1993). *Convex Bodies: The Brunn-Minkowski Theory*. Cambridge, UK: Cambridge University Press.

[42] Schneider, S.H. (2001). What is "dangerous" climate change? *Nature* **411** 17–19.

[43] Serra, J. (1982). *Image Analysis and Mathematical Morphology*. London, UK: Academic Press.

[44] Smith, R. L. and Robinson, P. J. (1997). A Bayesian approach to the modelling of spatial-temporal precipitation data. In: *Case Studies in Bayesian Statistics III. Lecture Notes in Statistics 121*, eds C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCullogh, P. Rossi and N.D. Singpurwalla, New York, NY: Springer Verlag, pp. 237–264.

[45] Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O. (2009). Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association* **104** 97–116.

[46] Stoyan, D., Kendall, W. S., and Mecke, J. (1995). *Stochastic Geometry and Its Applications (2nd ed.)*. New York, NY: Wiley.

[47] Tebaldi, C. and Sansó, B. (2009). Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayes approach. *Journal of the Royal Statistical Society, Series A* **172** 83–106.

[48] Tebaldi, C. , Smith, R. L., Nychka, D., and Mearns, L. O. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate* **18** 1524–1540.

[49] van den Berg, J., Meester, R., and White, D. G. (1997). Dynamic Boolean models. *Stochastic Processes and their Applications* **69** 247–257.

[50] Waymire, E. D. and Gupta, V. K. (1981). The mathematical structure of rainfall representations: 3, some applications of the point process theory to rainfall processes. *Water Resources Research* **17** 1287–1294.

[51] Wheater, H.S., Isham, V.S., Cox, D.R., Chandler, R.E., Kakou, A., Northrop, P.J., Oh, L., Onof, C., and Rodriguez-Iturbe, I. (2000). Spatial-temporal rainfall fields: modelling and statistical aspects. *Hydrology Earth System Science* **4** 581–601.

[52] Wigley, T. M. L. and Raper, S. C. B. (2001). Interpretation of high projections for global mean warming. *Science* **293** 451–454.

[53] Woolhiser, D. A. (1992). Modelling daily precipitation progress and problems. In *Statistics in the Environmental and Earth Sciences*, eds A. Waden and P. Guttorp, New York, NY: Halsted Press, pp. 20–22.

[54] Zhang, Z. and Switzer, P. (2007). Stochastic space-time regional rainfall modeling adapted to historical rain gauge data. *Water Resources Research* **43**, W03441.

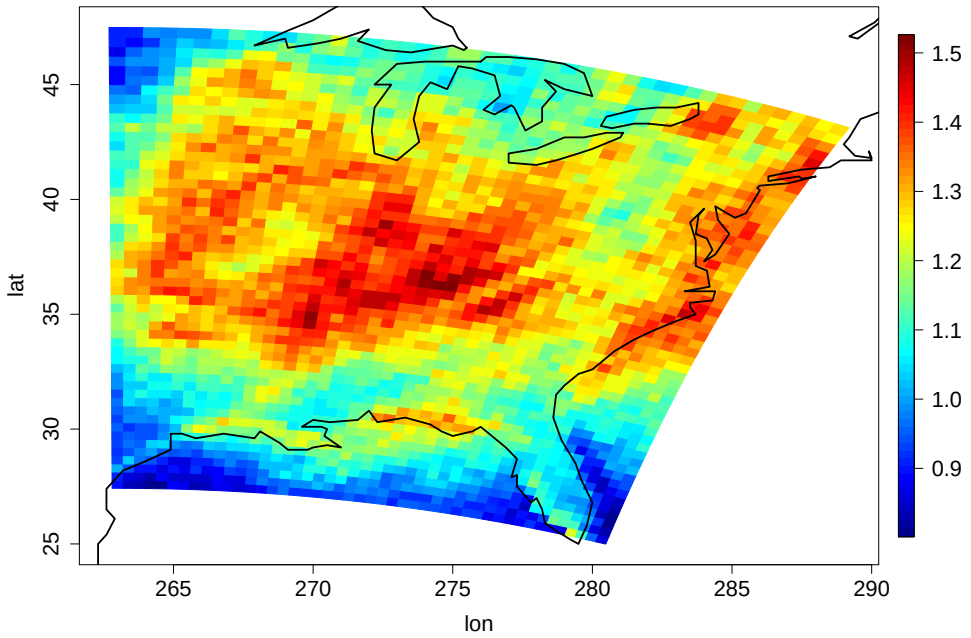

Figure 1: Map of threshold values $k(\cdot)$ in (10), from which concentrated precipitation sets $\{X_t\}$ are obtained.

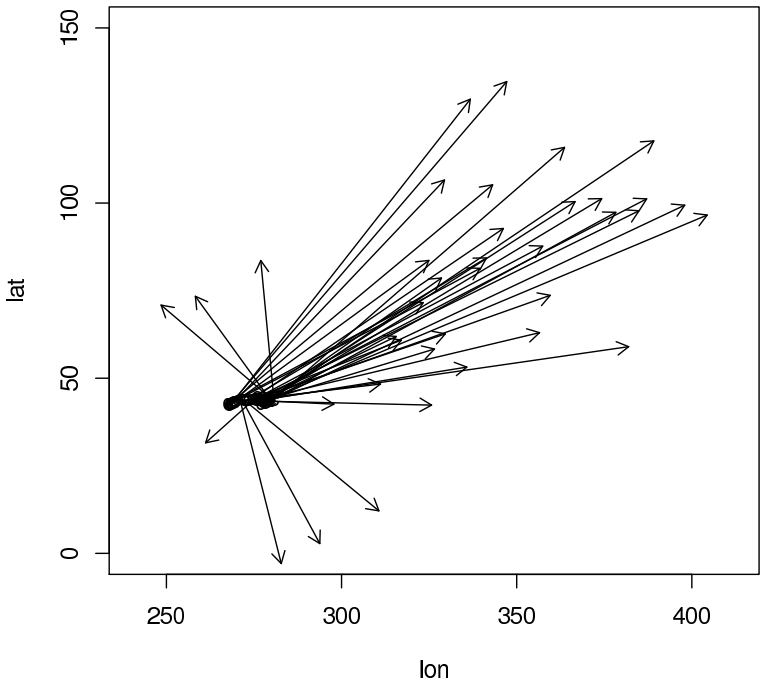

Figure 2: This plot shows pixels $\mathbf{x} \in M_3$, with their associated displacement vectors $\{d_3(\mathbf{x})\colon \mathbf{x} \in M_3\}$.

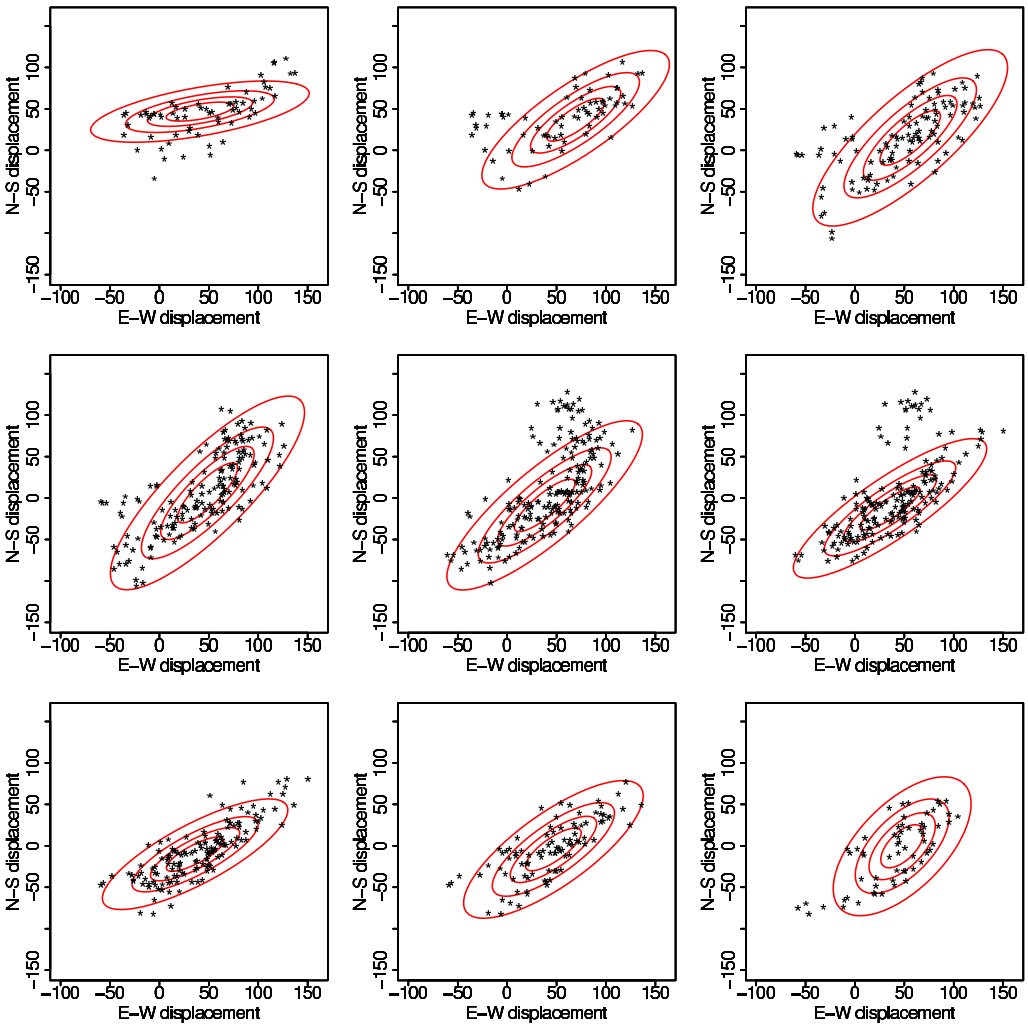

Figure 3: The plot shows 95, 75, 50, and 25 percentiles of robustly fitted bivariate Gaussian distributions describing the wind field dynamics for times 2 through 10. The top-left plot is for time 2, time 3 is the top-middle plot, . . . , and time 10 is the bottom-right plot.

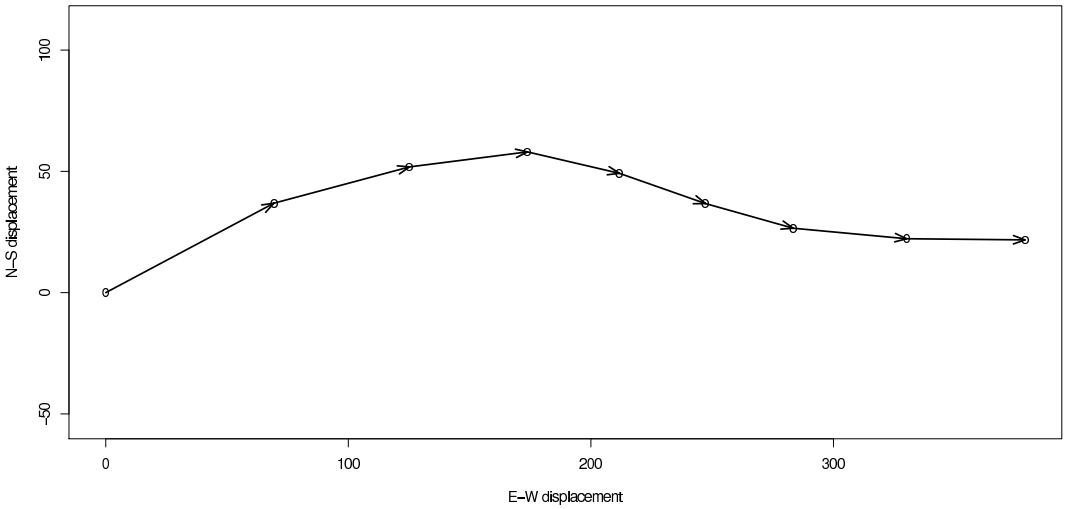

Figure 4: Passage of concentrated-precipitation centers $\{(\widehat{\mu}_{1,t+1}, \widehat{\mu}_{2,t+1}) \colon t = 1, \ldots, 9\}$

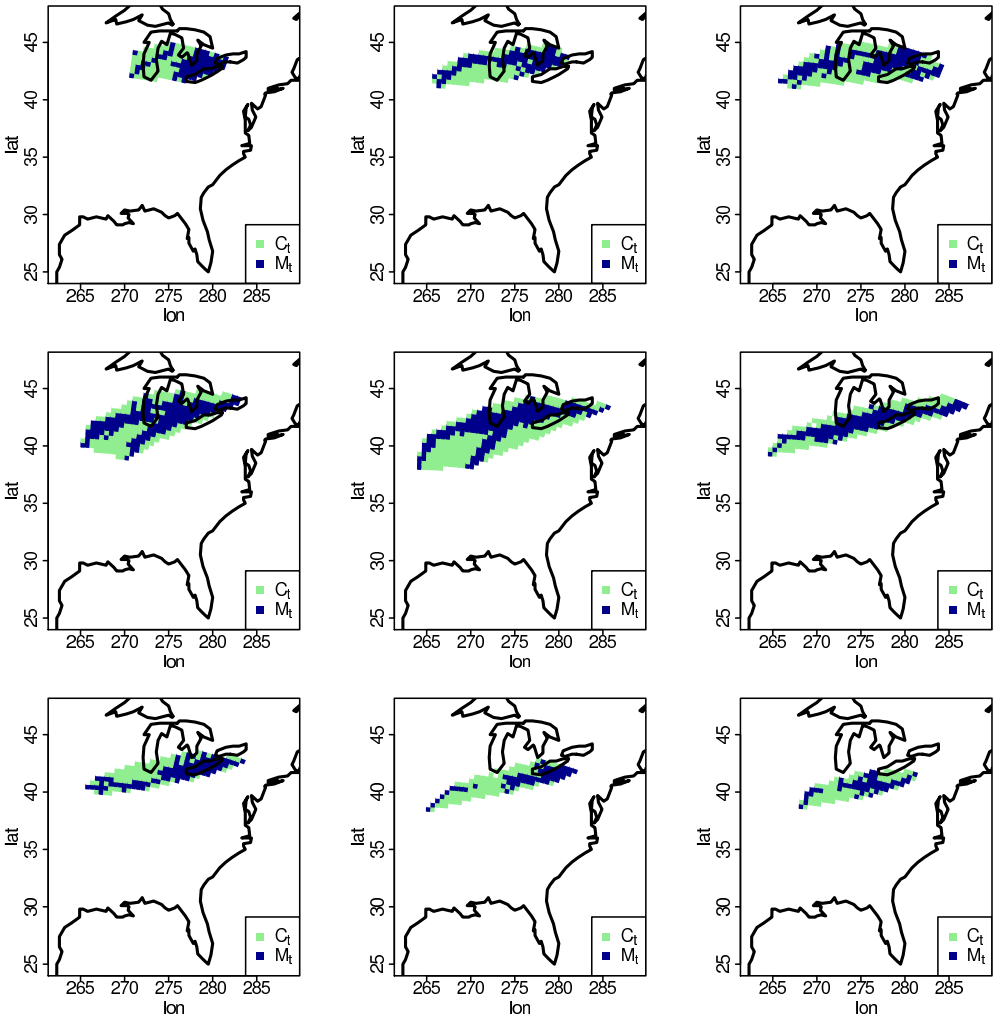

Figure 5: The "fattened" set M_t, given by (13), is superimposed on its convexification, $C_t(X_t)$, given by (14), for times 2 through 10. The top-left plot is for time 2, time 3 is the top-middle plot, ..., and time 10 is the bottom-right plot.

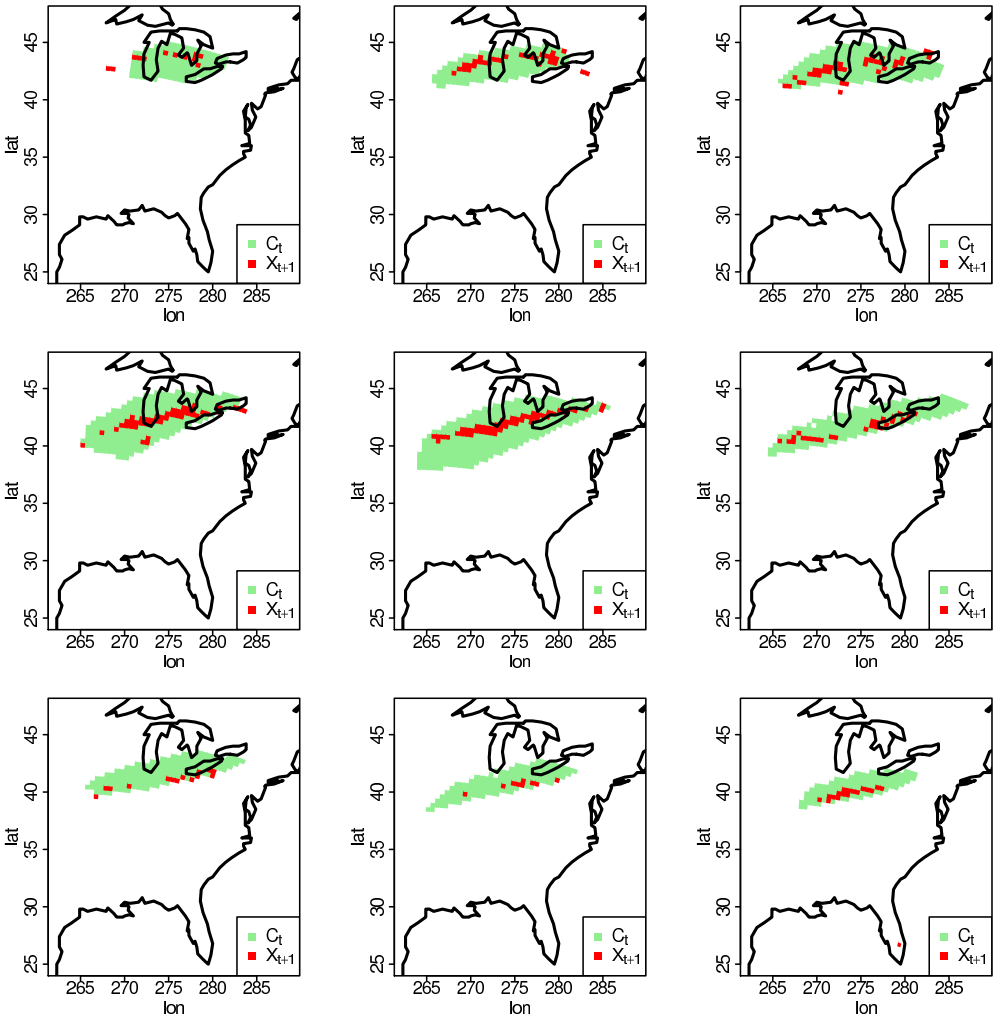

Figure 6: The set X_{t+1} is superimposed on $C_t(X_t)$, showing how X_{t+1} relates to X_t in the SVAR model (15), for times 2 through 10. The top-left plot is for time 2, time 3 is the top-middle plot, ..., and time 10 is the bottom-right plot.

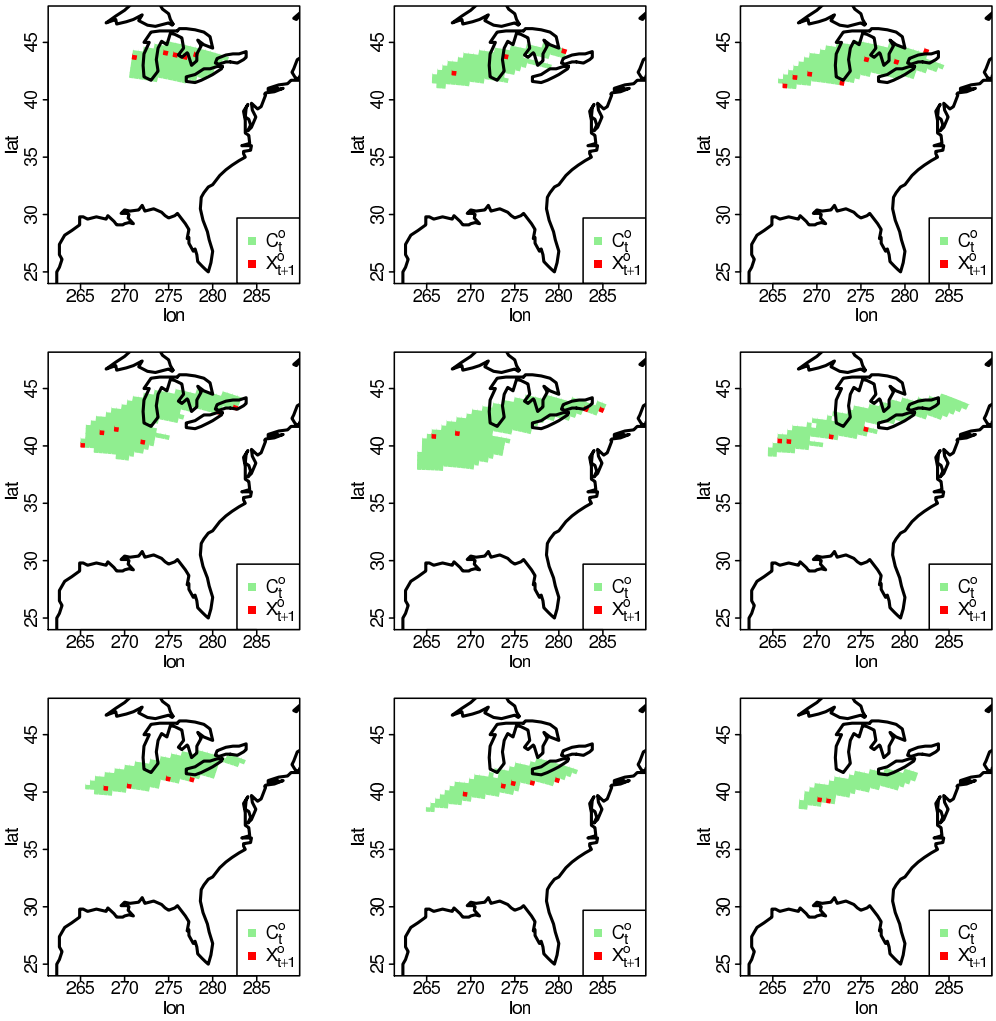

Figure 7: The set of exposed marker points, X_{t+1}^o, is superimposed on the transformed set $C_t^o(X_t)$. The ratio of the number of pixels in X_{t+1}^o to the number of pixels in $C_t^o(X_t)$, is an estimate, $\widehat{p}_{t+1}$, of the thinning probability, for times 2 through 10; see (19). The top-left plot is for time 2, time 3 is the top-middle plot, ..., and time 10 is the bottom-right plot.

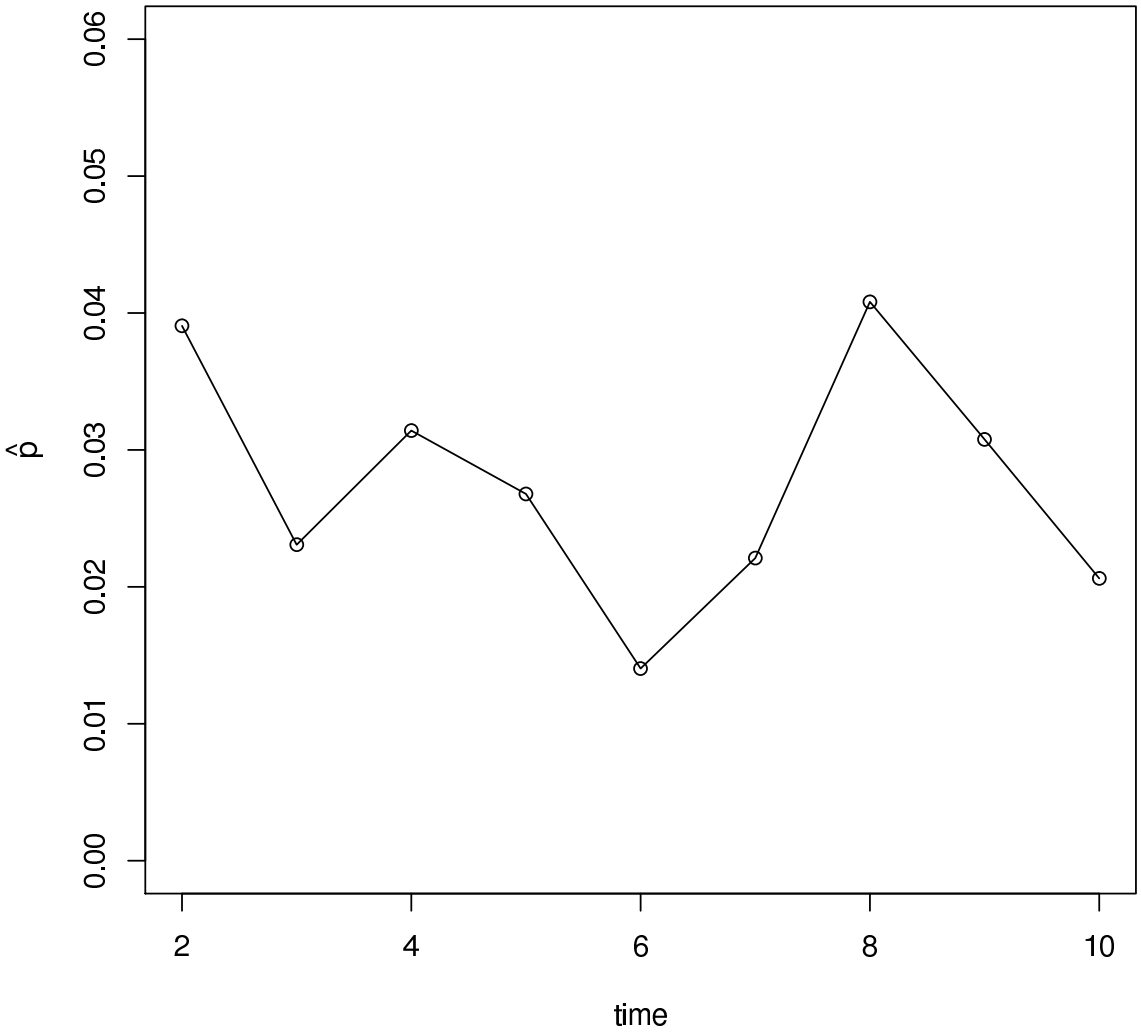

Figure 8: Time series of the estimates of the thinning probabilities, $\{\widehat{p}_{t+1} : t = 1, \ldots, 9\}$, given by (19).

Table 1: Parameter estimates of SVAR model for $\{X_{t+1}\colon t = 1, \ldots, 9\}$. The displacement means (μ_1, μ_2) and standard deviations (σ_1, σ_2) are in units of km; the displacement correlations (ρ) and thinning probabilities (p) are unitless.

Parameter Estimates				
$t+1$	$(\widehat{\mu}_{t+1,1},\, \widehat{\mu}_{t+1,2})$	$(\widehat{\sigma}_{t+1,1},\, \widehat{\sigma}_{t+1,2})$	$\widehat{\rho}_{t+1}$	$\widehat{p}_{t+1}$
2	(40.978, 46.649)	(45.172, 15.091)	0.597	0.039
3	(69.294, 36.862)	(38.768, 34.136)	0.800	0.023
4	(55.926, 14.973)	(40.263, 43.471)	0.799	0.031
5	(48.611, 6.191)	(40.138, 47.724)	0.832	0.027
6	(37.918, -8.838)	(40.411, 41.711)	0.839	0.014
7	(35.520, -12.470)	(39.951, 34.358)	0.854	0.022
8	(36.010, -10.158)	(38.335, 27.256)	0.797	0.041
9	(46.772, -4.340)	(37.128, 33.924)	0.776	0.031
10	(48.824, -0.472)	(27.709, 34.546)	0.629	0.021