Rapport de stage
de deuxieme année de Magistère de l'Ecole Normale Supérieure

10 Février - 10 Aout

Statistical synthesis of tropical cyclone tracks in a risk evaluation perspective

Synthèse de trajectoires de cyclones selon un modèle statistique dans une perspective d'estimation des risques

Camille Risi

sous la direction de

Kerry Emanuel

au

Massachussets Institute of Technology
Department of Earth, Atmospheric and Planetary Sciences
Program in Atmosphere, Oceans and Climate

Statistical synthesis of tropical cyclone tracks in a risk evaluation perspective

Camille Risi

February 10th - August 10th

Summary

During my 6-month internship, I was involved in a project to assess long-term wind risks related to tropical cyclones. While the tropical cyclone natural frequency is quite low and relatively few cyclones have been recorded in the past, this project relies on the synthesis of a large number of tropical cyclones that statistically conform to real ones. Cyclone tracks are first generated through a statistical model based on the historical cyclone track record. A deterministic and dynamical intensity and wind model is then applied along these synthetic tracks.

I worked on the development and implementation of the track synthesis algorithm. Cyclone tracks are modeled as a Markov chain. Genesis and conditional probabilities are first estimated from the historical cyclone tracks database. Tracks are then created during a Monte Carlo simulation.

Main problems faced are on the one hand the limited number of recorded tracks in the database from which to estimate probabilities, and on the other hand, the excessively large size of the probability arrays to estimate and store.

Résumé

Au cours de mon stage de 6 mois, j'ai participé à un projet d'estimation des risques, à long terme, de vents liés aux cyclones tropicaux. Tandis que la fréquence naturelle des cyclones tropicaux est assez faible et que relativement peu de cyclones ont été enregistrés dans le passé, ce projet repose sur la synthèse d'un très grand nombre de cyclones conformes statistiquement aux réels. Les trajectoires de cyclones sont d'abord synthetisées selon un modèle statistique fondé sur la base de données des cyclones enregistrés dans le passé. Le long de celles-ci est ensuite appliqué un modèle dynamique et déterministe d'intensité et de vents.

J'ai travaillé au cours de mon stage sur le développement et l'implémentation de l'algorithme de synthèse de trajectoires. Les trajectoires de cyclones sont modélisées sous la forme d'une chaine de Markov. Les probabilités conditionnelles et de génèse sont tout d'abord calculées à partir des données historiques. Les trajectoires sont ensuite créées au cours d'une simulation de Monte carlo.

Les principaux problèmes auquels j'ai du faire face sont d'une part le faible nombre de trajectoires enregistrées dans la base de donnée, à partir desquels estimer les probabilités, et d'autre part la taille excessivement importante des tableaux de probabilités à estimer et stocker.

Introduction

My 6-month internship took place at the Massachusetts Institute of Technology under the supervision of Dr. Kerry Emanuel, within the Program in Atmosphere, Oceans and Climate of the Department of Earth, Atmospheric and Planetary Sciences. I got involved in a project to assess long-term wind risks related to tropical cyclones.

A *tropical cyclone* (TC) is a cyclonically rotating storm with typical horizontal scales of 100-1000 km and which extends through the depth of the troposphere. Tropical cyclones originate over tropical oceans, are driven principally by heat transfer from the ocean surface, and die over cold water or land ([Emanuel, 2003, Emanuel, 1998]). They are also called *hurricanes* in North Atlantic and *typhoons* in Western Pacific.

Tropical cyclones are among the most destructive and lethal natural disasters. They cause loss of lives and enormous property damage around the world every year. While greatest risks to life commonly results from flooding and landslides following storm surges or torrential rains (300,000 people are estimated to have died from a TC-associated storm surge in Bangladesh in 1970 ([hrd, 2004])), the largest part of material and financial damage are usually caused by hurricane-related high winds, and resulting roof loss, building failure and air-borne debris ([Wyatt, 1999, Cochran, 2000]). As an example, the losses Hurricane Andrew in 1992 is responsible for reached US $26.5 billion ([hrd, 2004]).

It is therefore of high importance not only to be able to precisely forecast approaching TCs on the short term, but also to assess long term risks in perspectives of reinsurance, emergency planning, population warning in vulnerable coastal areas, land management and building codes establishment.

Generally, long term hurricane assessment models consist in, on the one hand, a meteorological component, to assess wind or storm surges probabilities, and on the other hand, loss projection models that take into account environment vulnerability ([Powell and coauthors, 2003]). The project I was involved in focuses on the meteorological component. The goal is to estimate long term TC-related wind probabilities through a simulation of a large number of TCs. TC tracks are synthesized through a statistical model. A dynamical intensity model is then applied along these synthetic tracks. I developped and implemented the track simulation algorithm.

This report is organized as follows: in the first section, I introduce the database used and explain how the track simulation algorithm I worked on integrates into the context of the wind risk assessment project. In a second section, I describe the methodology of this model, and techniques used to cope with the problem of the limited historical record to draw statistical inference from. In a third section, I present results obtained by this track simulation algorithm. Finally, I'll give a discussion of these results and the methods used, both from the quality of the synthetic tracks and in the perspective of the next steps toward the achievement of the risk assessment project. I will also describe some improvements that could have been brought to this algorithm, as well as alternatives to this Markov chain track model.

Contents

1 Presentation of the track simulation algorithm into the context of the wind risk estimation project **5**
- 1.1 The tropical cyclone database . . . 5
- 1.2 Previous methods used to estimate wind probabilities from the tropical cyclone database . . . 5
  - 1.2.1 Traditional methodology: wind field reconstruction of historical local cyclones . . . 5
  - 1.2.2 Problems related to the limited number of samples in the tropical cyclone database . . . 6
  - 1.2.3 Interest of Monte Carlo techniques to deal with the limited historical record . . . 6
- 1.3 The statistical-dynamical approach of the project . . . 6
  - 1.3.1 Methodology . . . 6
  - 1.3.2 The dynamical intensity model . . . 7
  - 1.3.3 The track simulation models . . . 7

2 The statistical track simulation algorithm: methodology and techniques **8**
- 2.1 Methodology: modeling tracks as Markov chains . . . 8
- 2.2 Initiation probabilities . . . 8
  - 2.2.1 Computation of the genesis probabilities: importance of smoothing . . . 8
  - 2.2.2 Sampling from the genesis probabilities . . . 9
  - 2.2.3 Uncertainties of genesis location before the satellite era . . . 9
- 2.3 Transition probabilities . . . 9
  - 2.3.1 Discretization of predictors . . . 10
  - 2.3.2 Choice of predictors and predictands . . . 11
  - 2.3.3 Computation of the transition probabilities . . . 15
  - 2.3.4 Sampling from the transition transition probabilities algorithm . . . 18

2.4 Termination probabilities . . . 19

3 Results: comparison of synthetic tracks to HURDAT tracks 19
3.1 Synthetic tracks' general aspect compared to HURDAT . . . 19
3.2 Comparison of speed, direction, acceleration and direction rate of change global and spatial distributions . 19
3.2.1 Termination pdf used to avoid the termination bias when comparing synthetic tracks and HURDAT tracks . . . 20
3.2.2 Comparison of synthetic tracks to all HURDAT tracks . . . 21
3.2.3 A significant difference between pre- and post-1970 HURDAT tracks . . . 21
3.2.4 A local example: tracks within 300km from Boston . . . 22
3.3 Comparison of the mutual information matrices between synthetic tracks and HURDAT tracks . . . 23
3.3.1 The mutual information matrix as a similarity measure between HURDAT and synthetic tracks . . . 23
3.3.2 Interpretation . . . 23
3.4 Number of sample to consider in a set . . . 24

4 Discussion of the method and alternatives 24
4.1 What could have been improved . . . 24
4.2 Advantages and drawbacks inherent to the methodology from the quality of synthetic tracks point of view . 26
4.2.1 Advantages and drawback of discretization . . . 26
4.2.2 Advantages and drawback of estimating all pdfs in advance . . . 26
4.2.3 An alternative with neither discretization nor pdf pre-estimation . . . 26
4.2.4 Uncertainties on track location before the satellite era . . . 29
4.3 Drawback of this track simulation method in the perspective of intensity computation along the tracks . . . 29
4.3.1 incompatibility between track and wind shear and its consequences . . . 29
4.3.2 An alternative: a more dynamical track simulation model based on the steering concept . . . 30
4.3.3 Comparison of tracks simulated through the purely statistical model and through steering concept in the perspective of the wind risk estimations . . . 31

Appendix 33

A Markov Chains 33

B Parametric and non-parametric representation of probability density functions 33

C Sampling from a probability density function 34

D Smoothing 34

E Mutual information and its measures 35

F Measures of similarity between two distributions 36

1 Presentation of the track simulation algorithm into the context of the wind risk estimation project

In the perspective of long term risk assessment, the goal is to accurately estimate TC-related wind speed probabilities for any given coastal location. A variety of methods exist to estimate these probabilities, but all draw statistical inference from the past tropical cyclone database.

1.1 The tropical cyclone database

The TC database used by all authors and which we also use is the “best-track” dataset. This historical record is a compilation of the past TCs’ geographical positions (latitudes and longitudes) and intensities[1] at 6 hour intervals.

For a given TC, the “best track“ is a post-storm reanalysis of the track, considering tools such as satellite imagery, radar depictions, or in earliest TCs, observation from ships or land stations, to determine the storm’s most likely tracks and intensity ([Jarvinen et al., 1984]).

The “best track” data-set is available for all basins. For the Northern Atlantic basin, for example, the database is called HURDAT and is maintained and updated by the National Hurricane Center in Miami. 1288 tracks are recorded form 1851 to 2002. 60 such tracks are shown in figure 1.

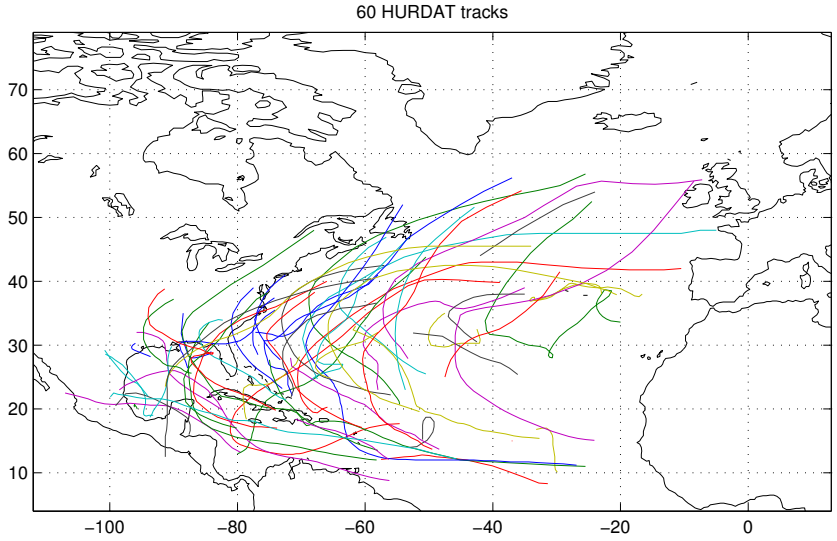

Figure 1: *60 random tracks drawn from the 1288 HURDAT tracks: the North Atlantic TC database.*

1.2 Previous methods used to estimate wind probabilities from the tropical cyclone database

1.2.1 Traditional methodology: wind field reconstruction of historical local cyclones

The most straight-forward methodology to estimate TC-related wind probabilities at any coastal location is summarized in figure 2 a.

To estimate wind speed probabilities in a particular location, accounting for spatially varying TC climatology, only historical TC tracks passing within a given radius (the scan radius) of that location are selected. Within this subregion, TC characteristics are assumed uniform.

Since the TC database only yields information about track and intensity, an empirical or analytical wind profile model is applied to deduce the wind caused by these selected historical TCs at the site of interest. From the obtained time series of winds at the site of interest, the wind speed probabilities are computed.

[1]The intensity of a TC is its maximum 1-min or 10-min averaged wind speed at an altitude of 10m. It can also be estimated from minimum central pressure, also available for most recent tracks.

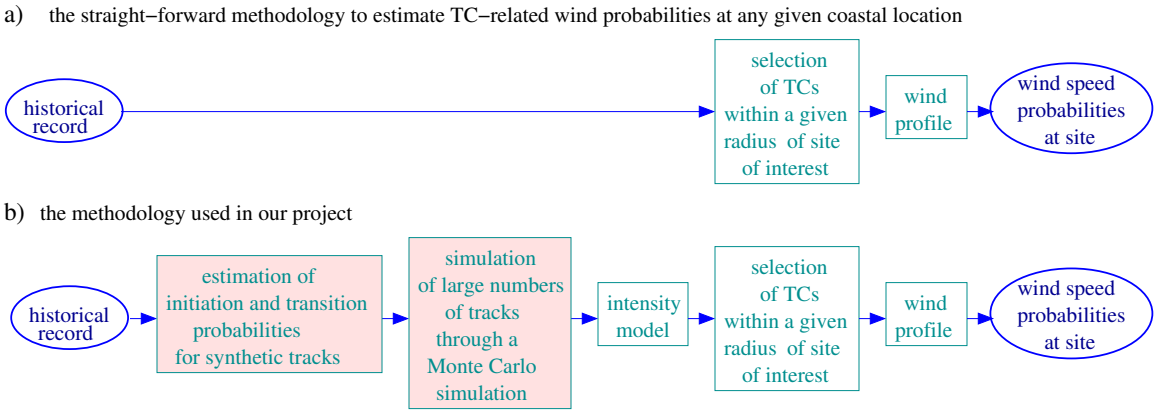

Figure 2: *Scheme comparing the most straigt-forward technique to estimating TC-related wind probabilities (a) to our method (b). Round blue boxes are the initial input (the historical record of TCs) and final output (the wind probabilities at any location). The part of the project I worked on is emphasized in pink filling.*

1.2.2 Problems related to the limited number of samples in the tropical cyclone database

However, the historical record is very limited for statistical inference: for the North Atlantic basin, for example, only 1288 tracks are available. This is particularly problematic when studying restricted areas: for example, only 25 tracks pass within a 100km radius of Boston. When drawing statistical inference from it, we might confuse sampling artifacts and actual TC patterns.

In addition, data becomes less certain the earlier it occurred in the record ([Murnane, 2000]). For example, there are uncertainties about position and intensity and it is likely that many storms went undetected at least partially before the 1970s, that is, before systematic satellite observations ([Jarvinen et al., 1984]).

Moreover, it would be impossible that way to estimate return periods for TC strength beyond the range of observation ([Darling, 1991, Chu and Wang, 1998]).

1.2.3 Interest of Monte Carlo techniques to deal with the limited historical record

Therefore, many authors resort to the use of Monte Carlo simulations to artificially enlarge the number of maximum winds from which exceedance probabilities are calculated. A Monte Carlo simulation refers to the use of a computer random number generator to simulate random processes, and is often used to find empirical solutions to rather complex mathematical problems ([Neumann, 1987]).

For example, some authors ([Russel, 1971, Batts et al., 1980, Georgiou, 1985, Rupp and Lander, 1996, Chu and Wang, 1998, Jagger et al., 2001]) fit an extreme value distribution ([Embrechts et al., 1999]) to the maximum wind distribution obtained from the historical record.

Some others ([Neumann, 1987]) calculate probability density functions (pdfs) of TC characteristics (such as distance of closest approach, intensity, radius of maximum winds...) from the historical storms within the scan radius of the site of interest. From these pdfs large numbers of characteristics are sampled and winds calculated.

Vickery ([Vickery et al., 2000]) suggests to directly enlarge the TC database in itself. A large number of synthetic TC tracks and intensities are produced through a statistical model based on a set of regression equations fitted to the historical record.

1.3 The statistical-dynamical approach of the project

1.3.1 Methodology

The method we use here is close to Vickery's in the sense that we also create a very large number of synthetic storms. However, since only 42% of the historical TCs recorded in the HURDAT database occurred after 1950, period in which the Doppler radar started to be used to estimate wind speeds ([Emanuel, 2003]), a large proportion of historical TCs have uncertainties on intensity records. Consequently, drawing statistical inference for storm intensity is problematic. So contrary to Vickery who synthesized both tracks and intensity through an all-statistical model, here only the tracks are created statistically while intensity is estimated deterministically by a dynamical model (figure 2b).

This approach, named statistical-dynamical, requires 2 components:

1. A statistical track simulation model, to create large numbers of different and diverse tracks whose key motion properties (geographical positions, translation speeds, headings, etc) conform statistically to historical data.
2. A dynamical, deterministic intensity model, which is ran along the tracks.

1.3.2 The dynamical intensity model

The dynamical model used is the Coupled Hurricane Intensity Prediction System (CHIPS)([Emanuel et al., 2003]). The input of this 3-D axisymetric balance model are:

1. monthly mean potential intensity[2] maps, calculated from the NCEP reanalysis data.
2. wind at 250hPa and 850hPa: this allows to approximate vertical wind shear, which is known to have a strong influence on TC intensity ([DeMaria and Kaplan, 1994]).
3. upper ocean thermal structure: the intensity model is coupled with a 1D ocean model, to account for the feedback from the ocean.
4. the TC's track.

Creating the input tracks is the role of the track simulation models, on which I focused during my internship.

1.3.3 The track simulation models

Two kind of track simulation approaches exist: statistical and dynamical. Both of these two approaches are used in the risk assessment project.

Statistical models The track generation technique I worked on is statistical. Statistical models draw inference from the statistical record to relate current and previous TC position (called *predictors*) to the next position and behavior (*predictands*: what we try to predict).

For example, Neumann ([Hope and Neumann, 1970]), in the HURRAN (Hurricane Analog) model, suggested to search historical analogs to the track being simulated, and calculate probabilities for the simulated tracks' next position from these historical analogs.

He later suggested a multiple regression model, CLIPER (Climatology-Persistence)([Neumann, 1972]) based on a set of polynomial regression to relate a selected number of predictors to predictands. Predictands consist in the TC next translation speeds and headings. Predictors include both climatology (geographical position) and persistence (TC's previous positions and displacements). Vickery ([Vickery et al., 2000]) uses a "simplified CLIPER" approach in which fixed predictors and predictands are linked by a set of linear regressions depending on geographical location.

The HURRAN and CLIPER models were initially devised for real time forecasting of particular approaching tracks. On the contrary, our goal is to synthesize large numbers of diverse and random tracks. Therefore, contrary to these traditional methods, we chose to generate tracks as Markov chains.

Dynamical models Dynamical models predict track movements from meteorological variables, such as environmental wind fields, making use of the "steering concept" ([hrd, 2004]), in which TCs are assumed advected by the environmental flow, with a correction to account for the beta effect[3]. In the project, such a approach was also used, with a statistically generated wind field.

Although the projects applies to all TCs basins (North Atlantic, Western and Eastern Pacific, Indian Ocean, Southern Hemisphere), all explanations and illustrations will be given for the North Atlantic basin, for which models were initially developed.

[2] When modeling a TC as an heat engine suggested as suggested by Emanuel ([Emanuel, 1988b, Emanuel, 1988a, Emanuel, 1995, M. and Emanuel, 1998]), the maximum intensity achievable (called potential intensity) by the TC is computable from the temperatures of the heat source (the sea) and the cold source (the stratosphere).

[3] The beta-effect is the effect the TC has on the environmental vorticity field. The disturbance of this vorticity field results in steering the TC towards the North-West at about 1 to 3 m/s.

2 The statistical track simulation algorithm: methodology and techniques

2.1 Methodology: modeling tracks as Markov chains

I developed a statistical track algorithm based on modeling tracks as Markov chains ([Vivant, 2003]). Definition and main properties of Markov chains are presented in appendix A.

The Markov chain assumption is supported by the fact that the autocorrelation spectra of key motion variables (e.g. translation speed, heading) become weak after four to five priors, indicating that the number of priors to condition the transition probabilities is likely to be finite.

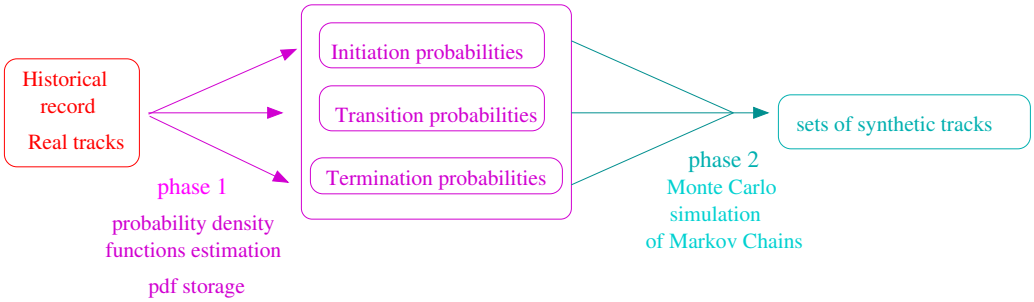

Figure 3: *Scheme summarizing the overall methodology used to synthesized tracks as Markov chains.*

Two sets of probability density functions (pdfs) must be calculated to create tracks as Markov chains:

- initiation pdfs, to initiate the tracks (e.g. first position and day of year)
- transition pdfs, to elongate the tracks: this is where predictands will be related to the predictors.

To these two sets of pdfs is added a third pdf: the termination pdf to stop the tracks.

All these pdfs are computed from the TC database, so as to produce tracks that conform to it.

Markov chains are generated through a Monte Carlo simulation. To save execution time during the simulation of large numbers of tracks, these pdfs are stored in advance (phase 1 on figure 3). From these pdfs are then sampled the successive positions through a Monte Carlo simulation (phase 2 on figure 3), as shown in figure 4.

2.2 Initiation probabilities

Tracks are initiated by their first position x_0, y_0 and date in year t_0 through the initiation probability $P_0(x_0, y_0, t_0)$.

2.2.1 Computation of the genesis probabilities: importance of smoothing

Let the space-time field of tropical cyclone genesis events be discretized to a $r_x \times r_y \times r_t$ grid, where r_x, r_y and r_t are respectively the resolutions in longitude, latitude and time of year. In practice, $r_x = r_y = 0.5^o$ and $r_t = 5 days$. The most straightforward way to calculate genesis probabilities would be to just count the number of genesis events in HURDAT that occurred in each box.

However, this is not satisfying. Figure 5a shows the marginal distribution $P_0(x_0, y_0)$ obtained by this method. The initiation probability was integrated over the year for illustration purpose. The distribution is very sparse and discontinuous. It reflects more the sampling artifacts than the actual underlying genesis distribution, which is expected not to vary as much at such a short spatial scale. For example, only 9 historical genesis events occurred above 40^o latitude. The raw histogram interprets this as a null probability everywhere above 40^o latitude, except for these 9 bins, in which the probability is bigger than the average. However, a more physical way to interpret this would be to set a very low but relatively uniform probability in the whole area.

Therefore, smoothing is used to derive a genesis distribution P_0 from the raw histograms H (see appendix D for a definition of smoothing). For all cells (x_l, y_m, t_n),

$$P_0(x_l, y_m, t_n) = \sum_{i,j,k} H(x_i, y_j, t_k) . G(x_l - x_i, y_m - y_j, t_n - t_k)$$

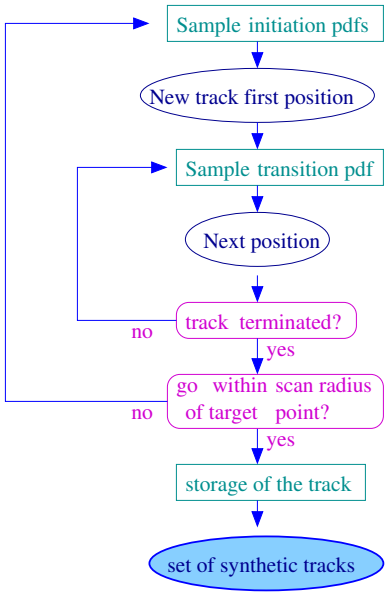

Figure 4: *Scheme summarizing the Monte Carlo simulation algorithm used to create tracks as Markov chains once initiation, transition and termination pdfs are estimated from the historical record. Inputs/outputs are in blue, algorithms in green and conditions in purple. The final output (the set of synthetic tracks) is emphasized.*

where G is the Gaussian smoothing kernel (appendix D).

Figure 5b shows the improvements brought by a spatially-isotropic constant Gaussian kernel smoothing compared to raw data (figure 5a). σ_x and σ_t, respectively the standard deviation in space and time dimensions, are constants. But we can see that the general aspect is still sparse.

Consequently, σ_x is calculated as a function of (x,y), so as to adapt to the data spatial density: it is all the widest as there are few data. σ_x is proportional to the size of the smallest square that contains at least a threshold of data points. σ_t arbitrarily remained constant: $\sigma_t = 5\,days$.

The result is much more precise in the tropics and more continuous in high latitudes, as shown in figure 5c.

This smoothing was retained in the final version of the program and used for the spatial smoothing of all pdfs.

2.2.2 Sampling from the genesis probabilities

Genesis events (x_0, y_0, t_0) are sampled from the genesis pdf P_0 using a 3-D "hit and miss" algorithm (recalled in appendix C).

2.2.3 Uncertainties of genesis location before the satellite era

Before the systematic satellite observation of TCs started to be performed in the 1970s ([Emanuel, 2003]), events were considered to have formed where they were first observed, by ships or coastal and island stations, both of which are distributed inhomogeneous in space. Indeed, there's a significant difference between the spatial distribution of genesis events for all tracks from 1851 to today and the tracks after 1970 only (figure 5d): tracks are now observed much earlier and are therefore much longer. This is of importance because when the intensity model is ran along the tracks, wind speed is initiated at 25knots, so the intensity of synthetic storms whom genesis positions are sampled from early historical TCs are biased. Consequently, the genesis pdf constructed from the satellite era TCs only should be used.

2.3 Transition probabilities

The transition probability to calculate is $P(X_{i+1} \mid X_i)$, where X_i is the state variable of the Markov chain. Inspired from the CLIPER notation ([Neumann, 1972]), let's call *predictand* (resp. *predictor*) the term on the left (resp. right) of the conditional probability: $P(predictands \mid predictors)$.

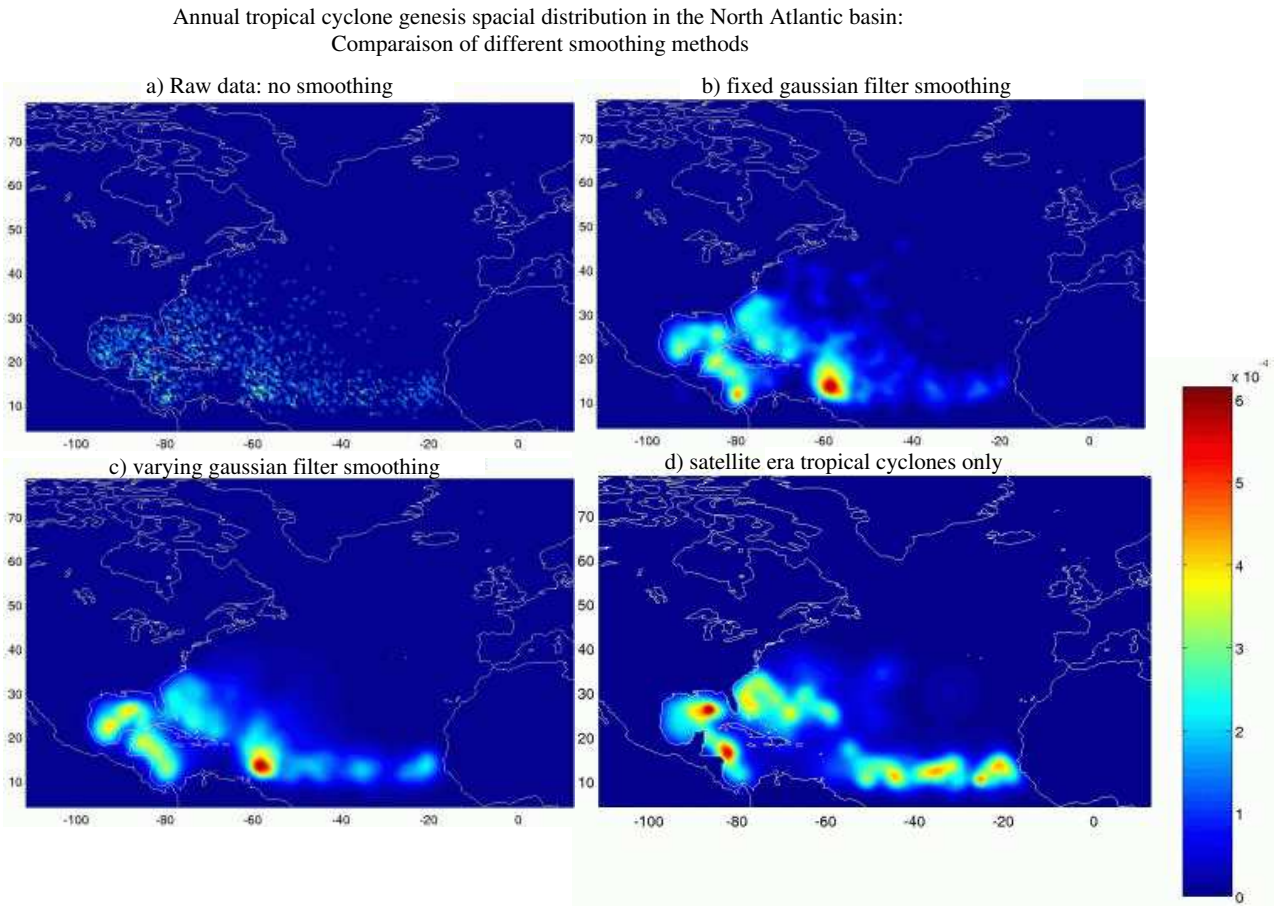

Figure 5: *Comparison of different annual genesis pdfs:*
a) no smoothing: counting genesis events in a 0.5^o *grid.*
b) 0.5^o *resolution with fixed-sized spatially isotropic Gaussian filter: standard deviation is* 1^o.
c) 0.5^o *resolution with varying-sized spatially isotropic Gaussian filter: standard deviation varies from 0.5 to* 7^o *according to genesis events density.*
d) same as c, but only for satellite era TCs (TCs that occurred after 1970).

2.3.1 Discretization of predictors

The transition probabilities are stored as follows: the predictors are discretized into a n-dimensional array, with n the number of predictors taken into account. Each cell of this multi-dimensional array represents a combination of predictors and contains the pdfs for the various predictands.

Although this storage method adds imprecisions from the discretization of predictors, it avoids the imprecisions brought by fitting regressions equation, like in a CLIPER-like approach ([Neumann, 1972]). The discretization approach is also faster to compute.

However, it is considerably space-consuming. Indeed, tracks are complicated processes that require lots of predictors to take into account. In the CLIPER model, for example, current and previous positions and displacements were taken as predictors. This could easily make the size of the multi-dimensional array very large, and, as summarized in figure 6, this is not desirable:

1. For memory reasons. All transition probabilities together could easily surpass the RAM capacity and even the disk memory capacity.

2. To insure a transition pdf is defined in all situation: if the predictor array is too large, since the historical record is limited, there won't be enough data to populate them. For example, during the simulation, one can fall in a situation in which no transition pdfs are defined, or lack statistical significance. This is very problematic, because such a track

would abort.

3. To avoid too spatially varying pdfs: if pdfs are calculated from only 1 or 2 data points, they are likely not to reflect the underlying distribution and to vary to chaotically with space.

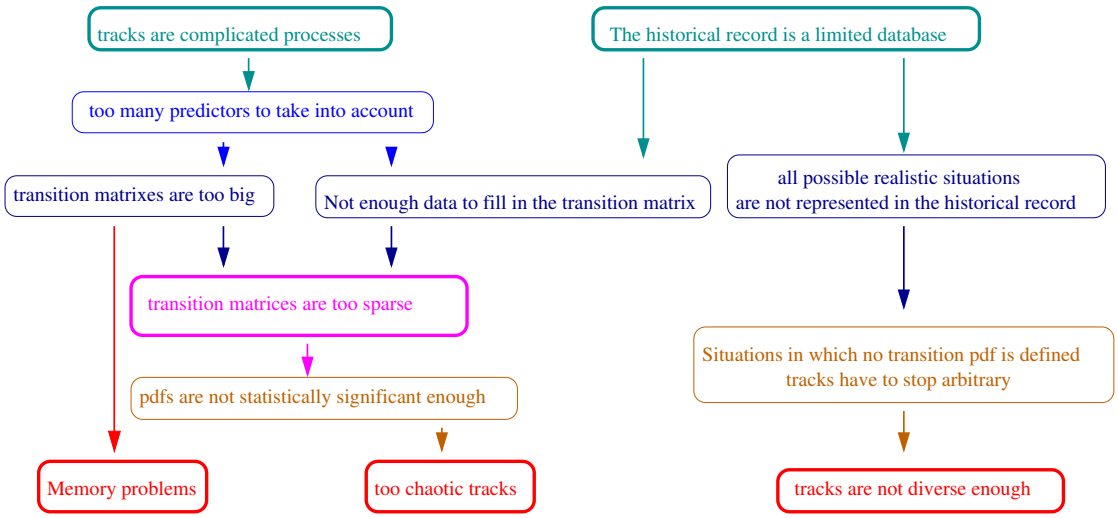

Figure 6: *Scheme explaining the difficulties raised by the methodology and their consequences.*

Number of priors to consider In the case of a track, the state of a Markov chain would naturally be defined as $X_i = (x_i, y_i, t_i)$, where x_i, y_i and t_i are respectively longitude, latitude and day of year. However, as explained in appendix A, the Markov chain assumption is generalizable to any finite number of priors by defining a new state $Y_i = (X_i, X_{i-1}, ...)$. As observed on figure 1, tracks are strikingly smooth and continuous phenomena. Therefore, to reproduce this smoothness, more than one prior should be taken into account. Two priors seems a good balance between reproducing the tracks smoothness and reducing the array size, and was chosen in this model. The results of the simulation will rule whether this was a good approximation.

Definition of the Markov chain state variable The state variable of the Markov chain is therefore $X_i = \left((x,y,t)_i, (x,y,t)_{i-1}\right)$. However, in practice, there are several ways to store the transition matrix on the computer, depending on the choice of predictors and predictands. This choice is made so that the predictor array size is minimum.

2.3.2 Choice of predictors and predictands

Displacement speed vectors as predictands Suppose we consider as predictors and predictands geographical positions directly. Let $\vec{e}_i = (x_i, y_i)$ be the geographical position (notations and indexations are summarized on figure 7). Then the spatial resolution should be very high to capture the smoothness of the tracks, hence an extreme size of the predictor array.

On the contrary, the track aspect suggests that displacement speed vectors $\vec{u}_i$ are much more homogeneous spatially, as also underlaid in figure 8: typically HURDAT tracks initiate in the tropics and head west-wards at low speed ($\sim 100 km/6h$). Then the largest proportion of tracks suddenly recurve north-east-wards and start accelerating when reaching middle latitudes ($> 35^o$). They eventually arrive at high latitudes , decelerate again, weaken and die. Therefore, considering displacement speed vectors allows to lower spatial resolution, and consequently the multi-dimensional arrays' size.

Let us parametrize the displacement speed vector $\vec{u}_i$ as (s_i, θ_i), where s_i is the translation speed and θ_i the direction[4] ([Vivant, 2003]).

The reduction of the array size this change of predictands and predictors allowed can be assessed through the following resolution discussion. Let N be the predictor array size.

[4]Direction at a given position is here defined as the angle between the track speed vector at this given position and a vector pointing to the east. Angles are counted positively counterclockwise, and range in $[-180^o, 180^o]$. Translation speed at a given position is here defined as the track speed vector's norm at this given position, in km/6h. Translation speed will often be called just speed along the report.

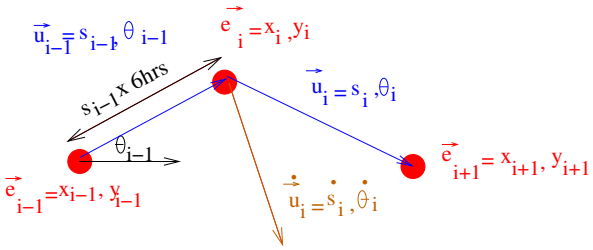

Figure 7: *Figure defining variable notations used and the way they are indexed. In red are represented the 6-hourly positions (x,y), in blue are the 6-hour-averaged speed vectors, defined by its norm (translation speed s) and direction* θ *, and in orange is the rate of change of the speed vector, defined by its norm (the acceleration* $\dot{s}$*) and direction (direction rate of change* $\dot{\theta}$*), and the way they are indexed.*

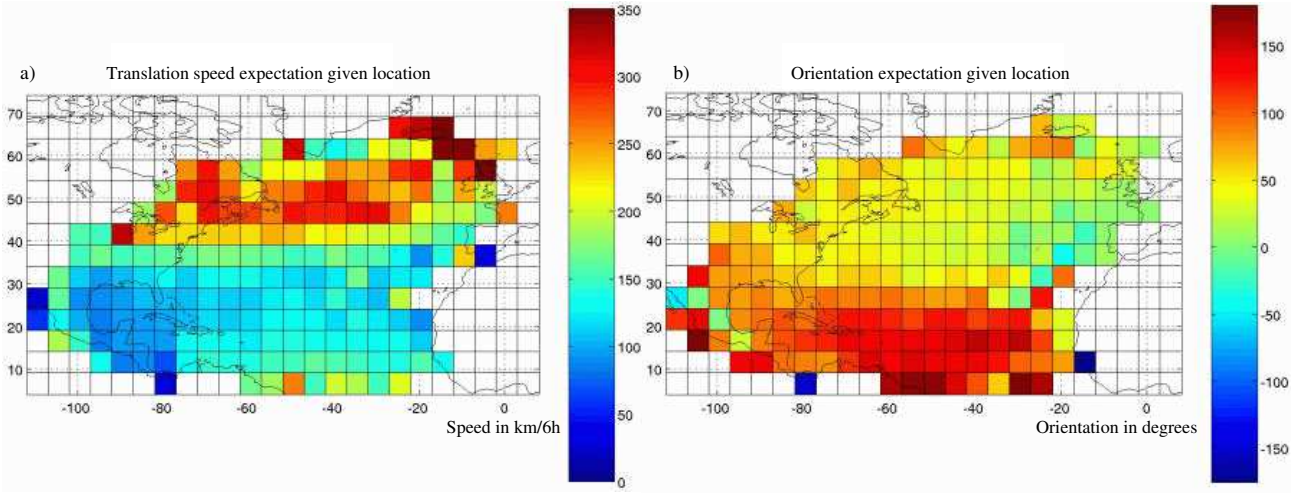

Figure 8: *HURDAT motion as a function of geographical location: a) speed spatial distribution. b) direction spatial distribution. Maps were calculated directly from HURDAT, by computing the expectation for each* 5×5^o *box.*

Suppose we directly consider geographical positions and express the transition probability with 2 priors as $P(\vec{e}_{i+1} \mid \vec{e}_i, \vec{e}_{i-1})$. For tracks to look smooth, it is observed that a spatial resolution of at least 0.1^o is required. Let the North Atlantic basin spread from -110^o to 10^o in longitude and from 0 to 60^oin latitude. Then we have $N = (1200 \times 600)^2 \simeq 5 \times 10^{11}$. On the contrary, suppose we consider the transition probability $P(\vec{u}_i \mid \vec{e}_i, \vec{u}_{i-1})$. Speeds range between 0 and $800\,km/6h$ approximately, and directions between -180^o and 180^o. To insure the same 0.1^o precision, the resolution in speed should be approximately $10\,km/6h$ and 5^o in direction. Besides, a spatial resolution of 0.5^o is sufficient to capture the spatial variations of the displacement speed vectors. Hence, a array size of $N = 240 \times 120 \times 80 \times 72 \simeq 2 \times 10^8$cells.

Rate of change of displacement speed vectors as predictands Figure 9 shows the conditional probabilities $P(s_i \mid s_{i-1})$ and $P(\theta_i \mid \theta_{i-1})$. One notice that all values are clustered around the first diagonal: speed and direction are very smoothly varying. Therefore, most of the space is wasted in this representation. A more efficient representation would be to consider the variations from the first diagonal, that is, take the derivatives: $\dot{\vec{u}}_i = (\dot{s}_i, \dot{\theta}_i)$, where $\dot{s}_i$ is the acceleration and $\dot{\theta}_i$ the direction rate of change:

$$\vec{u}_i = \vec{u}_{i-1} + \dot{\vec{u}}_i \tag{1}$$

This short array size calculation shows the improvements brought by considering the transition probability $P(\dot{\vec{u}}_i \mid \vec{e}_i, \vec{u}_{i-1})$: to get the same 0.1^o precision on $\vec{e}_{i+1}$, as previously explained, the resolution on $\vec{u}_i$ and thus $\dot{\vec{u}}_i$ should be of $(10\,km/6h, 5^o)$ approximately. However, a resolution of $40km/6h$ and of 20^o for speed and direction respectively is observed to be sufficient to have the same resolution if calculating $\vec{u}_i$ from equation 1. Consequently, $N = 240 \times 120 \times 20 \times 18 \simeq 10^7$.

To conclude, to limit the predictor array size, the transition probability is stored as:

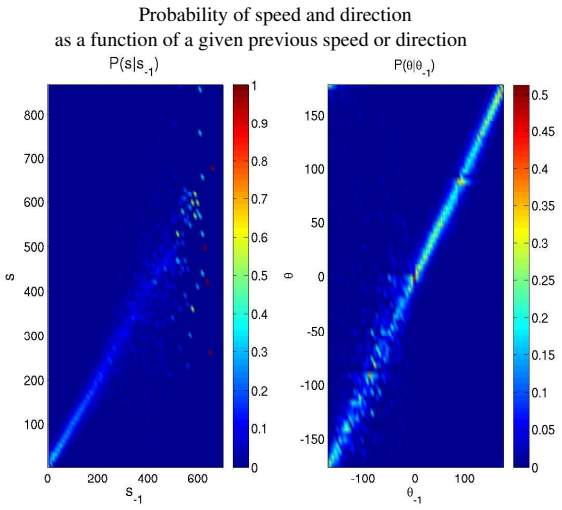

Figure 9: *a)* $P(s_i|s_{i-1})$ *and b)* $P(\theta_i|\theta_{i-1})$ *. The main information conveyed by these pdfs is that speed and direction should vary slowly with time, as proved by the diagonal aspect of the plots.*

$$P\left(\dot{s}_i, \dot{\theta}_i \mid t_i, x_i, y_i, s_{i-1}, \theta_{i-1}\right)$$

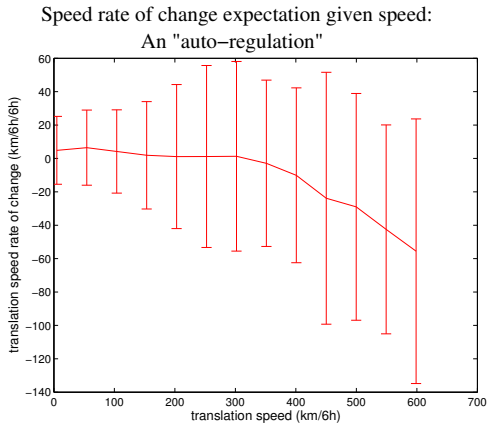

Figure 10: *Speed rate expectation given previous speed* $E\left(\dot{s}_i \mid s_{i-1}\right)$*: speed rate seems to act as a speed regulation: for example, storms tend to accelerate when their speed is low and to decelerate when their speed is high.*

A risk of divergence A risk we face when choosing acceleration and direction rate of change as predictands is that some values of translation speed might diverge. I tried to introduce a retro-action on speed and orientation, but it gave quite poor results.

Actually, it turns out that no retro-action was needed. For example, in the case of speed, the relationship between $\dot{s}_i$ and s_{i-1} already acts as a natural retro-action, as shown in figure 10: TC tend to accelerate when they are slow and decelerate when they are fast. Such a retro-action holds for orientations as well. To insure that this retro-action is efficient, s_{i-1} and θ_{i-1} were always used as predictors, with the relatively high resolution of 40km/6h in s_{i-1} and 20^o for θ_{i-1}.

Minimizing the bias due to termination pdfs The choice of acceleration rather than speed directly as predictand has another advantage, from a more theoretical point of view.

As already explained in section 2.3.2, typically, storm tracks, in the Atlantic basin for example, initiate in the tropics, go west-wards and slightly north-wards quite slowly, then suddenly recurve north-wards and eastwards and accelerate when

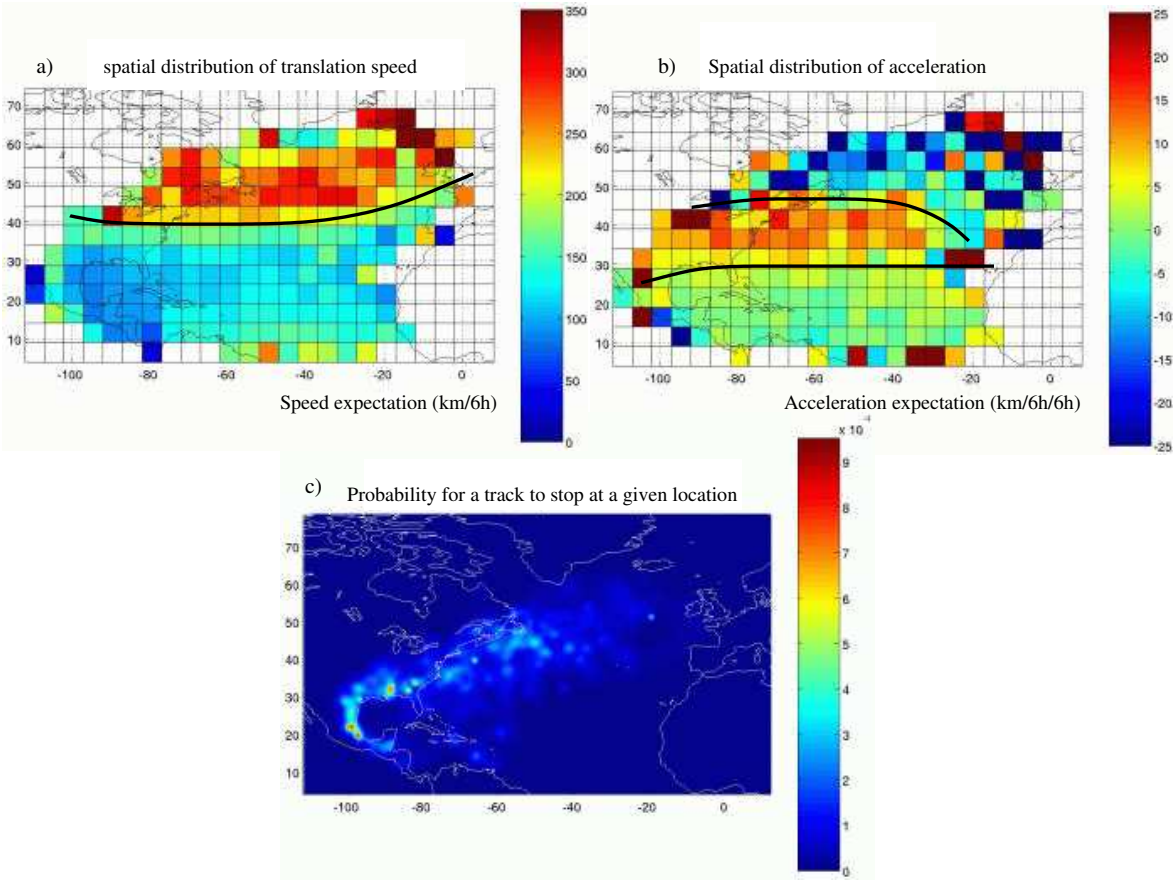

Figure 11: *a) Annual spatial distribution of the speed expectation, phrased on a* 5×5^o*grid with no smoothing. 2 distinctive belts can be identified, as plotted in black.*
b) Annual spatial distribution of the acceleration expectation, phrased on a 5×5^o*grid with no smoothing. 3 distinctive belts can be identified, as plotted in black.*
c) Probability for a track to die as a function of location, phrased on a 0.5×0.5^o *grid with the same smoothing as used for the computation of the genesis pdfs. Tracks are most likely to die over land or cold water.*
The fact that real tracks that move slowly out of the tropics are more likely to die leads to a significant difference between the acceleration spatial distribution and the speed distribution.

reaching latitudes about 30 - 35^o, and eventually slow down when reaching latitudes about 50^o .

However, while the acceleration map clearly shows these 3 belts (figure 8b), the speed mean map displays only 2 belts: low speeds bellow 40 - 45^o latitude, and high speeds above (figure 8a). This is because slow storms that have spent a long time over cold water die (figure 8c), so even if storms tend to decelerate in high latitudes, only fast storm survive, thus the high speeds in high latitudes.

Consequently, drawing inference directly from track speeds should lead to an overestimation of storm acceleration in mid-latitudes, and a underestimation of storm deceleration in high latitudes.

The bias at a given location (x,y) can be formulated as follow:

Let $P_{th}(s \mid x,y)$ be the theoretical speed pdf, from which the HURDAT speed distribution $P_H(s \mid x,y)$ is derived by truncating tracks if the event "dead" has occurred. Thus, we can calculate that

$$P_H(s \mid x,y) = P_{th}(s \mid x,y) \times \frac{1 - P_{th}(\text{"dead"} \mid s,x,y).P_{th}(x,y)}{1 - P_{th}(\text{"dead"} \mid x,y)} = P_{th}(s \mid x,y) \times f(s)$$

with $f(s)$ a normalizing factor that all the more disrupts the theoretical distribution as $f(s)$ varies a lot with s, that is, if $P_{th}(\text{"dead"} \mid s,x,y)$ varies a lot with s.

In our statistical model, statistical inference should be drawn from the unknown theoretical distribution, but is actually drawn from the HURDAT known distribution. In other words, we implicitly assume in this model that $P_H(s \mid x,y) \simeq P_{th}(s \mid$

$x,y)$, that is $P_{th}(\text{"}dead\text{"} \mid s,x,y)$ doesn't vary too much with s. This assumption is obviously not valid.

Suppose now we consider $\dot{s}$. The assumption made now becomes that $P_{th}(\text{"}dead\text{"} \mid \dot{s},x,y)$ doesn't vary too much with $\dot{s}$. Although acceleration $\dot{s}$ is related with speed s and therefore with the survival of a track in high latitudes, $\dot{s}$ is in practice much less related to the survival of the storm than the translation speed directly.

Therefore, the choice of acceleration as predictand minimizes the bias due to the termination pdfs.

2.3.3 Computation of the transition probabilities

As justified in section 2.3.2, the transition pdf to compute is

$$P\left(\dot{s}_i, \dot{\theta}_i \mid t_i, x_i, y_i, s_{i-1}, \theta_{i-1}\right)$$

One also notices that upon this state variable definition, the initiation pdf must contain the pdfs to sample the first speed s_0 and direction θ_0 as well. They are sampled from the conditional pdf $P(s \mid x,y,t)$ and $P(\theta \mid x,y,t)$. These conditional pdfs are estimated and sampled from using exactly the same methods and techniques as for the transition pdfs, detailed in this section.

2.3.3.1 Independence assumptions Suppose we directly compute pdfs as $P\left(\dot{s}_i, \dot{\theta}_i \mid t_i, x_i, y_i, s_{i-1}, \theta_{i-1}\right)$ for the Northern Atlantic basin. As explained in section 2.3.1, $N \simeq 1.10^7 \times N_t$ cells (if N_t is the number of bins in the time of year dimension). But the database only contains about $3.5.10^4$data points. Data is therefore not abundant enough to populate the transition matrix.

A way to reduce the pdf matrix dimension is to make some independence assumptions.

Speed-direction independence assumption For example, we can assume that s is independent from θ ([Vivant, 2003]):

$$P\left(\dot{s}_i, \dot{\theta}_i \mid t_i, x_i, y_i, s_{i-1}, \theta_i\right) \simeq P\left(\dot{s}_i \mid t_i, x_i, y_i, s_{i-1}\right).P\left(\dot{\theta}_i \mid t_i, x_i, y_i, \theta_{i-1}\right)$$

This allows a great reduction of the transition pdf array. This approximation is questionable: it is known that TC move much faster when heading north-eastwards. However, it can be justified by the fact that a large part of the relation between s and θ might be explained by the relation both have to latitude. Results of the simulations will rule whether this is actually a valid approximation.

Climatology-persistence independence assumption Another stronger assumption had also been put forward in earlier versions of the algorithm: assuming that probabilities given climatological factors and given persistence factors are independent. In these versions of the algorithm, the predictands used were actually s and θ . However, such an assumption applied to the predictands $\dot{s}$ and $\dot{\theta}$ would have the same affect and would be expressed as:

$$P\left(\dot{s}_i | t_i, s_{i-1}, x_i, y_i\right) \simeq P\left(\dot{s}_i | t_i, s_{i-1}\right).P\left(\dot{s}_i | t_i, x_i, y_i\right)$$

and

$$P\left(\dot{\theta}_i | t_i, s_{i-1}, x_i, y_i\right) \simeq P\left(\dot{\theta}_i | t_i, s_{i-1}\right).P\left(\dot{\theta}_i | t_i, x_i, y_i\right)$$

The speed and direction distributions of tracks created under this assumption differed from real tracks. For example, in the case of speed, high probabilities are exaggerated and low probabilities under estimated (figure 12 a). This is because the probability is calculated as a product. This proves that the transition probabilities given priors and given location are not independent. When calculating the real probability the program samples from, making the product of the two pdfs, we verify that it fits created tracks but it is different from the HURDAT distribution (figure 12 b, magenta curve).

This is particularly problematic in the perspective of wind risk assessment in extratropical locations such as Boston: only fast moving tracks are able to maintain a large intensity in such high latitudes. Therefore, an underestimation of high speeds would underestimate wind risks in such coastal locations. This approximation was therefore abandoned.

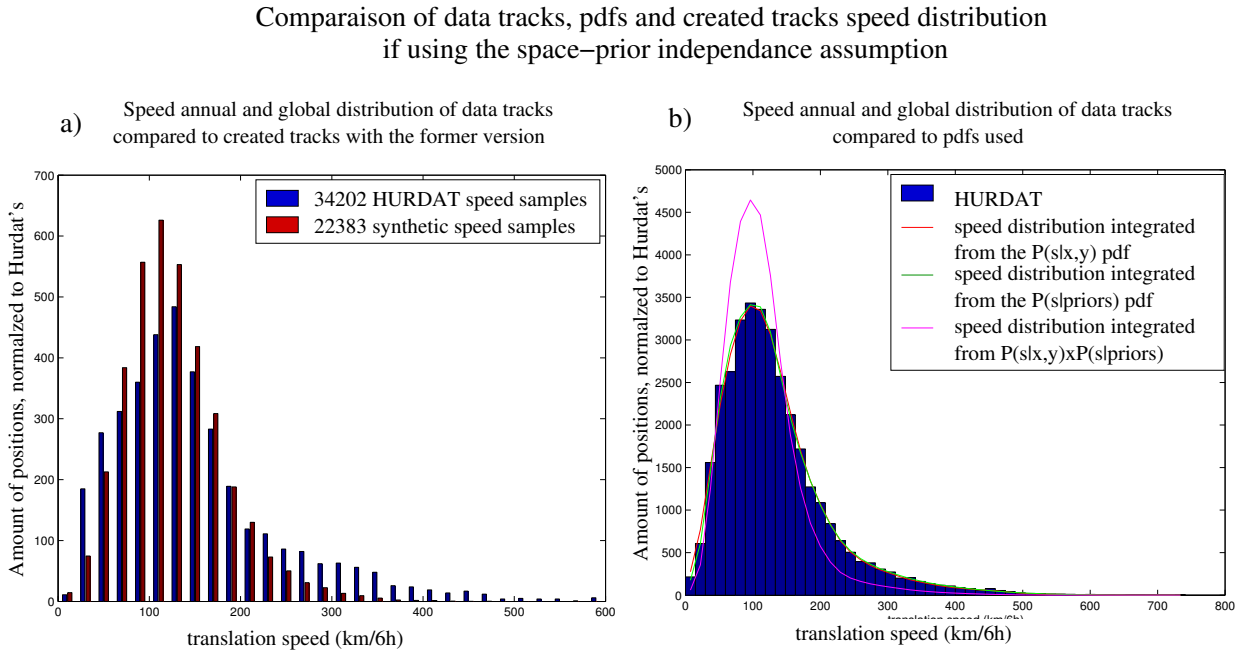

Figure 12: *a) Comparison of data tracks and created tracks annual and global speed distribution using the approximation that* $P(s_i|t_i,s_{i-1},x_i,y_i) \simeq P(s_i|t_i,s_i-1).P(s_i|t_i,x_i,y_i)$*. We notice that less likely speeds (low and high speeds) are underestimated by this model.*
b) Comparison of data speed distribution with pdfs the model samples from if using the space-prior independence assumption.
In red is P(s) calculated from the speed pdf given space by $P(s) = \sum_{x,y} P(s|x,y).P(x,y)$*. In green is P(s) calculated from the speed pdf given prior speed by* $P(s) = \sum_{s_{-1}} (Ps|s_{-1}).P(s_{-1})$*.*
In magenta is P(s) from which created tracks are sampled under the independence assumption, calculated by

$$P(s) = \sum_{x,y,s_{-1}} P(s|x,y).P(s|s_{-1}).P(x,y,s_{-1})$$

Although the pdfs given priors (green) and space (red) separately fit the HURDAT distribution, the effective pdf the model samples from (their product) is different.

2.3.3.2 kernel-smoothed histograms A way to make the transition pdfs smaller to store could be to store them parametrically (see appendix B for a short review of different ways to store pdfs). For example, in an earlier version of the algorithm, $P(\dot{s}_i \mid s_{i-1})$ and $P(\dot{\theta}_i \mid \theta_{i-1})$ were assumed Gaussian, since many of the pdfs did look Gaussian ([Vivant, 2003]). However, it turns out that it is not sufficient in many cases. Figure 13 shows the shape of 50 random conditional pdfs $P(\dot{s}_i \mid s_{i-1})$ (a) and $P(\dot{\theta}_i \mid \theta_{i-1})$ (b), centered on 0 for illustration purpose. One can see although most of these pdfs look Gaussian, some differ strongly from a Gaussian shape and have several peaks. For example, in the Gulf of Mexico, where tracks recurve, as already explained, the direction distribution should be bimodal, which can't be carried by a Gaussian pdf.

Moreover, the inaccuracy of this approximation can yield to more situations that never happened in the past increasing the number of cases in which pdfs no transition pdfs are defined and tracks abort.

Therefore, the non-parametric representation was chosen for acceleration as well as direction rate of changes. The histograms were then smoothed using a Gaussian kernel as explained in appendix B. Histogram resolution were 8km/6h/6h and 3^o for acceleration and direction rate of change respectively. The Gaussian smoothing kernel standard deviation was equal to the resolution in both cases.

2.3.3.3 Spatial smoothing As explained for initiation pdfs in section 2.2.1, spatial smoothing is a good way to reduce matrix sparsity and to avoid reflecting sampling artifacts in the desired pdf.

If resolution in space is too low, information is lost. If the resolution is too high, the matrices are too sparse, leading to situations where no pdf are defined, or with insufficient statistical significance. Therefore, the size of the smoothing kernel

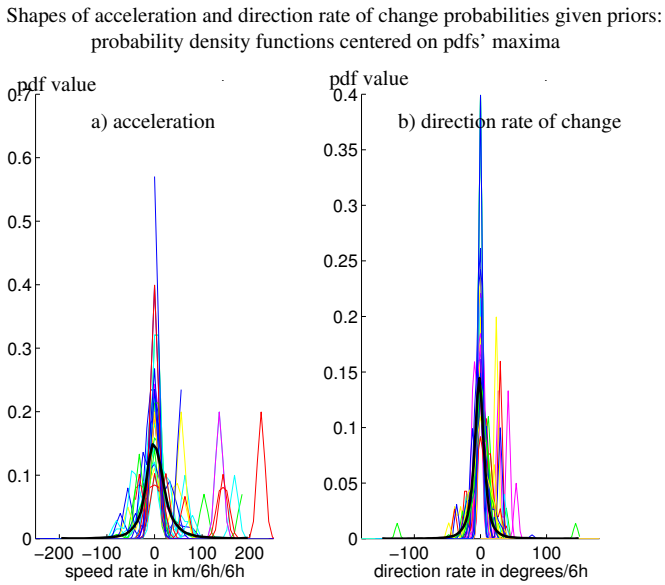

Figure 13: *Shape of 50 random conditional pdfs $P(\dot{s}_i \mid s_{i-1})$(a) and $P\left(\dot{\theta}_i \mid \theta_{i-1}\right)$ (b) . Pdfs were centered on 0 so that their maximum value was on 0 and shapes can easily be compared. Notice although most pdfs and their average (plotted in black) look Gaussian, a significant number of pdfs show more than one peak.*

varies in space to adapt to the density of data samples, and the same spatially varying Gaussian kernel presented in section 2 was used to smooth the pdfs in space. Arbitrarily, no smoothing was performed neither in time nor previous speed and direction dimensions.

To sum up, the formula used to compute the transition pdf is as follow: Let q stands for either s or θ. For a given time period t , previous speed or direction q_{-1}, position (x_m, y_n) and predictand $\dot{q}_l$:

$$P(\dot{q}_l \mid x_m, y_n, q_{-1}, t) = \frac{P(\dot{q}_l, x_m, y_n, q_{-1}, t)}{P(x_m, y_n, q_{-1}, t)} = \frac{\sum_{i,j,k} H(\dot{q}_k, x_i, y_j, t, q_{-1}) . e^{-\frac{1}{2}\left(\frac{(x_m - x_i)^2 + \left(y_n - y_j\right)^2}{\sigma_x\left(x_i, y_j\right)} + \frac{(\dot{q}_l - \dot{q}_k)^2}{\sigma_h}\right)}}{c(x_m, y_n, q_{-1}, t)}$$

where $H(\dot{q}_k, x_i, y_j, t, q_{-1})$ is raw number of data points falling in the cell $(\dot{q}_k, x_i, y_j, t, q_{-1})$, $c(x_m, y_n, q_{-1}, t)$ is a normalizing constant that insures that the conditional pdf integrates to 1, $\sigma_x(x_i, y_j)$ is the spatially varying standard deviation of the Gaussian smoothing kernel and σ_h is the standard deviation of the histogram smoothing kernel.

2.3.3.4 variable grid size To insure that a pdf meet statistical significance standards even in periods of the year when very few historical events were recorded, the resolution is made variable in the time dimension. The year is divided into 9 time periods (figure 14) containing at least a given number of data points.

To insure a physical significance as well, 3 other conditions were added, which prevail over the first one:

1. a maximum length, above which track behaviors get very different,
2. a minimum length, bellow which it would be waste of memory to make separate periods. A length of 15 days was chosen here,
3. the first period starting date is manually chosen so that it separate 2 time periods in which tracks show different behaviors.

2.3.3.5 Variable window method Multiple resolutions for the predictor array are available. That way, if no transition pdf is defined, or was estimated from to few data points, then a coarser resolution is used, and so on.

The resolution used is the finest possible that satisfied a threshold of statistical significance. To choose between lowering the resolution in one predictor or the other, priority rules are set. Priority rules used in standard simulations are shown in table 1.

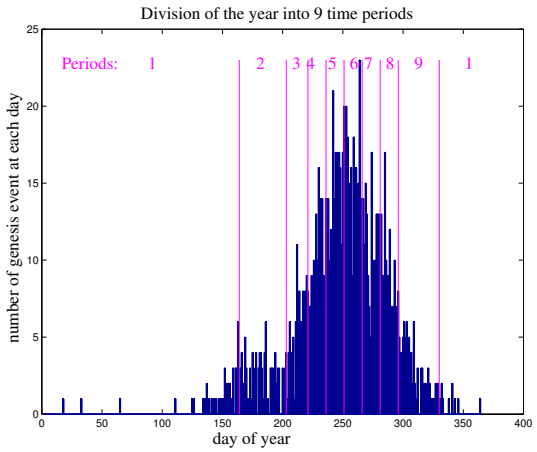

Figure 14: *Illustration of the variable bin size technique: division of the year into 9 time period, each containing a threshold of data tracks to insure statistical significance while reducing the transition matrix size.*

spatial resolution→ temporal resolution↓	0.5^o(with Gaussian smoothing)	5^o (no smoothing)	belts (no smoothing)	global
9 time periods	1	2	3	7
annual	4	5	6	8

Table 1: *simulation default priority rules for the transition pdfs of either acceleration or direction rate of change.*

As explained in section 4, only the highest resolution is available in the previous speeds and directions dimensions, so as to prevent divergence.

Two time resolutions are available: annual and seasonal (periods defined as explained in section 2.3.3.4). For space, 4 resolutions are available: high (0.5^o resolution with smoothing), low (5^o resolution with no smoothing), "belts" and global.

The "belts" resolution is a space discretization pattern in which space is decomposed into 3 or 4 manually defined areas ("belts") of relatively uniform track behavior. Belts are defined by belt boundaries parametrized as third order polynomials.

Coupled with the default threshold parameters, these priority rules make the model use seasonal pdfs in about 80% of cases, and high spatial resolution in about 70% of cases.

2.3.4 Sampling from the transition transition probabilities algorithm

Sampling from a probability density function algorithm As recalled in appendix C, the "hit and miss" and inverse cumulative density function (CDF) are two methods for sampling from a pdfs. They are equivalent in the case of non-parametric pdfs, except in the execution time. While the CDF is the fastest method in the case of very skewed pdfs, the "hit and miss" method is often faster in many other cases.

Therefore the algorithm used in the simulation to draw random numbers from the non parametric transition pdfs is as follows:

1. try the "hit and miss" method for a given amount of time (for example, 1 second),
2. if no sample is drawn, use the inverse cumulative density function method.

Computation of the next track position Once the predictands $(\dot{s}_i, \dot{\theta}_i)$ are sampled, the displacement speed vector is computed from:

$$s_i = s_{i-1} + \dot{s}_i.dt \tag{2}$$

$$\theta_i = \theta_{i-1} + \dot{\theta}_i.dt \tag{3}$$

with $dt = 6h$. The next position (x_{i+1}, y_{i+1}), 6 hours later, is then computed using this set of formulas:

$$y_{i+1} = y_i + \frac{180}{\pi.R_t}.s_i.dt.sin\left(\theta_i.\frac{\pi}{180}\right) \tag{4}$$

$$x_{i+1} = x_i + \frac{180}{\pi.R_t.cos\left(\frac{y_i.\pi}{180}\right)}.s_i.dt.cos\left(\theta_i.\frac{\pi}{180}\right) \tag{5}$$

where where R_t is the earth radius and dt is the 6-hour interval.

2.4 Termination probabilities

In our case, a third set of pdfs is needed: the termination pdfs to stop the tracks, so that time is not wasted in prolongating the tracks where it is not likely to survive. Since an intensity model will rule whether a TC still exists or dies (by convention, a TC is ruled "dead" if its intensity fall below 17m/s), a termination pdf is not of greatest importance, as long as track are never stopped before they being rule dead by the intensity model. Therefore, a simple termination pdf based on hurricane activity was chosen: $P(stop \mid location) = 1$ if $P(location) < threshold$, and 0 otherwise.

Two other conditions might lead to track termination as well: if the maximum number of 6-hourly position (120) is reached, or if no transition pdf have been found even at the lowest resolution (which happens in 0.05% of cases).

3 Results: comparison of synthetic tracks to HURDAT tracks

The goal of the statistical track algorithm described is to produce large numbers of different tracks that statistically conform to HURDAT. In this section are presented some comparison between real and synthetic tracks to assess their similarity.

3.1 Synthetic tracks' general aspect compared to HURDAT

Figure 15 shows 60 random synthetic tracks compared with HURDAT tracks. The general aspect is quite well reproduced. Some loops appear in created tracks at about the same frequency as in HURDAT (8% of tracks). However synthetic tracks look slightly more chaotic. This might be explained by a insufficient number of priors considered in the new model, or a lack of statistical significance of transition pdfs (see section 2.3.1), or by imprecisions due to discretization of the predictands.

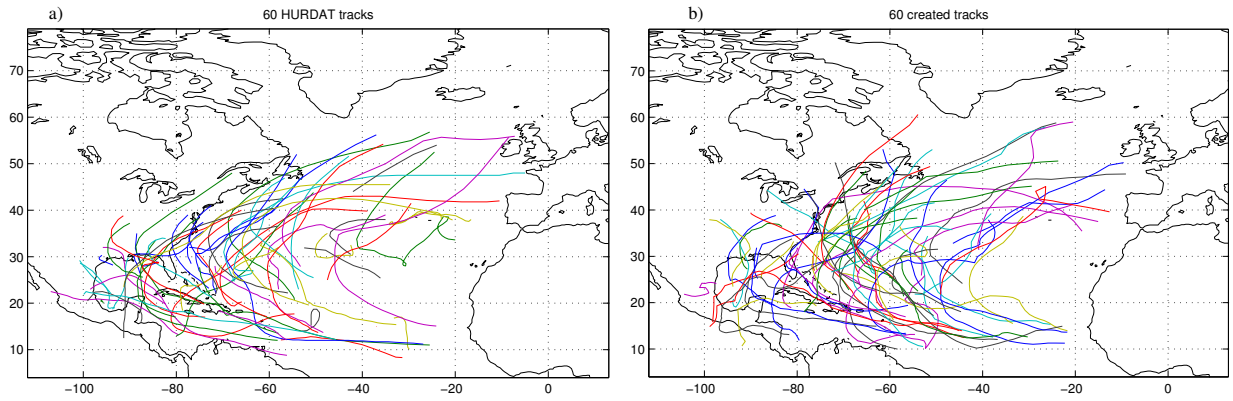

Figure 15: *general aspect of 60 randomly drawn synthetic tracks (a) compared to 60 random HURDAT tracks.*

3.2 Comparison of speed, direction, acceleration and direction rate of change global and spatial distributions

To assess the statistical similarity between HURDAT and synthetic tracks, histograms and expectation maps of motion variables (s, θ, $\dot{s}$, $\dot{\theta}$) are qualitatively compared. To more quantitatively compare the HURDAT and synthetic distribution, the Kullback-Leibler divergence was also calculated. As explained in appendix F, the Kullback-Leibler divergence is a

measure of the distance between two distribution ([Cover and Thomas, 1991]). These measure was chosen here because either the whole HURDAT dataset or the post-1970 HURDAT dataset may be considered as the true distribution we try to reproduce through a model.

3.2.1 Termination pdf used to avoid the termination bias when comparing synthetic tracks and HURDAT tracks

As explained in section 4, the fact that synthetic tracks are continued until $P(x,y) < threshold$ whereas HURDAT tracks stop over cold water, which occurs usually much earlier, adds a bias on some distributions, translation speed in particular.

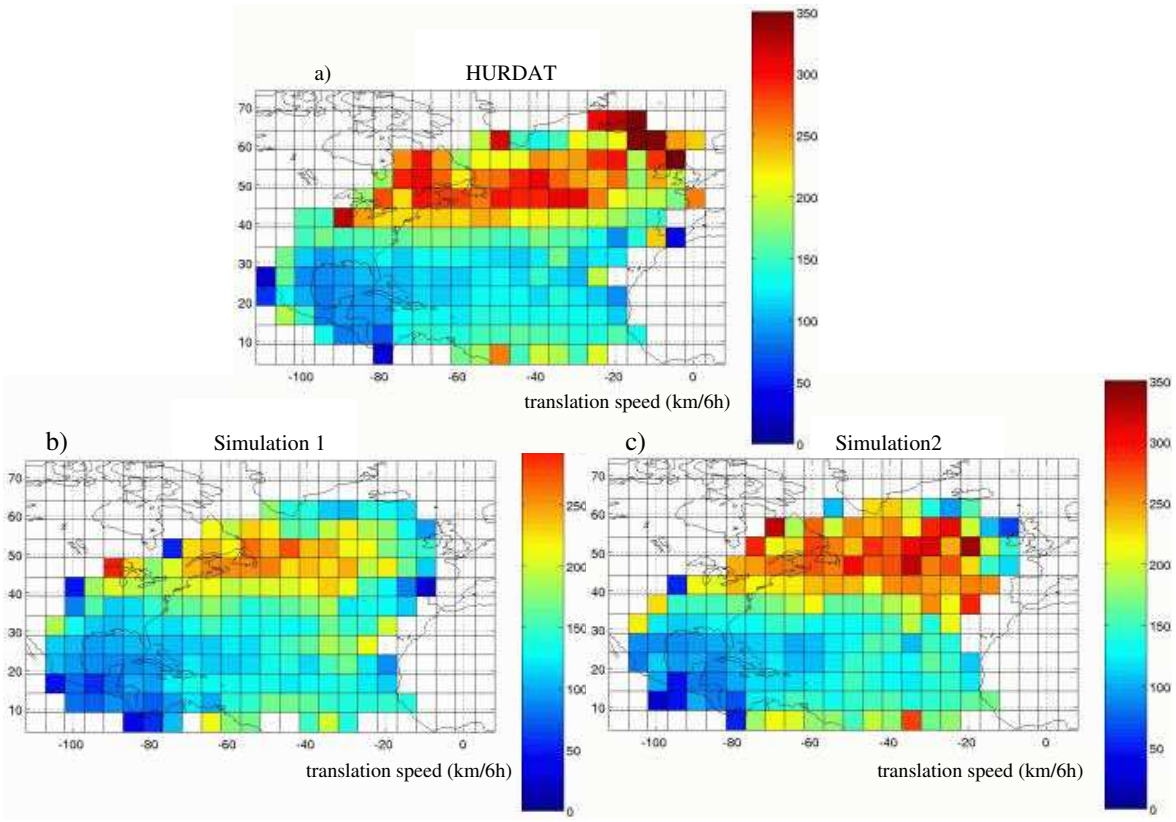

Figure 16: *Comparison of the spatial (b and c) speed distributions for 2 different runs of the model: in simulation 1, the regular termination criterion was used:* $P(stop \mid x,y) = 1$ if $P(x,y) < threshold$ and 0 otherwise. *In simulation 2, a more physical termination pdf is applied:* $P(stop \mid time\,spend\,over\,cold\,water)$
Both are compared to HURDAT (a)

So as to better compare HURDAT and synthetic tracks, a new termination pdf was here tested. Figures 16 and 17 compare 2 simulations with different termination probability:

1. In simulation 1 was use the regular termination criterion: $P(stop \mid x,y) = 1$ if $P(x,y) < threshold$ and 0 otherwise (figure 16b and 17d).

2. In simulation 2, a more physical termination pdf is applied: $P(stop \mid time\,spend\,over\,cold\,water)$, cold water being arbitrarily defined as higher latitudes than a threshold (20^o was here chosen) (figures 16c and 17b). Thus, simulation 2 tracks in average die much sooner.

Both simulations results are compared to HURDAT (figure 16a). As expected, there are globally to many high speed samples in simulation 1 (figure 17a), as tracks last much longer thus have the time to accelerate. Besides, speeds in high latitudes are too low compared to HURDAT (figure 16b), since slow tracks were continued in the simulation whereas they die in HURDAT.

On the contrary, the more physical termination pdf yields results very similar to HURDAT (figures 16c and 17b). This new termination criterion was used for all direct comparisons between HURDAT tracks and synthetic tracks.

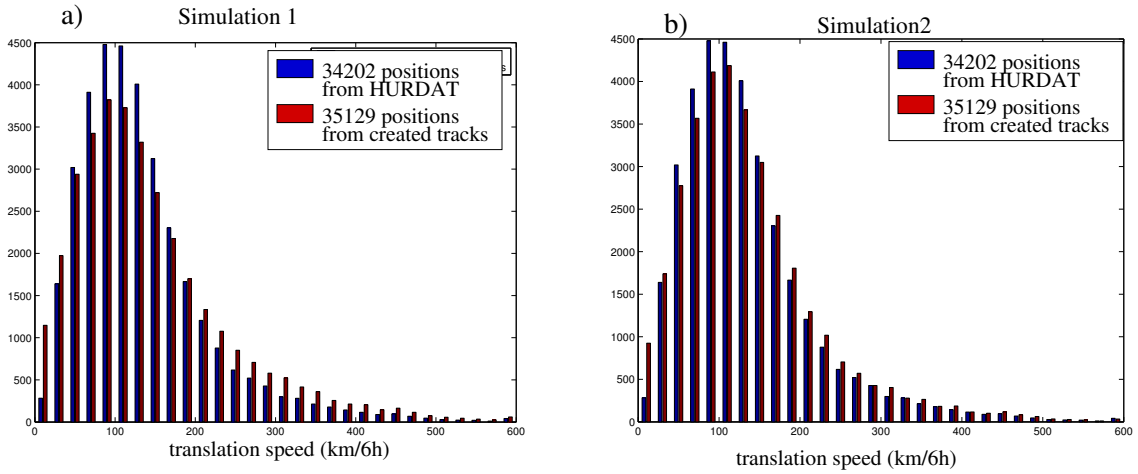

Figure 17: *Comparison of the global (d and e) speed distributions for the same 2 runs as in figure 16: n simulation 1, tracks were stopped until a statistical significance threshold is reached, whereas in simulation 2 tracks were stopped with a more physical criterion.*
Histograms compare HURDAT in blue and created tracks in red.

3.2.2 Comparison of synthetic tracks to all HURDAT tracks

The speed and direction global distribution (figures 19a and 19b: blue for HURDAT, red for synthetic tracks) are quite close to HURDAT's, as well as their spatial distributions (figure , 16c and 18b). However, one notice that synthetic tracks have too frequent very low speeds. The discontinuity at 180^o for direction is also poorly reproduced in the simulation, in which the distribution is much smoother.

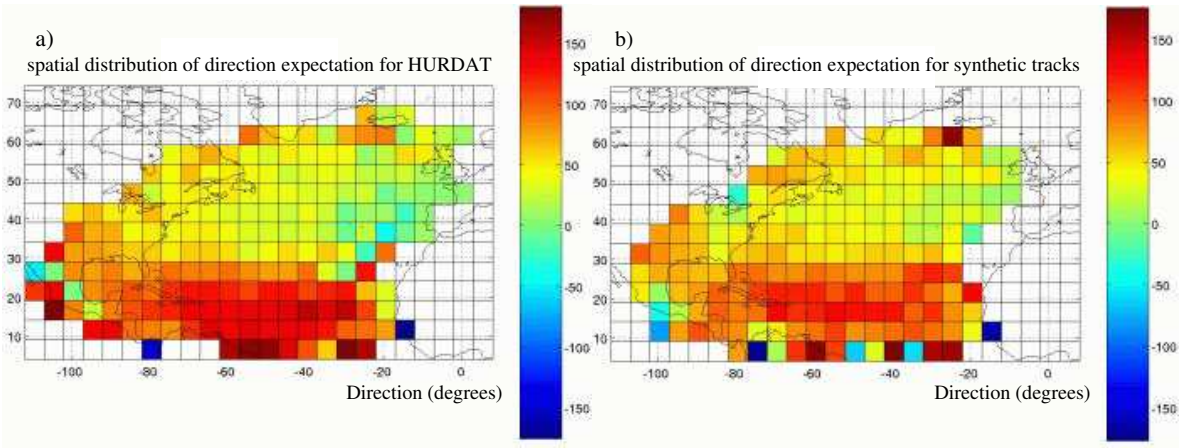

Figure 18: *Comparison of spatial direction distribution for HURDAT (a) and synthetic tracks (b).*

The global distribution of acceleration and direction rate of change shows a slightly broodier distribution than in HURDAT (figure 20: blue for HURDAT, red for synthetic tracks). This might at least partly explain why tracks look a bit more chaotic. An hypothesis to explain this is that the discretization in the predictands dimensions ($\dot{s}$ and $\dot{\theta}$) cast a bias on the distribution to sample.

3.2.3 A significant difference between pre- and post-1970 HURDAT tracks

Although the motion variables are quite well reproduced by the Markov chain model, there are some differences between HURDAT and synthetic distributions. However, one notice that there's a significant difference between the distribution of all HURDAT tracks and post-1970 tracks only: due to imprecisions in position records before this date, tracks

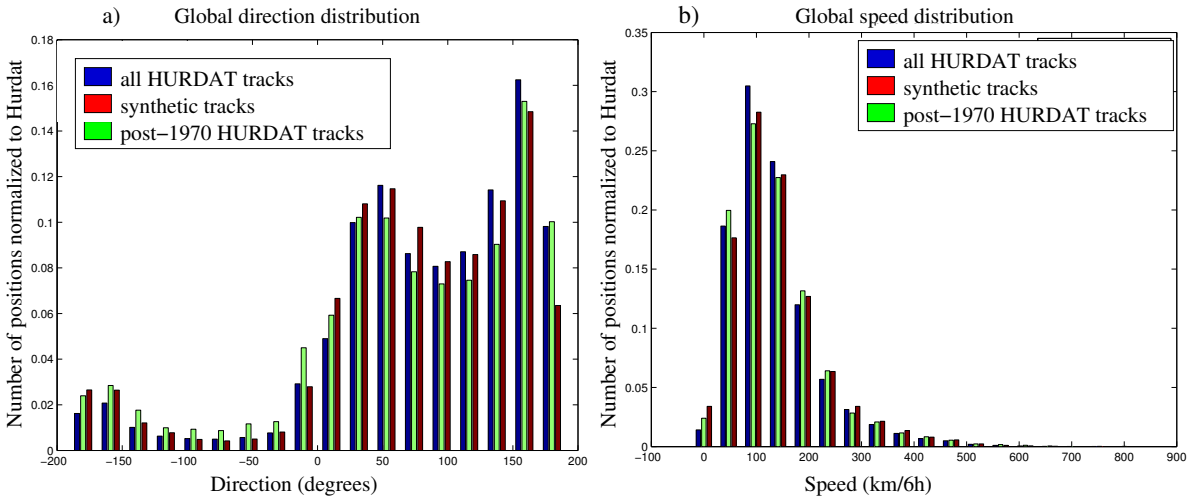

Figure 19: *The histograms compare all HURDAT tracks in blue , created tracks in red, and post-1970 HURDAT tracks in green. Synthetic tracks were stopped using P(stop|time spent over cold water).*
a) speed distribution
b) direction distribution

appear smoother and less erratic before this date (as shown by broader direction, acceleration and direction rate of change histograms: green on figures 18a and 20), and the discontinuity at 180^o is not as clear in post-1970 tracks only.

Table 2 compares the Kullback-Leibler divergence between synthetic tracks distributions and on the one hand, all HURDAT tracks, and on the other hand, post-1970 HURDAT tracks.

The synthetic distributions are expected to be closer to the all HURDAT tracks histogram than the post-1970 tracks only. However, it is the contrary for speed and accelerations: maybe the use of second order predictands were able to reproduce some intrinsic properties of TC tracks that were not fully captures in the pre-satellite tracks.

HURDAT set→ distribution compared↓	whole HURDAT	post-1970 only
translation speed	0.0090	0.0050
direction	0.0194	0.0323
acceleration	0.0057	0.0031
direction rate of change	0.0054	0.0287

Table 2: *Kullback-Leibler divergence values applied to the comparison of the speed, direction, acceleration and direction rate of change distributions obtained from the synthetic tracks, with the distributions obtained from the whole HURDAT dataset (first column) and for post-1970 tracks only (second column).*

3.2.4 A local example: tracks within 300km from Boston

Inn the wind risk estimation model perspective, tracks coming within a given radius of a given location are selected. It is therefore important to assess the spatial precision of the statistical model.

For illustration, the translation speed distribution within 300km of Boston obtained with synthetic tracks is compared to HURDAT in figure 21. The translation speed is a key variable for risk assessment in extratropical locations: exceptionally high translation speeds can make some hurricanes maintain high intensities and be very damaging even in extratropical regions in which the local potential intensity (defined in section 1) is very low. The 1938 New England “Long Island Express” is an example: although hitting a region of low potential intensity, this fastest moving hurricane ever recorded ranks among the most fatal hurricanes in the US history ([Valee and Dion, 1998]).

The translation speed distribution within 300km of Boston is extremely well reproduced by the statistical model.

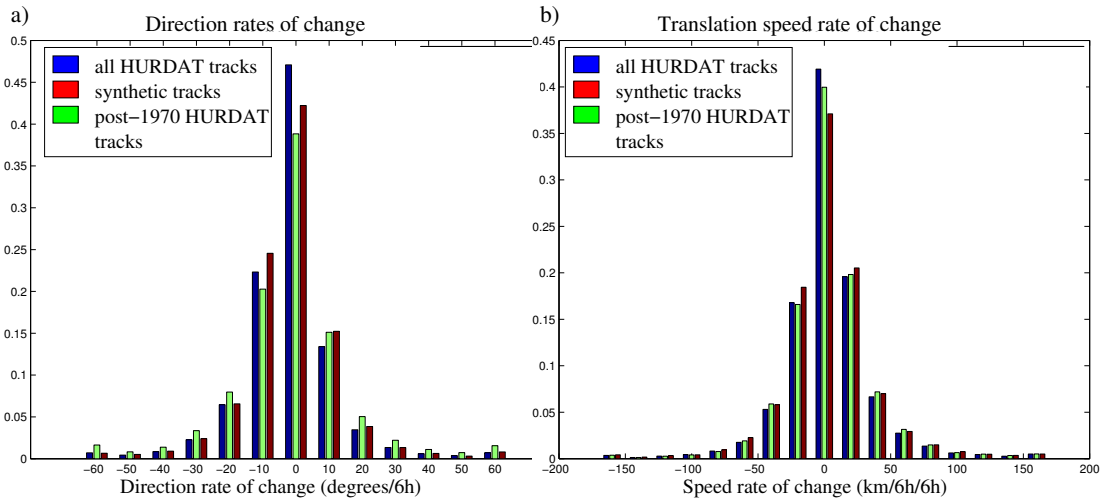

Figure 20: *acceleration (a) and direction rate of change (b) global distributions obtained with the Markov chain model. The histogram compare HURDAT in blue, created tracks in red, and post-1970 HURDAT tracks in green. Synthetic tracks were stopped using P(stop|time over cold water).*

3.3 Comparison of the mutual information matrices between synthetic tracks and HURDAT tracks

3.3.1 The mutual information matrix as a similarity measure between HURDAT and synthetic tracks

To deeper compare the relationships between the variables in synthetic and data tracks, the mutual information matrix was calculated. Appendix E recalls the definition, computation, interpretation and interest of mutual information. This matrix (figure 22) actually represents for each couple of variable (Y_i, Y_j), the quantity $\sqrt{\frac{I(Y_i,Y_j)}{min(H(Y_i),H(Y_j))}}$, where

I stands for mutual information and H for entropy. The mutual information was preferred to the correlation because it detects all kinds of relations. In our case, many relations are not linear (acceleration given latitude, for example).

Globally, the information matrix for synthetic tracks (figure 22) is relatively similar to HURDAT[5]. However, a more detailed analysis of this figure assesses the effect of the approximations and low resolutions in this model.

3.3.2 Interpretation

dependence in time The resolution used in the day of year dimension is quite low (figure 14), and seasonal pdfs are only used in about 80% of cases (section 2.3.3.5). Moreover, to run the program within memory limits, the seasonal pdfs have to be loaded separately, which is very time consuming. To try not to waste to much time, the time period of t_0 is used for the whole track. Since the average duration is 29 positions, this leads in average to an uncertainty up to 7 days.

However, as shown in figure 19, the time influence on direction and activity seems well reproduced. That means that the influence of time is quite low, and that diversity within a given time periods is large enough to make the low resolution used sufficient.

speed-direction interdependence The main assumption made was that speed and direction are independent. As explained in section 3.3.1, that's a very questionable approximation, since there's a strong relation between s and θ, as underlaid in figure 22a).

The comparison between the mutual information matrix shows that the model fails to completely reproduce this relation: there's actually a direct relation between speed and direction in real tracks.

Number of priors considered As explained in section 2.3.1, the selection of only 2 prior to account for persistence was questionable, since there's a striking continuity between positions, speed and direction, underlaid by figure 22a. Here we

[5]The information matrix for all HURDAT tracks and post-1970 tracks are extremely similar

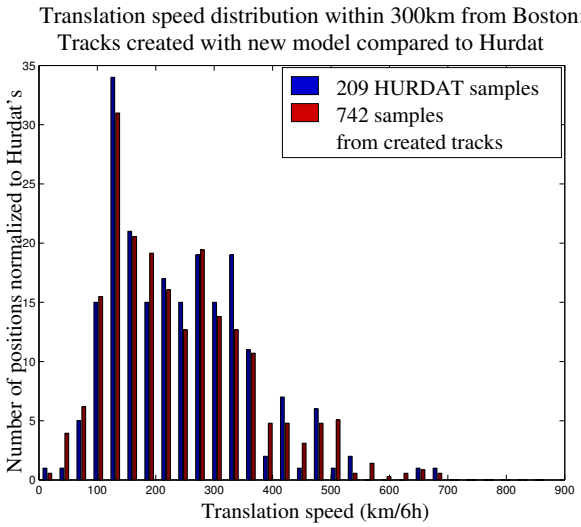

Figure 21: *Translation speed distribution within 300km of Boston for tracks created with the Markov Chain model (red), compared to HURDAT (blue). Synthetic tracks were stopped using P(stop|time over cold water).*

can see that persistence in speed, direction and location is surprisingly well reproduced. We can therefore consider that the reduction of persistence predictors to 2 priors only was far a viable approximation.

However, figure 22b shows that relations between $\dot{\theta}$ and θ, and between $\dot{\theta}$ and $\dot{\theta}_{-1}$, are underestimated by the model. The persistence in direction rate of change is problematic. The relations between $\dot{\theta}$,θ and $\dot{\theta_{-1}}$ is important in the recurvature and looping processes: when a track starts recurving, it keeps on recurving, or when a track starts looping, it finishes its loop. The lack of persistence in direction rate of change might partly explain that synthetic tracks are more chaotic than HURDAT's.

3.4 Number of sample to consider in a set

When creating a set of 100 tracks, the distribution of any synthetic variable may strongly differ from HURDAT due to sampling artifacts. In the example of the genesis month distribution, figure 23 shows the standard deviation of histograms bars for 100 sets of 100 tracks.

What is the minimum number of tracks to create in a set to insure that the set is representative, and its distributions are close to HURDAT? As shown in figure 23, the more samples in a set, the closer in average the synthetic distribution from HURDAT. Such tests were conducted in the case of genesis month, and it seems that at least 500 or 600 tracks per set are needed for the set to be satisfyingly representative.

When wind risk probabilities are computed, a set of 1,000 tracks was used to compute wind exceedance probabilities in the whole Atlantic basin, and sets of 15,000 tracks were generated to compute probabilities on specific coastal sites.

4 Discussion of the method and alternatives

4.1 What could have been improved

Lower resolution in longitude HURDAT's mutual information matrix (figure 22a) shows that speed and direction are much more linked to latitude than longitude. Therefore, the resolution in longitude could be reduced from 0.5^o to 5^o without much loss of information. The consequent gain of space of a factor of 10 could be used to increase the pdfs statistical significance and/or add new predictors. For example, one could take into account the dependency between direction and speed and the dependency of previous direction rate of change on current direction rate of change. Both of these relationships are lacking in the current algorithm.

Generalize variable bin size to all predictors The variable bin size used to define time periods (section 2.3.3.4) could be generalized to all predictors. For example, in the case of predictor s_{i-1}, the resolution is 40km/6h and speeds range in [0,800] km/6h. However, very few storm reach translation speeds greater than 400km/6h. Consequently, half of the space

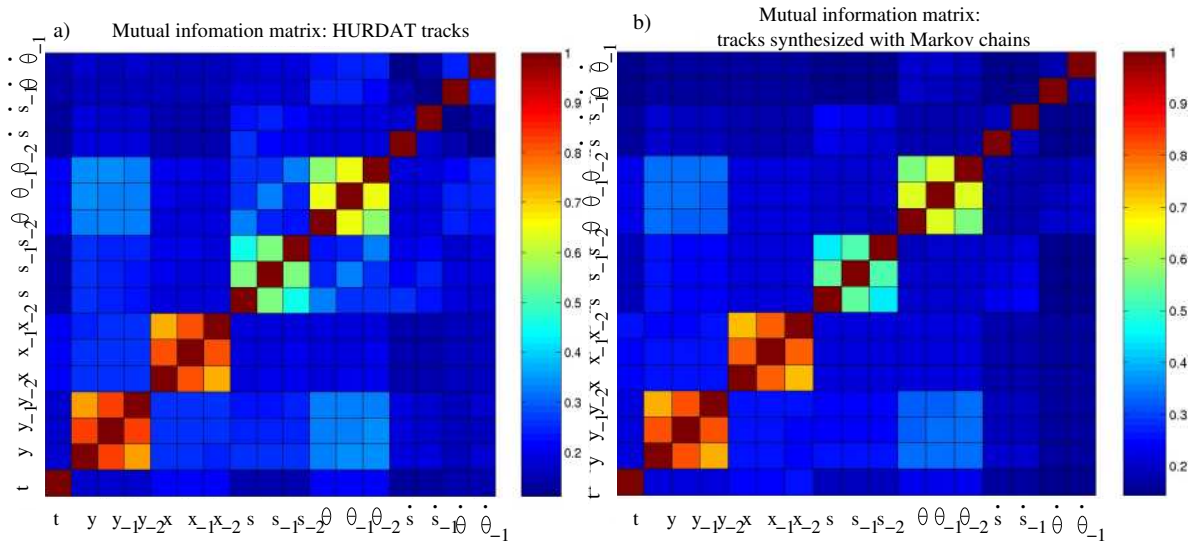

Figure 22: *Comparison of the matrix of the square root of the normalized mutual information between key TC variables, calculated from HURDAT (a) and for tracks simulated with the Markov chain algorithm (b).*
From left to right and bottom to top: time (t), longitudes and its 2 priors (x, x_{-1}, x_{-2})*, latitude and its 2 priors* (y, y_{-1}, y_{-2})*, translation speed and its 2 priors* (s, s_{-1}, s_{-2})*, direction and its 2 priors* $(\theta, \theta_{-1}, \theta_{-2})$ *, acceleration and its first prior* $(\dot{s}, \dot{s}_{-1})$ *and direction rate of change and its first prior* $(\dot{\theta}, \dot{\theta}_{-1})$*. A value of 0 means the two variables are independent. A value of 1 means there's an exact deterministic relation between the two variables.*

is wasted for very rare TCs. Resolution could be decreased in high speeds, and calculated using a similar algorithm as for time period definitions (section 2.3.3.4).

Interpolations Suppose we reduce the resolution in high speeds as just suggested, and that the last bin extends from 600km/6h to 800km/6h. If a synthetic storm has a current speed of 600km/6h, then the pdf used to draw its acceleration could have been made from a historical TC that accelerated from 600km/6h to 800km/6h. This could lead to an anomalously high speed of 1000km/6h.

To provide against this problem, a reduction in resolution should come together with interpolation techniques. For example, a relationship between previous speed and acceleration could be first calculated so as to adapt the conditional pdfs to the synthetic storm current speed.

Improved precision in predictand sampling The variable bin size should also be extended to the predictands dimension. When sampling from a pdf, the algorithm used is equivalent to considering the probability uniform within the bin. This is all the more inaccurate as bins are large, and as pdf values vary quickly with the predictand in question. This latter condition applies in most acceleration and direction rate of change pdf, which have a skewed Gaussian shape (figure 13). This inaccuracy in the predictand sampling is an hypothesis to explain why acceleration and direction rate of change are broader in synthetic tracks.

Therefore, the bins should be made variable, so as to make bins much smaller around the pdf means in cases of Gaussian-like pdfs. The sampling algorithm could also be improved by representing the discretized pdf not by rectangle histograms, but with trapezoids.

Discussion of the variable window method The variable window method is questionable, since the choice of the resolution to use during the simulation (see 2.3.3.5) is completely deterministic, and could therefore be made during the first phase of pdf estimation. Only the highest resolution would be kept, which could allow memory saving. Indeed, the pdfs are very heavy to store: 17GB par basin! However, the memory savings by suppressing the variable window method would be small, as the high resolution contains more than 95% of all pdfs weight. Moreover, the variable window method has the advantage to allow user to tune some resolution parameters through the priority rules (see section 2.3.3.5). Otherwise, one would have to re-estimate completely the pdfs to change a resolution, which might takes several days.

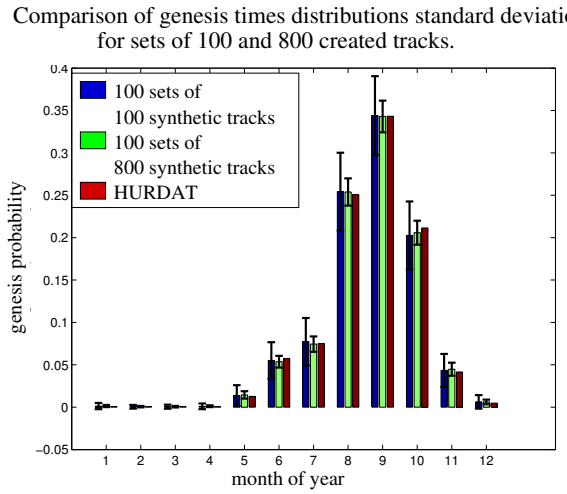

Figure 23: The red histogram represents HURDAT genesis dates distribution. In blue (resp green) is represented the mean and standard deviation of 100 histogram bars value produced for each of the 100 sets of 100 (resp 800) synthetic tracks

Utilization of low-level file-access functions to reduce simulation execution time During the simulation, more than 5 hours are lost in loading the large pdfs files, time period after time-period (since all pdfs would surpass the RAM capacity). The file-loading is accomplished using a high-level matlab function, which is very convenient to use, but is very slow. Some time might be saved by using lower-level file access functions, that enable to directly read a value in a binary file without the need of loading the entire files.

4.2 Advantages and drawbacks inherent to the methodology from the quality of synthetic tracks point of view

Although many improvement could have made the synthetic tracks more realistic and closer to HURDAT, there are some drawbacks that are inherent to the methodology used, in particular the decision to discretize variables and to estimate all pdfs in advance and store them.

4.2.1 Advantages and drawback of discretization

The discretization of variable offers a very convenient way to store the pdfs, in an explicit manner. Moreover, it avoids the errors related to regression equations or parametric representation of the relationships. However, discretization adds imprecisions from the uniformity assumption both in the predictor and predictand dimensions, as explained in section 4.1.

Besides, it also make the pdfs considerably heavy. One can wonder how come we end up with 17GB of pdfs, while the historical record has hardly 5MB...

4.2.2 Advantages and drawback of estimating all pdfs in advance

Estimating all pdfs in advance presents the advantage that the simulation should then be very fast. However, this is not the case because of the huge amount of pdf data to load.

Besides, the pdf estimation phase takes about one week. This is a drawback when one wants to change the resolutions, the smoothing parameters or the choice of predictors and predictands.

4.2.3 An alternative with neither discretization nor pdf pre-estimation

I suggested and implemented an alternate algorithm to the Markov chain approach, based on analog selection. Thereafter, the Markov chain algorithm and this alternate algorithm will be named algorithm 1 and 1' respectively.

Methodology The idea is inspired from the analog motion model HURRAN ([Hope and Neumann, 1970]). The algorithm used is schemed in figure 24. Track are initiated as in the previously described algorithm . Then, for each forward step, an historical analog case sampled from the HURDAT database, with a probability proportional to a score S measuring the similarity between HURDAT potential analogs and the track being synthesized.

Let $\overrightarrow{X^H} = \left(t^H, x^H, y^H, s^H_{-1}, \theta^H_{-1}, \dot{\theta}^H_{-1}\right)$ be the vector of predictors for any potential HURDAT analog H and $\overrightarrow{X^s} = \left(t^s, x^s, y^s, s^s_{-1}, \theta^s_{-1}, \dot{\theta}^s_{-1}\right)$ the situation of the track being synthesized. Let $\overrightarrow{r} = \left(dt, dx, dy, ds, d\theta, d\dot{\theta}\right)$ be a vector of resolution parameters. Then the score S is calculated as

$$S\left(\overrightarrow{X^H}, \overrightarrow{X^s}, \overrightarrow{r}\right) = e^{-\Sigma_{i=1}^{n}\left(\frac{X_i^s - X_i^H}{r_i}\right)^2}$$

with n the number of components in the predictors vectors.

Note that there is no more independence assumption between speed and direction, and that the previous direction rate of change, which was lacking in the previous algorithm, is now a predictor.

Let $\dot{\theta}_{analog}$ and $\dot{s}_{analog}$ be the direction rate of change and acceleration of the selected analog respectively respectively. Then the acceleration and direction rate of change for the track being synthesized, $\dot{\theta}_i$ and $\dot{s}_i$, are assumed to follow a normal distribution $\mathcal{N}\left(\dot{\theta}_{analog}, \sigma_{\dot{\theta}}\right)$ and $\mathcal{N}\left(\dot{s}_{analog}, \sigma_{\dot{s}}\right)$ respectively.

Tracks are terminated with the same termination pdfs as in algorithm 1.

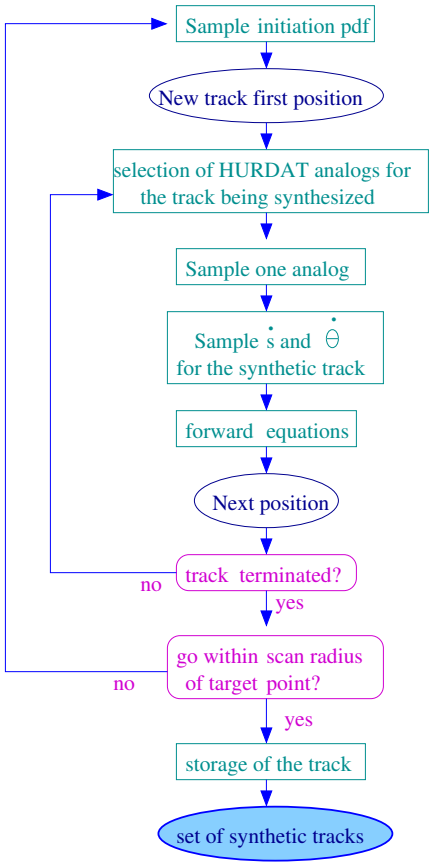

Figure 24: *Scheme summarizing the alternate algorithm 1' based on analog selection. Inputs/outputs are in blue, algorithms in green and conditions in purple. The final output (the set of synthetic tracks) is emphasized. The structure of the synthesis algorithm is very similar to algorithm 1, except that a transition algorithm based on analog selection is used instead of transition pdfs directly. Initiation and termination pdfs are the same as in the Markov chain algorithm.*

Similarity between this approach and the explicit Markov chain approach Although these two algorithms are very different in approach, they are very similar in practice.

- The analog selection through a Gaussian score is comparable to the Gaussian smoothing used in space in algorithm 1, except there is no discretization of predictors here.
- The pdf estimation is similar in the way they both use Gaussian kernel to smooth the historical data points, except no more discretization of predictands is carried out here.
- A big difference however is that this algorithm insures that both $\dot{s}$ and $\dot{\theta}$ are sampled from the same analog, and are consequently mutually consistent.

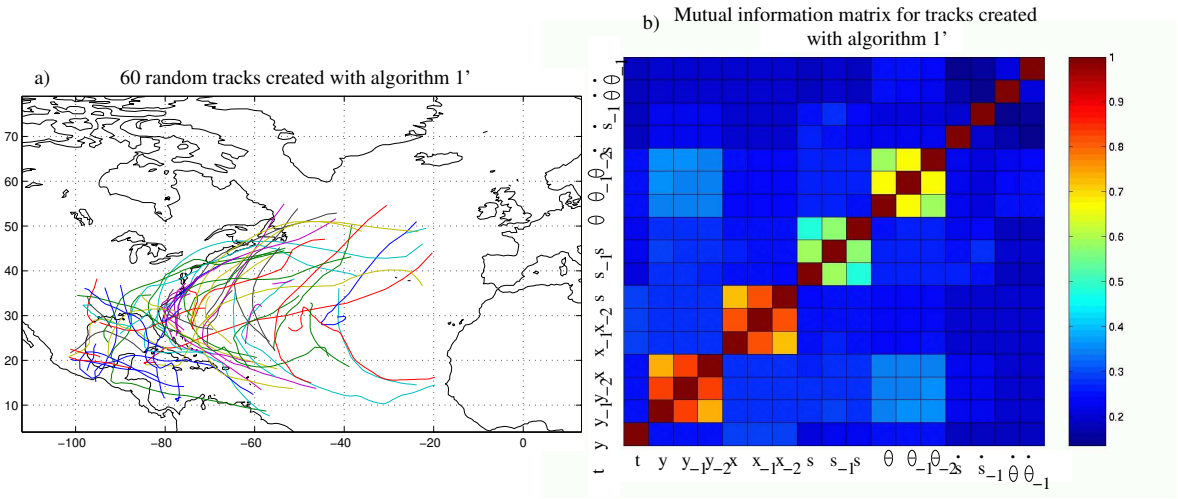

Figure 25: *Results obtained with the alternate analog algorithm 1':*
a) general aspect of 60 randomly drawn synthetic tracks. This can be compared with HURDAT (figure 8a) and algorithm 1 (figure 15).
b) Matrix of the square root of the normalized mutual information between key TC variables. This can be compared with HURDAT and algorithm 1 (figure 22).

results and discussion The aspect of the tracks is similar to HURDAT tracks (figure 25), although they also look slightly more chaotic. As in algorithm 1, this might be due to an insufficient statistical significance of the pdfs.

track synthesis algorithm→ distribution compared↓	algorithm1 (Markov chain)	algorithm 1' (analog selection)	algorithm2 (steering concept)
translation speed	0.0090	0.0052	0.0489
direction	0.0194	0.0139	0.0299
acceleration	0.0057	0.0007	0.1063
direction rate of change	0.0054	0.0018	0.0893

Table 3: *Values of the Kullback-Leibler divergence applied to compare synthetic distributions of speed, direction, acceleration and direction rate of change to these same distributions for all HURDAT tracks. Algorithm 1 is the algorithm based on an explicit Markov chain. The 2 last columns are alternatives for this model: algorithm 1'i s the algorithm based on analog selection. Algorithm 2 is the algorithm based on steering concept.*

The speed, direction, acceleration and direction rate of change are presented in figure 26 (red histograms). Table 3 compares the Kullback-Leibler divergence obtained for this algorithm (second column) compared to algorithm 1 (first column) (the third column is an another alternative described later). The distributions obtained with this alternative are more similar to HURDAT for all the four motion variables analyzed.

The biggest improvement is for the acceleration and direction rate (figure 26c and d, table 3) probably because no imprecisions were added through discretization.

This algorithm allows to have closer distributions to HURDAT tracks. However, the errors are slightly larger when compared to satellite-era tracks only (not shown).

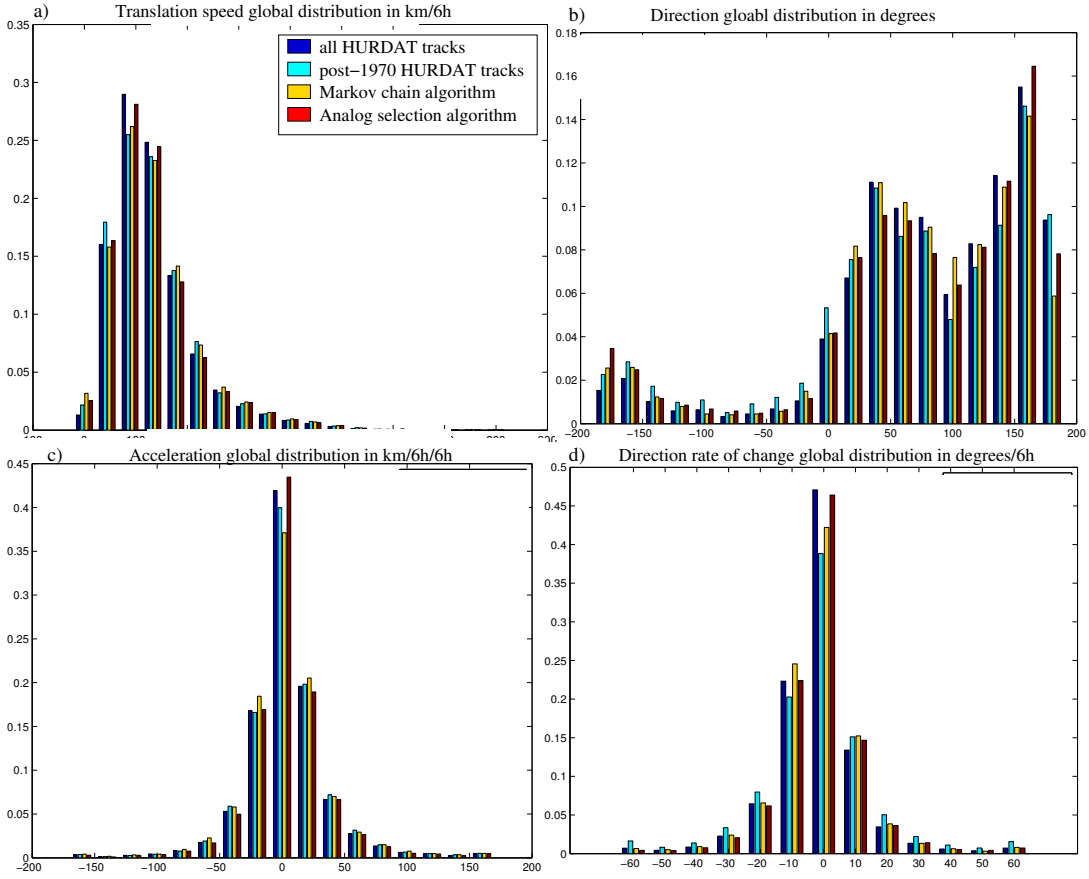

Figure 26: *Speed, direction, acceleration and direction rate of change global distribution obtained with the alternate analog algorithm 1' (in red) compared to all HURDAT tracks (in dark blue), post-1970 HURDAT tracks only (in cyan) and synthetic tracks from the explicit Markov chain model (in yellow). Synthetic tracks were stopped using P(stop|time over cold water).*

The mutual information matrix is improved (figure 25b): the relationship between s and θ is stronger, as well as the relationship between θ and θ_{-1}. However, the relationship between s and θ is more confused than in HURDAT.

Another big advantage of this simulation is that only initiation and termination pdfs are needed, and they are very small compared to transition pdfs. Moreover, the execution time is faster than the previous algorithm for less than 10,000 tracks. A drawback is that execution time is linear with the number of tracks to synthesize. It is therefore longer to run to synthesize very large numbers of tracks, in particular for specific sites.

4.2.4 Uncertainties on track location before the satellite era

As we showed, tracks after 1970, where satellite observation became routine, display significant differences from earlier tracks. Therefore, using the whole database might not be representative of the real tracks.

However, restricting the database to draw inference from would be inapplicable under this statistical methodology, since we already face important sparsity problems. HURDAT have only been recorded 340 tracks since 1970.

4.3 Drawback of this track simulation method in the perspective of intensity computation along the tracks

4.3.1 incompatibility between track and wind shear and its consequences

The goal was to create synthetic tracks in the perspective of applying a dynamical intensity model. Storm intensity is deeply influenced by wind shear, in particular the wind shear between 850 and 250hPa ([DeMaria and Kaplan, 1994]).Therefore,

in addition to tracks and monthly mean climatology data, the intensity model requires a time series of the wind at these two levels, hereafter noted $\overrightarrow{v_{250}}$ and $\overrightarrow{v_{850}}$. This time series is synthesized by an independent statistical model.

The problem is, this approach amounts to considering that wind shear and storm motion are independent, which is not valid. For example, the inaccuracy of this independent assumption are revealed by the study case of New Orleans: a significant proportion of storms hitting New Orleans originated in Western Gulf of Mexico and move North-Eastwards. But wind shear, which moderates intensity, is significantly stronger when tracks move North-eastwards. Therefore, because of an absence of direct connection between TC motion and wind shear in this methodology, the intensity model yields more frequent intense events on New Orleans than in historical data ([Emanuel et al., 2004]).

A solution could be to add constraints from track motion in the wind field statistical model. However, another alternative was favored: to synthesize directly tracks mutually consistent with wind shear.

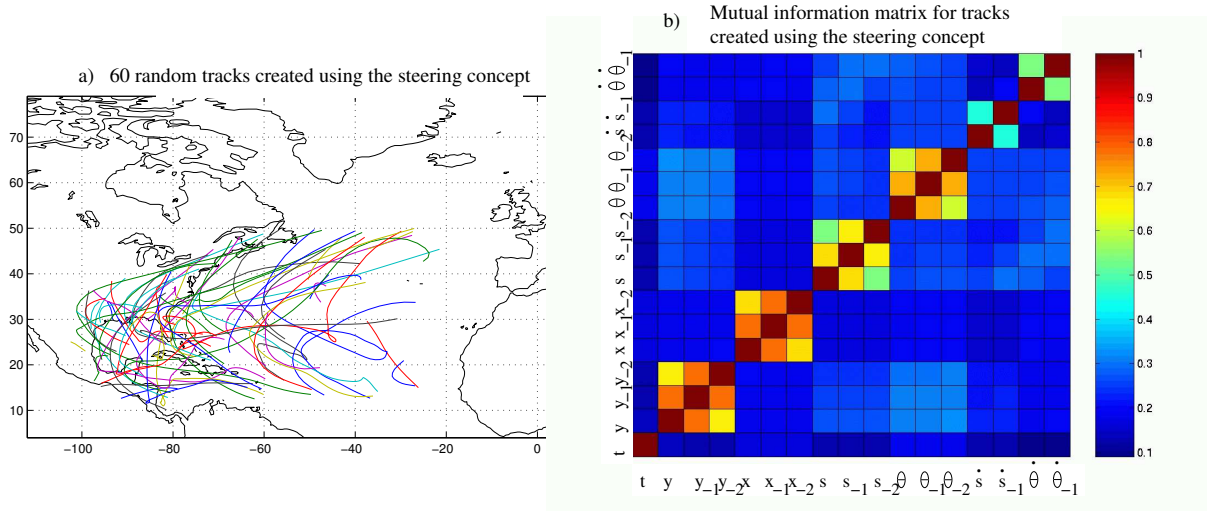

Figure 27: *Results obtained for tracks synthesized using the steering concept algorithm (algorithm 2):*
a) aspect of 60 random tracks. These can be compared with HURDAT tracks (figure 1) and tracks synthesized through algorithm 1 (figure 15) and algorithm 1' (figure 25a).
b) Matrix of the square root of the normalized mutual information between key TC variables. This can be compared with HURDAT (figure 22a) , algorithm 1 (figure 22b) and algorithm 1' (25b).
Both the P(x,y)<threshold termination condition and the $P(stop \mid time\,over\,cold\,water)$ pdf were applied to stop the tracks.

4.3.2 An alternative: a more dynamical track simulation model based on the steering concept

Methodology The methodology to create tracks and wind shear that are mutually consistent is to use a more dynamical model for TC motion, based on the steering concept: TCs are assumed to move with the altitude-averaged flow in which it is embedded, to which is added a vector pointing to the North-West, owing to the beta-effect $\overrightarrow{v_\beta}$ (see section 1.3.3). The altitude-averaged flow is here approximated as a linear relation between $\overrightarrow{v_{250}}$ and $\overrightarrow{v_{850}}$. Therefore, if $\overrightarrow{X}$ is the vector position of the storm, its motion is given by:

$$\frac{d\overrightarrow{X}}{dt} = \alpha\overrightarrow{v_{850}} + (1-\alpha)\overrightarrow{v_{250}} + \overrightarrow{v_\beta}$$

where $\alpha \in [0,1]$.

$\overrightarrow{v_{250}}$ and $\overrightarrow{v_{850}}$ are constructed by a statistical model, so that the mean, variance, covariance of $\overrightarrow{v_{250}}$ and $\overrightarrow{v_{850}}$ conforms to historical wind data, and so that there kinetic energy as a function of time follows a realistic spectral distribution.

The historical wind database used is the 4-times daily NCEP/NCAR reanalysis data set. My work in this part of the project was to built the monthly covariance maps of $\overrightarrow{v_{250}}$ and $\overrightarrow{v_{850}}$ meridional and zonal components. This consisted in importing the wind data netCDF files from 1948 to 2003, extracting them and compute the monthly covariances maps.

results and discussion Figure 27 displays 60 tracks created through the steering concept model. They look slightly smoother than historical tracks.

Tracks created using this algorithm also have slightly less accurate s, θ, $\dot{s}$ and $\dot{\theta}$ distributions (see values of the Kullback-Leibler divergence in table 3, third column). Note that in the steering concept algorithm, tracks are stopped when moving out of a rectangle domain extending from -110^o to -20^o in longitude and 0^o to 50^o in latitude. To better compare the results with HURDAT and the other synthesis algorithms, a combination of all used termination criteria for the different simulation methods were used: rectangle domain, $P(stop \mid time\,over\,cold\,water)$ and $P(x,y) < threshold$.

This method have the advantage not to do any assumption about which predictors to select, how many priors to consider or about the independence between s and θ . Therefore, as expected, the information matrix (figure 27) displays high mutual information values between most of the motion variables, in particular consecutive speeds, directions, directions rates of change, accelerations, and combinations of these variables. However, theses values are strikingly larger than in real tracks.

The main advantage of this algorithm is to create tracks and wind shear that are mutually consistent. Nevertheless, similarity between the resulting tracks and HURDAT, although not as good as the statistical models, are surprisingly good if considering that HURDAT tracks were never taken directly into account in the synthesis algorithm.

4.3.3 Comparison of tracks simulated through the purely statistical model and through steering concept in the perspective of the wind risk estimations

Both the purely statistical and the steering concept methods were used in the wind risk estimation model and compared.

They yields similar results in the study cases such as Miami, in which historical data is abundant. Besides, in this location, TC undergo no particularly strong interactions with wind shear and follow the steering concept well.

However, in many other cases, the two methods yield interestingly different wind risk estimates. For example, in a study case such as New Orleans, where a proportion of tracks are affected by strong wind shear, wind risks estimates from purely statistically simulated tracks are flawed by the incompatibility between tracks and wind shear.

On the other hand, in a study case such as Boston, tropical cyclone often undergo an extratropical transition, that is, they switch from a tropical regime to an extratropical one. In particular, the translation speed increases, which is responsible for maintaining high intensities over cold waters. But during extratropical transition, the steering assumption is no more sufficient to model the storm motion. Therefore, extreme hurricane events are underestimated when evaluating wind probabilities on Boston using tracks synthesized through the steering concept. On the contrary, the extratropical transition effects on motion are reflected by the historical database, and thus in the purely statistically simulated tracks.

Conclusion

In the perspective of calculating wind probabilities over given coastal locations, a statistical model was developed to simulate large number of hurricane tracks whose motion properties are conform to the historical data. Simulated as Markov chains whose initiation and transition probabilities are computed from the historical database, this track simulation technique yields motion properties extremely similar to the database. However, the Markov chain assumption together with various independence assumption, on the one hand between motion characteristics and on the other hand between motion and wind shear, is a drawback of this technique. Moreover, the robustness of the statistics is questionable in areas with few data.

These drawbacks are avoided in a more dynamical alternative to this statistical model, based on the steering concept. Tropical cyclone motion is slightly less similar to the historical data, which was never directly taken into account, but displays more consistency between different motion characteristics, as well as between motion and wind shear.

Both methods were used in the wind risk estimation model, compared and contrasted. They turn out to be complementary: depending on environmental influences undergone and historical data abundance, each method has its strength and weakness.

This coupled statistical-deterministic approach is new and contrasts with all previous wind probability estimation models. The synthesis of a large number of tracks provides a convenient way to interpolate track properties in location, and the deterministic intensity model allows to avoid the paucity of reliable historical intensity estimates.

Perspectives for this project are to further develop the track synthesis algorithm, in particular, accelerate the algorithm based on analog selection. Besides, it is also planned to integrate interdecadal variability, to which tropical cyclone activity is very sensitive ([Gray et al., 1992]).

An important question one can ask is: how can one evaluate this model? The model results were here compared to direct estimates from historical data, and the results were positive. However, a validation on independent data might be interesting, in particular for extreme events that are beyond the range of observation. For example, some historical reconstitution yields some information about landfalling storm back to the 1600s in the US. Some geological information could be also used to validate wind probabilities on a very long term ([Murnane, 2000]).

Acknowledgments

I would like to warmly thank my supervisor Kerry Emanuel for welcoming me for this internship. I enjoyed the opportunity he gave me to work with a lot of independence. I also appreciated his availability to provide explanations when needed. This internship really taught me a lot.

I'm also very grateful to Sai Ravela. He spent a lot of time to give me judicious advice and useful explanations. His help, both during my internship and for the redaction of this report, was very precious.

I thank Emanuel Vivant for sending me documentations about the former versions of the track simulation routines.

Appendix

A Markov Chains

A Markov process is a series of states $(X_i)_{i\in[1,n]}$, discrete or continuous, in which $\forall i \in [1,n]$, the transition from state X_i to $X_{i+1}$is defined by a transition probability conditioned only upon the immediately preceding state X_i (called the first prior), independently of the path by which the preceding state was reached and of time i ([Lange, 2003]). Let a, b, c, ... be possible values of the state X. Then $\forall i \in [1,n]$,

$$P(X_{i+1} = a \mid X_i = b, X_{i-1} = c, ...) = P(X_{i+1} = a \mid X_i = b) = P(a \mid b)$$

This is called the Markov assumption. This assumption allows to greatly simplify many complicated problems.
To build a Markov process, two sets of probability density functions (pdfs) are required ([Lange, 2003]):

1. Initiation probability $P(X_0)$, to initiate the process
2. transition probabilities $P(a \mid b)$ to elongate the process.

A Markov process in which states X are discrete is called a Markov chain. Transition probabilities in a Markov chain can be represented as a transition matrix T, so that $\forall (a_j, a_k) \in \Omega^2$, where Ω is the set of possible states, and $\forall i \in [1,n]$,

$$T(j,k) = P(X_{i+1} = a_j \mid X_i = a_k)$$

It's important to note that the definition of a Markov chain can be generalized to any series of states for which the transition probability depends on a finite number of preceding states (called *priors*). Suppose the transition probability is conditioned upon τ priors. Then the transition probability is $P(X_i \mid X_{i-1}, X_{i-2}, ...X_{i-\tau})$. A new state Y can be defined as $Y_i = (X_i, X_{i-1}, ...X_{i-\tau+1})$ and the transition probability can be expressed as $P(Y_i|Y_{i-1})$. The Markov chain assumption is thus satisfied.

B Parametric and non-parametric representation of probability density functions

Let P(X) be the distribution of a random variable X from which $(x_i)_{i\in[1,n]}$ samples are observed. P(X) can be estimated and stored in two ways:

1. Parametrically

P(X) is assumed to have a given shape, for example: uniform , Gaussian, Weibull... Function parameters can be estimated to fit the sample distribution the best. Such probability density functions (pdfs) are very small to store, since only the estimated function parameters are stored. However, precision is lost if the assumption of the chosen shape is not valid.

An advantage to parametrically store a pdf is that the distribution always reflect the smoothness and continuity of the underlying distribution, even if n is very small. However, small sample size might cast a bias on the function parameter estimation.

2. Non-parametrically

No assumption is made on the pdf shape. The pdf is stored as a histogram $(H_j)_{j\in[1,N]}$, with N the number of histogram bins. Errors are added from the discretization of the variable to predict, but the resulting pdf is in many cases more accurate, as long as the resolution is sufficiently high. The main drawback is that it uses much more storage (N histogram values, plus information about the bin centers $(b_j)_{j\in[1,N]}$).

The histogram might be calculated by just counting the number of sample that fall into each bin. However, this raw histogram might not reflect the smoothness of the underlying distribution, since histograms are very sensitive to the sample size, in particular if $n \ll N$.

To avoid this problem, histograms can be smoothed using a smoothing kernel ([Wand and Jones, 1995]): for each bin j,

$$P(X \in bin\ j) = \sum_{k=1}^{N} H_k . G(b_j - b_k)$$

The most common smoothing kernel is the Gaussian kernel of band-width σ:

$$G(b_j - b_k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(b_j - b_k)^2}{2\sigma}}$$

C Sampling from a probability density function

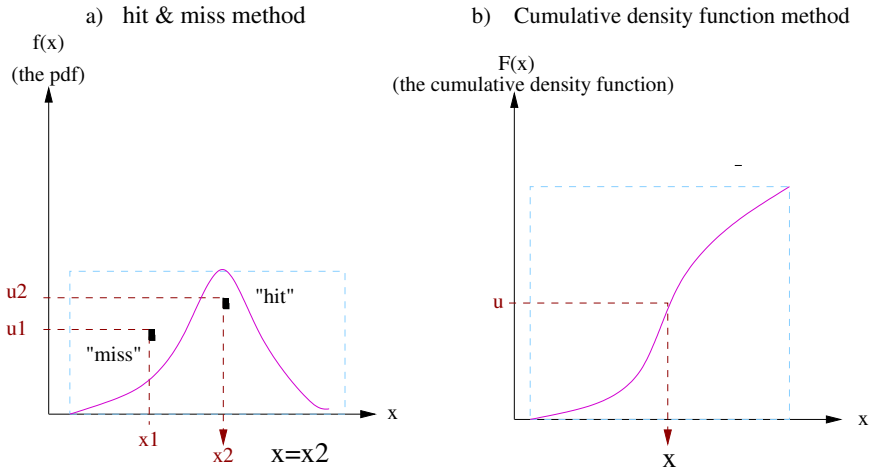

Figure 28: *Scheme summarizing two methods to draw random numbers from a pdf: a) the "hit and miss" method; b) the inverse cumulative density function method. Both are used in the simulation.*

To draw numbers from a given pdf, two methods among others can be used ([Gentle, 1943])(figure 28):

1. the inverse cumulative density function method makes use of the fact that if u is a random number following a uniform law on [0,1], then x that satisfied $u = \int_{-\infty}^{x} f(t)dt$ follows the law f(x). This method is fast for parametric method, when the integral of f(x) is known in advance.

2. For non-parametric methods, this method can still be used, but requires a numerical integration of f(x), which might be long for histogram pdfs with many bins. The acceptance rejection algorithm (or "hit & miss") often leads to faster results: u is drawn uniformly within the domain, and $x = u$ is accepted with probability f(x). However, this methods might be very slow or even lead to infinite loops in the case of very skewed pdfs.

D Smoothing

Let F be a field, and $H(\vec{p})$a function of this field, where $\vec{p} \in F$.
P is the result of a kernel smoothing of H if:

$$P(\vec{p}) = \sum_{\vec{q} \in F} H(\vec{q}) . G(\vec{p} - \vec{q})$$

where G is called the smoothing kernel. The most common smoothing kernel is the Gaussian kernel ([Ravela and Mammatha, 1999]):

$$G(\vec{p}) = \frac{1}{\left(\sqrt{2.\pi\,|det\Sigma|}\right)^3} e^{-\frac{1}{2}\vec{p}^{\,T}\Sigma^{-1}\vec{p}}$$

where Σ is the covariance matrix. For example, in a 3-D case and expressed on the principal component basis,

$$\Sigma = \begin{matrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_t^2 \end{matrix}$$

with σ_x, σ_y and σ_t the standard deviations of the Gaussian kernel in x,y,and t dimensions respectively. If these standard deviations are equal, then the smoothing is said isotropic. Standard deviation can be whether constant or functions of $\overrightarrow{p}$ as well.

E Mutual information and its measures

To estimate the relation between two random variables X and Y with observed values $(x_i)_{i\in[1,n]}$ and $(y_i)_{i\in[1,n]}$, the correlation coefficient r, defined as

$$r^2 = \frac{cov(X,Y)}{\sqrt{V(X).V(Y)}}$$

where *cov* stands for covariance and V for variance is often used.

However, to detect relations that are not linear, the mutual information ([Shannon, 1948]) is of great interest.

The entropy H of a random variable X, with a distribution represented as an histogram discretized as $(X_i)_{i\in[1,N]}$, where N is the number of bins in the histogram, is calculated as

$$H(X) = -\sum_{i=1}^{N} P(X = X_i).log\left(P\left(X = X_i\right)\right)$$

Entropy gives the uncertainty of a random variable X. For example, it is maximum for a uniform distribution and minimum (0) for a Dirac distribution.

The uncertainty of variable X given Y can be similarly measured:

$$H(X \mid Y) = -\sum_{j=1}^{M} P(Y = Y_j) \sum_{i=1}^{N} P(X = X_i \mid Y = Y_j).log(P(X = X_i \mid Y = Y_j))$$

where Y's distribution is discretized on a M-bin histogram.

The mutual information I, defined as

$$I(X,Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

is a measure of the relation existing between X and Y. $I(X,Y) = 0$ if there is no relation at all, and $I(X,Y) = H(X)$ if there is a deterministic relation between X and Y.

The mutual information can be normalized by its maximum possible value as

$$I_N(X,Y) = \frac{I(X,Y)}{min\left(H(X),H(Y)\right)}$$

if $min\left(H(X),H(Y)\right) \neq 0$ (or $I_N(X,Y) = 0$ otherwise), to obtain a coefficient I_N between 0 and 1 and comparable from random variables to others ([Li et al., 2003]).

To study the relation between a set of p random variables $(Y_i)_{i\in p}$, the mutual information matrix M can be computed so that $\forall (i,j) \in [1,p]^2$,

$$M(i,j) = I_N\left(Y_i, Y_j\right)$$

In practice, the mutual information tends to yield either values very close to 1, or values very close to 0. Therefore, the square root of the normalized mutual information was used in this report to better visualize the results.

F Measures of similarity between two distributions

Many measures exist to assess the similarity between two non-parametric distributions: Bhattacharyya distance, the χ^2 distance, the Kullback-Leibler divergence, root-mean-squares errors...

Let $(H_i)_{i\in[1,n]}$ and $(G_i)_{i\in[1,n]}$ be two pdfs stored as n-bins histograms.

The Kullback-Leibler divergence, also called relative entropy, is a measure of the inefficiency of assuming that the distribution is G while the true distribution is H ([Cover and Thomas, 1991]). Therefore, it is the most natural measure of similarity to use when one of the distribution is assumed true and the other is an approximation of it. The Kullback-Leibler divergence is expressed as:

$$D(H \mid G) = \sum_{i=1}^{n} H_i.log\left(\frac{H_i}{G_i}\right)$$

It is equal to 0 if the two distributions are equal, and is all the largest as distributions are different. It can be seen as a "distance" between two distributions, although it is not a true distance since it is not symmetric.

The Kullback-Leibler divergence was chosen in this report to assess the similarity between synthetic and historical tracks.

References

[hrd, 2004] (2004). National Oceanic and Atmospheric Administration's web page, Hurricane Division Center: http://www.aoml.noaa.gov.

[Batts et al., 1980] Batts, M., Cordes, L., Russell, L., Shaver, J., and Simiu, E. (1980). Hurricane wind speeds in the united states. Technical report, U.S. Dept. of Commerce - National Bureau of Standards, NBS Builing Science Series, 124.

[Chu and Wang, 1998] Chu, P.-S. and Wang, J. (1998). Modeling return periods of tropical cyclones intensities in the vicinity of hawaii. *J. Appl. Meteor.*, 37, 951-960.

[Cochran, 2000] Cochran, L. (2000). Wind engineering as related to tropical cyclones. C.p.p.'s technical publications, routledge press, Cermak Peterka Petersen (CPP), Inc.

[Cover and Thomas, 1991] Cover, T. and Thomas, J. (1991). *Elements of information theory*. John Wiley and Sons.

[Darling, 1991] Darling, R. W. R. (1991). Estimating probabilities of hurricane wind speeds using a large-scale empirical model. *Journal of Climate*, 4,1035-1046.

[DeMaria and Kaplan, 1994] DeMaria, M. and Kaplan, J. (1994). A statistical hurricane intensity prediction scheme (ships) for the atlantic basin. *Weather and Forecast*, 9, 209-220.

[Emanuel, 1995] Emanuel, K. (1995). Sensitivity of tropical cyclones to surface exchange coefficients and a revised steady-state model incorporating eye dynamics. *J. Atmos. Sci.,52, 3969-3976*.

[Emanuel, 1998] Emanuel, K. (1998). Lecture notes: Tropical cyclones. http://wind.mit.edu/ emanuel/geosys/geosys.html.

[Emanuel, 2003] Emanuel, K. (2003). Tropical cyclones. *Annu.Rev. Earth Planet. Sci. 2003 31:75-104*.

[Emanuel et al., 2003] Emanuel, K., DesAutel, C., Holloway, C., and Korty, R. (2003). Environmental control of tropical cyclone intensity. *Journal of the Atmospheric Sciences: Vol. 61, No. 7, pp. 843-858*.

[Emanuel et al., 2004] Emanuel, K., Ravela, S., Vivant, E., and Risi, C. (2004). A combined statistical-deterministic approach to hurricane risk assessment.

[Emanuel, 1988a] Emanuel, K. A. (1988a). The maximum intensity of hurricanes. *J. Atmos. Sci.*, 45, 1143-1155.

[Emanuel, 1988b] Emanuel, K. A. (1988b). Toward a more general theory of hurricanes. *Amer. Sci.*, 76, 370-379.

[Embrechts et al., 1999] Embrechts, P., Resnick, S., and Samadorodnitsky, G. (1999). Extreme value theory as a management tool. *North American Actuarial Journal*, 3, 30-41.

[Gentle, 1943] Gentle, J. (1943). *Random Numbers generation and Monte Carlo methods*. New York: Springer-Verlag, c2003.

[Georgiou, 1985] Georgiou, P. N. (1985). *Design wind speeds in tropical cyclone-prone regions.* PhD thesis, University of Western Ontario.

[Gray et al., 1992] Gray, W. M., Landsea, P., Mielke, J. P., and Berry, K. (1992). Predicting atlantic seasonal hurricane activity 6-11 months in advance. *Wea. Forcasting*, 7, 440-455.

[Hope and Neumann, 1970] Hope, J. R. and Neumann, C. J. (1970). An operational technique for relating the movement of existing tropical cyclones to past tracks. *Monthly weather review*, vol 98,12, 925-933.

[Jagger et al., 2001] Jagger, T., Elsner, J. B., and Niu, X. (2001). A dynamic probability model of hurricane winds in coastal counties of the united states. *J. Appl. Meteor.*, 40, 853-863.

[Jarvinen et al., 1984] Jarvinen, B. R., Neuman, C., and Davis, M. (1984). A tropical cyclone data tape for the north atlantic basin 1886-1983: Contents, limitations and uses. *NOAA Tech. Memo.*, NWSNHC-38.

[Lange, 2003] Lange, K. (2003). *Applied probability*. New York: Springer-Verlag, c2003.

[Li et al., 2003] Li, M., Chen, X., ma, B., and Vitanyi, P. (2003). The similarity metric. *Proceedings of the 14th ACM-SIAM Symposium, held in Phoenix, USA*, pp. 1889-1894.

[M. and Emanuel, 1998] M., B. and Emanuel, K. A. (1998). Dissipative heating and hurricane intensity. *Meteor. Atm. Physics 55, 233-240*.

[Murnane, 2000] Murnane, R. e. a. (2000). Hurricane landfall probabilities along the east and gulf coasts of the united states. *Eos, Trans. Amer. Geophys. Union*, 81, 433-438.

[Neumann, 1987] Neumann, C. (1987). The national hurricane center risk analysis program (hurisk) (reprinted with corrections 1991. *NOAA Tech. Memo.*, NWS NHC-38, 57pp.

[Neumann, 1972] Neumann, C. J. (1972). An alternate to the hurran (hurrican analog) tropical cyclone forecast system. *NOAA Tech. Memo.*, NWS SR-62, 22pp.

[Powell and coauthors, 2003] Powell, M. and coauthors (2003). State of florida hurricane loss projection model: atmospheric science component. Technical report, Florida Commission on hurricane loss projection methodology.

[Ravela and Mammatha, 1999] Ravela, S. and Mammatha, R. (1999). *Gaussian filtered representation of images*. Encyclopedia of Electrical and Electronical Engeneering.

[Rupp and Lander, 1996] Rupp, J. A. and Lander, M. A. (1996). A technique for estimating recurrence intervals of tropical cyclone-relted high winds in the tropics: Results for guam. *J. Appl. Meteor.*, 35, 627-637.

[Russel, 1971] Russel, L. J. (1971). Probability distributions for hurricane effects. *J. Waterw. Harbors Coastal Eng. Div. Amer. Soc. Civ. Eng.*, 97, 139-154.

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, vol. 27, pp. 379-423 and 623-656.

[Valee and Dion, 1998] Valee, D. R. and Dion, M. R. (1998). *Southern New England Tropical Storms and Hurricanes, A Ninety-Eight Year Summary (1909-1997)*. National Weather Service, Taunton, MA.

[Vickery et al., 2000] Vickery, P. J., Skerjl, P., Steckley, A. c., and Twinsdale, L. (2000). Simulation of hurricane risk in the united states using an empirical storm track modeling technique. *Journal of structural engineering*, 126, 1222-1237.

[Vivant, 2003] Vivant, E. (2003). Statistical hurricane tracks simulation methods as applied to the atlantic hurricane database. Master's thesis, MIT. Rapport de stage de fin de scolarite a l'Ecole Polytechnique.

[Wand and Jones, 1995] Wand, M. and Jones, M. (1995). *Kernel Smoothing*. Monographs on Statistics and Applied Probability.

[Wyatt, 1999] Wyatt, T. (1999). Wind effects. Dealing with natural disasters. Meeting between the Royal Society and the Royal Academy of Engineering.