Predictive Statistical Modeling of Atlantic Hurricanes

Thomas Cohn

I. Introduction

A hurricane is one of the most destructive weather events in the United States. Every fall, major storms threaten states along the Gulf Coast and Atlantic ocean with high winds, heavy rain, and storm surge. Some hurricanes proceed to make landfall, causing significant damage to life and property, while others avoid land altogether, having almost zero human impact. So understandably, the modeling and prediction of Atlantic hurricanes is an important topic of study. The goal of this paper is to present a statistical method of estimating the probability that a hurricane makes landfall based on its formation and early motion.

In this paper, I will begin by presenting the revised Atlantic hurricane database, called HURDAT2 (Landsea et al.), which will provide the data used for our algorithms. I will discuss some of the key variables within the data that my analysis will focus on, provide some high-level information about the data, and briefly look at how I process the data into my code. Next, I will present the prediction method. I will discuss the time series similarity metrics presented by Ho et al., and explain how I use them to perform dimension reduction on the set of hurricane tracks. I then explain how I compute a local linear kernel regression esimate of the projected data, with respect to whether or not a storm made landfall. I also mention the procedure for using the learned model to predict whether or not a hurricane will make landfall. At this point, I will also compare methods of generating new hurricane data, and perform some basic analysis to demonstrate that the generated data is realistic. I will then perform a full analysis demonstrating the strengths and weaknesses of my method. The paper will conclude with a brief discussion on the implications of my work, and directions of future research.

II. Data

After an Atlantic hurricane dissipates, the National Hurricane Center (NHC) determines the most accurate track of the storm, and collects various other information about the storm throughout its lifetime. Both modern and historical data have been combined to produce a database containing information on every single Atlantic hurricane since 1851, known as the revised Atlantic hurricane database, or HURDAT2 (Landsea et al.). The available data includes information about the hurricane taken every six hours, from the formation of the storm to its dissipation. This includes:

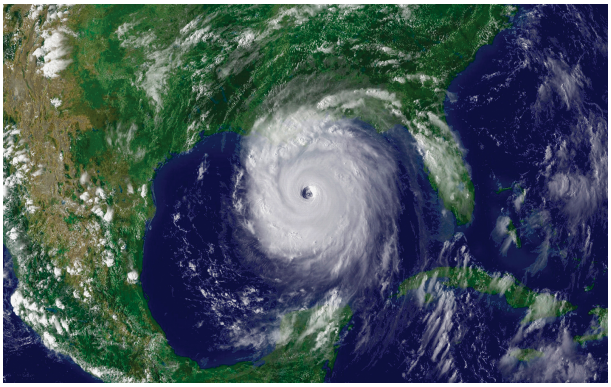


Fig. 1: A satellite image of Hurricane Katrina, taken in 2005. Over 1200 people were killed, and the storm did over $160 billion dollars of damage. *NOAA*

Fig. 2: A flooded neighborhood of New Orleans, in the aftermath of Hurricane Katrina. *Paul Morse*

- Latitude and longitude of the system center
- Landfall events
- Maximum sustained wind
- Minimum barometric pressure
- Maximum extent of hurricane and tropical storm winds

On occasion, there will be additional measurements besides the six hour updates, to mark important events in the storm's lifetime. For example, the data for Hurricane Irene (from 2011) contains an entry at 9:35 AM on August 28 to indicate a landfall that occurred at that time. Another example is Hurricane Gordon (from 2018); the dataset contains an entry at 9:00 AM on September 3 to provide additional detail on the intensity of the storm during a period of time where that intensity is rapidly

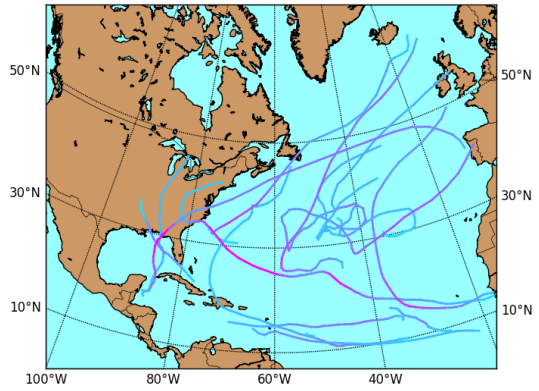


Fig. 3: A map of the track of every Atlantic tropical storm in 2018. The color represents the maximum wind speed of the hurricane at that point in time – magenta being the most intense.

changing.

The tracked latitude and longitude of the system center is one very important component of the dataset, as it allows us to examine the path a hurricane has taken throughout its lifetime. It also implicity gives us information on where the hurricane formed, and can be used to compare different storms to see how similar they are. For these reasons, the track of a hurricane is the most useful information to us, and will be the variable we focus on for prediction.

The maximum sustained wind and minimum batrometric pressure are two pieces of information that can be helpful for tracking how powerful a hurricane is. Both represent the peak strength of a hurricane at a given time, albeit in different ways. Maximum sustained wind speed and minimum barometric pressure are direct measurements of the power of a storm. In addition, the change in barometric pressure indicates whether a storm is strengthening or weakening. However, the focus of this paper is primarily predicting whether or not a storm will landfall, so the strength of historical storms is only tangentially relevant.

The extent of hurricane and tropical storm winds for a given hurricane gives us valuable information about how large a storm is. In addition, because those measurements are separated by the quadrant of the storm, they provide knowledge of the shape of the storm. Utilizing this information to improve my predictions could be successful, but it is beyond the scope of this paper.

Finally, tracking whether or not a hurricane made landfall, and the conditions when it made landfall, allows us to esimate how dangerous the hurricane was. A strong hurricane which never makes landfall is much less hazardous than a weak hurricane that does. I will specifically use this information as the dependent variable when training my statistical models.

A. High-Level Analysis

Figure 3 plots the track of every Atlantic hurricane in 2018, in order to demonstrate the most common paths for hurricanes to take. The majority of storms form off the west coast of Africa, initially move west across the Atlantic, then gradually turn north, and eventually back east. Depending on how this change occurs, and where exactly the storm formed, there are several common outcomes. Some storms turn to the north early and remain in the mid-Atlantic. Others turn later, moving up the Atlantic coast. The remaining hurricanes pass into the Gulf of Mexico before turning north towards the Gulf Coast and continental United States. There are also some storms which form within the Carribean or Gulf of Mexico – they tend to move north-northeast towards the United States.

B. Using the Data

As a part of constructing the HURDAT2 database, its creators took the time to examine historical records to ensure that the dataset is as complete as possible. There are no empty cells in the track data, which means all historical track data is usable without requiring any sort of interpolation. There are some empty cells in other variables, such as barometric pressure and wind speed extent, but this information is not relevant to this paper's analysis. Historical research is currently in progress to attempt to find some of these values, but ultimately, some measurements just don't exist.

The data from every storm, from 1851 to present, is stored in a single comma-separated values (CSV) file. The size of the file can be inconvenient, so I've written shell scripts to separate out individual years or specified ranges of years. In addition, each hurricane has a header row with different column specifications then the best track data lines that follow. I created a custom `Hurricane` object to hold all information about a given storm, and then wrote code to read in all the hurricanes from a given data file.

There are several other slight changes we make to the data. We reformat latitude and longitude as ordinary cartesian coordinates, and date and time information as a python `datetime` object. We also explicitly record several aspects of the storm that are only implicitly mentioned in the data, such as the maximum wind speed, minimum barometric pressure, and weather the storm made landfall.

III. Method

There are four parts to the methodology. First, I discuss the time series similarity metrics presented by Ho et al. in their paper *Manifold Learning for Multivariate Variable-Length Sequences With an Application to Similarity Search*. Then, I explain how the metrics are used to

perform dimensionality reduction on the hurricane tracks. Next, I present the local linear extension to Nadaraya-Watson kernel regression. Finally, I explain how to use the learned model to predict whether a hurricane will make landfall.

A. Time Series Similarity Metrics

In order to compare two hurricanes, we model their tracks as arbitrary-length multivariate spatiotemporal data sequences. Consider two hurricanes with track lengths r and s. Then their data sequences are

$$A = [(t_{A,1}, x_{A,1}, y_{A,1}), \ldots, (t_{A,r}, x_{A,r}, y_{A,r})]$$

and

$$B = [(t_{B,1}, x_{B,1}, y_{B,1}), \ldots, (t_{B,r}, x_{B,r}, y_{B,r})]$$

(respectively). The t_* are translated so that $t_{*,1} = 0$, and the x and y coordinates correspond to longitude and latitude (respectively).

I now present slightly modified versions of two similarity metrics used in the paper. There are parameters $\delta, \varepsilon > 0$, where δ controls the temporal similarity range, and ε controls the distance similarity range. Let $\mathscr{L}(A)$ give the length of A, and let $\mathscr{R}(A)$ return all but the last term of the time series A. Indexing by -1 is taken to mean accessing the last element in a list.

Define

$$M_1(A, B, \delta, \varepsilon) = \frac{\mathrm{LCSS}(A, B, \delta, \varepsilon)}{\min\{\mathscr{L}(A), \mathscr{L}(B)\}} \tag{1}$$

and

$$M_2(A, B, \delta, \varepsilon) = \frac{\mathrm{SLC}(A, B, \delta, \varepsilon)}{\min\{\mathscr{L}(A), \mathscr{L}(B)\}} \tag{2}$$

$\mathrm{LCSS}(A, B, \delta, \varepsilon)$ is the approximate longest common subsequence, and $\mathrm{SLC}(A, B, \delta, \varepsilon)$ is the soft longest common subsequence. I explicitly write out their equations in Figure 4, and implement them using bottom-up dynamic programming.

B. Dimensionality Reduction

Now that I have a metric with which I can compare two hurricanes, compute a modified dissimilarity matrix $\mathcal{S}$ for hurricanes $H_1, \ldots, H_n$

$$[\mathcal{S}]_{i,j} = -\frac{1}{2}(M_2(H_i, H_j, \delta, \varepsilon))^2 \quad i, j \in \{1, \ldots, n\} \tag{6}$$

M_2, the SLC metric, is used because it is more effective than the LCSS metric according to Ho et al. In Figure 5, I plot several sets of similar hurricanes to demonstrate the function of the metric. Because SLC is a symmetric function, $\mathcal{S}$ is symmetric, so I only compute half of it and then reflect across the diagonal.

I assume that there is some $k \in \mathbb{N}$ such that I can project our hurricanes into an $\mathbb{R}^k$ vector space, where the Euclidean distance between two hurricanes in $\mathbb{R}^k$ is approximately equal to their dissimilarity. Let P be this projection operator. To compute the P, I use kernel principal component analysis (KPCA) Schölkopf et al., with $\mathcal{S}$ as a precomputed kernel. This is implemented by Pedregosa et al. in the python package `scikit-learn`.

C. Kernel Regression

In order to infer a relationship between the probability a storm makes landfall and its projection, I use Nardaraya-Watson kernel regression, presented jointly in papers by Nadaraya and Watson. For L the indicator random variable of whether or not a hurricane H makes landfall, assume that $L = \mu(P(H)) + e$ where e is an error term and $E(e) = 0$. Let $H^{(i)}$ for $i = 1, \ldots, n$ be the Hurricanes used to train the model, with $L^{(i)}$ whether or not each of them made landfall. Then the Nardaraya-Watson estimator for kernel function K and bandwidth h is

$$\hat{\mu}(P(H)) = \frac{\sum_{i=1}^{n} K\left(\frac{P(H^{(i)}) - P(H)}{h}\right) L^{(i)}}{\sum_{i=1}^{n} K\left(\frac{P(H^{(i)}) - P(H)}{h}\right)} \tag{7}$$

Code to compute the estimator is implemented by Seabold and Perktold in the python package `statsmodels`. I use a local linear regression, to mitigate bias issues at the edge of the support.

D. Predicting Landfall

Now that I have described the full model, I'll walk through the process of actually making a prediction for a new hurricane. Let H be a recently-formed hurricane or tropical storm (for our purposes, I assume at most 3 days of track information). First, I compute the dissimilarity s_i with respect to each hurricane $H^{(i)}$, $i = 1, \ldots, n$ in the training data sets in the same manner as Eq (6). This gives us a vector S, which I can then project using our KPCA embedding to some vector $x = P(S) \in \mathbb{R}^k$. Finally, plugging x into the Nardaraya-Watson estimator gives us some $y \in [0, 1]$, which we take to be the probability that the hurricane will make landfall.

IV. Simulations

In order to generate more data from the limited real-world data available, I can add Gaussian noise to existing data. One way to do so is simply adding Gaussian noise to each point in a track. A more complex way is generating random walks of the same length, and then add them pointwise to the track. I will analyze how realistic the data generated by these methods are, and use the larger dataset to examine the quality of my estimators.

The first way to add noise to a track is to just independently add Gaussian noise to each point. Let $A = [(\vec{x}_1), \ldots, (\vec{x}_n)]$ (with $\vec{x}_i \in \mathbb{R}^2$) be the track of a

$$\mathrm{LCSS}(A,B,\delta,\varepsilon) = \begin{cases} 0 & \mathscr{L}(A) = 0 \text{ or } \mathscr{L}(B) = 0 \\ 1 + \mathrm{LCSS}(\mathscr{R}(A), \mathscr{R}(B)) & |t_{A,-1} - t_{B,-1}| < \delta \text{ and} \\ & ||(x_{A,-1}, y_{A,-1}) - (x_{B,-1}, y_{B,-1})||_2 < \varepsilon \\ \max\{\mathrm{LCSS}(\mathscr{R}(A), B, \delta, \varepsilon), & \\ \quad \mathrm{LCSS}(A, \mathscr{R}(B), \delta, \varepsilon)\} & \text{otherwise} \end{cases} \tag{3}$$

$$\mathrm{SLC}(A,B,\delta,\varepsilon) = \begin{cases} 0 & \mathscr{L}(A) = 0 \text{ or } \mathscr{L}(B) = 0 \\ C + \mathrm{SLC}(\mathscr{R}(A), \mathscr{R}(B)) & |t_{A,-1} - t_{B,-1}| < \delta \text{ and} \\ & ||(x_{A,-1}, y_{A,-1}) - (x_{B,-1}, y_{B,-1})||_2 < \varepsilon \\ \max\{\mathrm{SLC}(\mathscr{R}(A), B, \delta, \varepsilon), & \\ \quad \mathrm{SLC}(A, \mathscr{R}(B), \delta, \varepsilon)\} & \text{otherwise} \end{cases} \tag{4}$$

$$C = \min\left\{1, 1 - \frac{\varepsilon - ||(x_{A,-1}, y_{A,-1}) - (x_{B,-1}, y_{B,-1})||_2}{\varepsilon}\right\} \tag{5}$$

Fig. 4: Equations for LCSS and SLC, used in computing the similarity between spatiotemporal data sequences. Note the usage of Eq (5) in the computation of Eq (4).

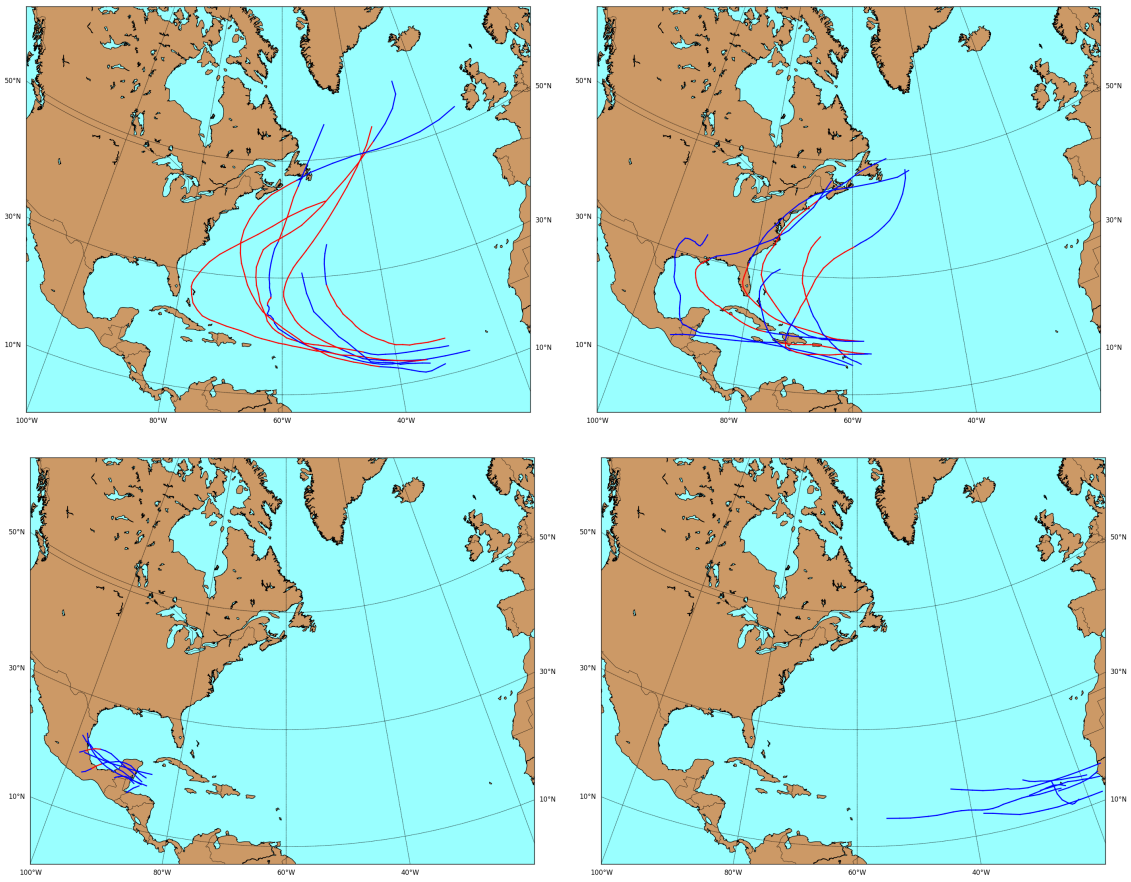
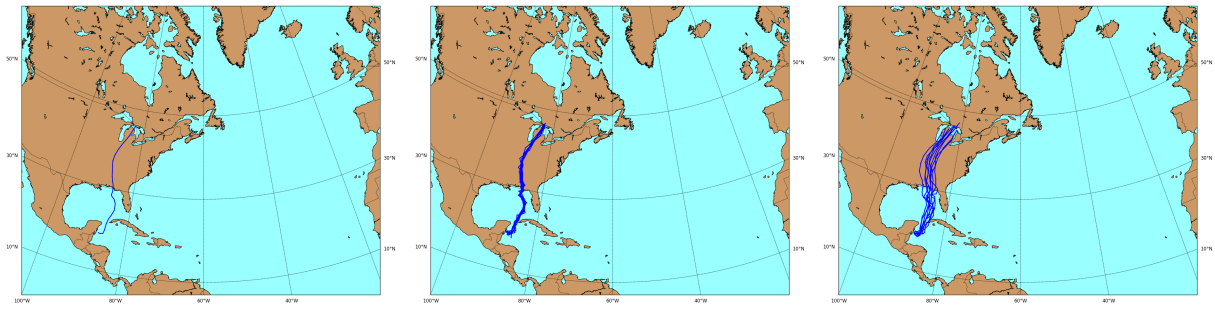


Fig. 5: The tracks from several sets of similar storms, according to the metric M_2. The red portion of the track is when the storm was hurricane strength, whereas the blue portion is when the winds were tropical storm force or lower. Qualitiatively, it's clear that the metric works as intended.



(a) Original Hurricane Track **(b)** Pointwise Noise **(c)** Random Walk

Fig. 6: A comparison of the impact of adding noise to copies of the track of a single hurricane with the pointwise noise method and the random walk method. The random walk clearly is more desirable, as it both appears smoother, and has greater variance.

	Number of Landfalls ℓ	Number of Hurricanes n	ℓ/n
Original Data	81	798	0.102
Gaussian Noise	80	798	0.100
Random Walk	85	798	0.107

TABLE I: Comparing the number of direct landfalls

hurricane. For $i = 1, \ldots, n$, let $\varepsilon_i \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^2$ and $\Sigma \in \mathbb{R}^{2\times 2}$ are some mean and covariance. Then our new hurricane track is

$$A' = [\vec{x}_1 + \varepsilon_1, \ldots, \vec{x}_n + \varepsilon_n]$$

A more advanaced way to add noise to a track is to compute a Gaussian random walk of the same length as the track, and then add them together. Let A be the track of the hurricane as before. For $i = 1, \ldots, n$, let $\varepsilon_i \sim \mathcal{N}(\vec{x}_{i-1} + \varepsilon_{i-1}, \Sigma)$ and $\varepsilon_0 \sim \mathcal{N}(\mu, \Sigma)$ for some initial mean μ and covariance Σ. Once again, our new hurricane track is

$$A' = [\vec{x}_1 + \varepsilon_1, \ldots, \vec{x}_n + \varepsilon_n]$$

Investigating the means and covariances that would be mimic real-world data is beyond the scope of this paper, but would represent an interesting direction of future work. For now, I keep $\mu = 0$, and let $\Sigma = 0.1I$, for the purposes of comparing the two methods of adding noise. Given every hurricane from 1970 through 2018, I make duplicate copies and add noise from both methods. In Table I, we measure the number of direct landfalls from each category, and compare the two noised totals to the original. (A direct landfall is where the epicenter of the storm directly passes over land – this allows us to use basemap software to directly determine whether or not a hurricane track with added noise has made landfall.) We find that both methods of adding noise to hurricane data preserve the average number of direct landfalls to a high degree of accuracy.

Although quantitatively, both simulation methods have similar results, Figure 6 provides the basis of a qualitative comparison of the two ways of adding noise. Specifically,

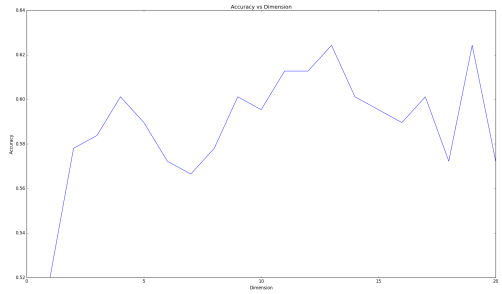


Fig. 7: A plot of accuracy of the estimator versus the number of dimensions the hurricanes are embedded into by KPCA.

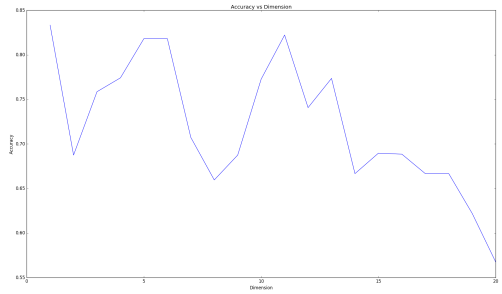


Fig. 8: A plot of accuracy of the estimator for only points in which it has high confidence (at least 90%) versus the number of dimensions the hurricanes are embedded into by KPCA.

it's clear that adding noise via random walk better accomplishes the goal of producing realistic hurricanes. The simulated tracks from the random walk method are smoother than those of the pointwise noise method, and they tend to diverge further from the seed hurricane track.

V. Analysis

I perform two major experiments to analyze the efficacy. First, I need to determine the underlying dimensionality of the vector space the hurricanes lie in.

To do this, I try embedding a subset of the full dataset into $\mathbb{R}^i$ for $i = 1, \ldots, 15$, and then regressing the Nardaraya-Watson estimator for each case. I then predict a different subset of the data to evaluate the accuracy of the model, for each number of dimensions. In Figure 7, I plot the accuracy by the number of dimensions i. In Figure 8, I plot the accuracy (but this time only including points for which the estimator has at least 90% confidence) by the number of dimensions. The overall accuracy is roughly improving up to $i = 15$, but not after. On the other hand, the accuracy for high confidence points appears to be optimal at $i = 10$. This makes it difficult to judge the underlying dimension of the data, so I will experiment with both choices of dimension.

First, I will test with $i = 10$. In order to fully evaluate how accurate my algorithm is, I train it on every huurricane from 2000 through 2018, and then test it on every hurricane from 1970 through 1999. Out of 479 hurricanes in the test dataset, the learned model correctly predicts 344; this is an accuracy of 0.72.

Now, I will test with $i = 15$. In order to fully evaluate how accurate my algorithm is, I train it on every huurricane from 2000 through 2018, and then test it on every hurricane from 1970 through 1999. Out of 479 hurricanes in the test dataset, the learned model correctly predicts 335; this is an accuracy of 0.70.

Additionally, in the case where $i = 15$, I plot all of the successfully predicted storms in Figure 9, and all of the not successfully predicted storms in Figure 10. There's a clear trend – storms further to the north are significantly easier to predict in the early stages of their lifetime just using statistics, whereas storms further to the south and in the Gulf of Mexico are more challenging.

VI. Discussion

In this paper, I've presented a new statistical method for predicting whether or not a hurricane will make landfall. I use time series similarity metrics to compare hurricanes, and use KPCA to find a representation for them in a Euclidean space. I then use Nardaraya-Watson kernel regression to approximate the relationship between these embedded coordinates and the probability a hurricane will make landfall. The analysis section provides evidence supporting the efficacy of my method.

Of course, this is a relatively primitive model, and there's a wealth of available information about a hurricane that I don't utilize. Factors such as the size of a storm, the extent in various directions of its winds, the barometric pressure, and much more could help tune an extension of the model to be much more accurate.

Another challenge my technique faces is speed. Computing the similarity between large numbers of hurricanes becomes expensive very quickly. The number of comparisons is $O(n^2)$ for the number of hurricanes in the training dataset, and each comparison is $O(\mathscr{L}(A)\mathscr{L}(B))$, even when utilizing dynamic programming. With more computational resources, it would be possible to train and test on larger datasets, and make full use of the simulation capabilities to generate new data.

Another ineresting area for future research would be high level modeling of hurricane behavior. This could be useful for further tweaking the simulation work, such as choosing a better mean and/or covariance.

Obviously, only 70% accuracy is not enough to make life-or-death predictions, and any purely statistical model will always be inferior to a more complex model that takes present weather patterns into account. However, this paper makes it clear that there are merits to using statistical techniques to model and understand hurricanes.

Works Cited

Ho, Shen-Shyang, et al. "Manifold learning for multivariate variable-length sequences with an application to similarity search". *IEEE transactions on neural networks and learning systems*, vol. 27, no. 6, 2015, pp. 1333–1344.

Landsea, Christopher, et al. "The revised Atlantic hurricane database (HURDAT2)". *NOAA/NHC.[Available online at nhc. noaa. gov]*, 2015.

Nadaraya, Elizbar A. "On estimating regression". *Theory of Probability & Its Applications*, vol. 9, no. 1, 1964, pp. 141–142.

Pedregosa, F., et al. "Scikit-learn: Machine Learning in Python". *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.

Schölkopf, Bernhard, et al. "Kernel principal component analysis". *International conference on artificial neural networks*, Springer, 1997, pp. 583–588.

Seabold, Skipper and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python". *9th Python in Science Conference*, 2010.

Watson, Geoffrey S. "Smooth regression analysis". *Sankhyā: The Indian Journal of Statistics, Series A*, 1964, pp. 359–372.

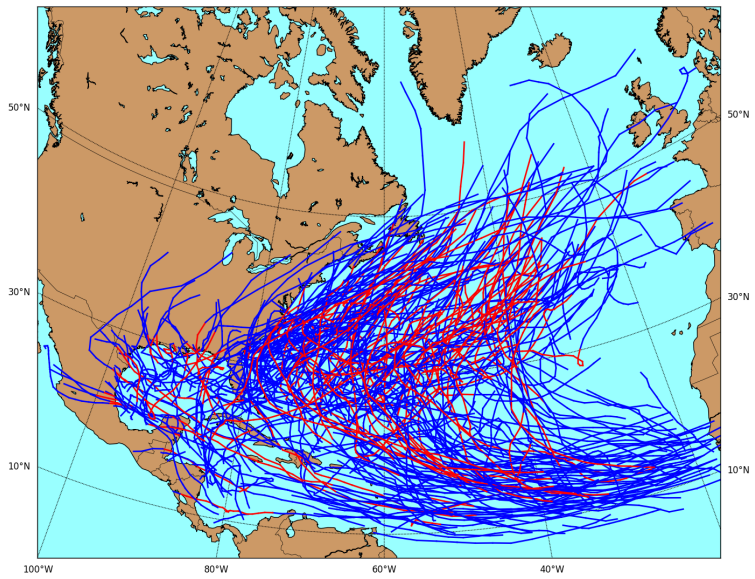


Fig. 9: A plot of all hurricanes successfully predicted by the dimension-15 model.

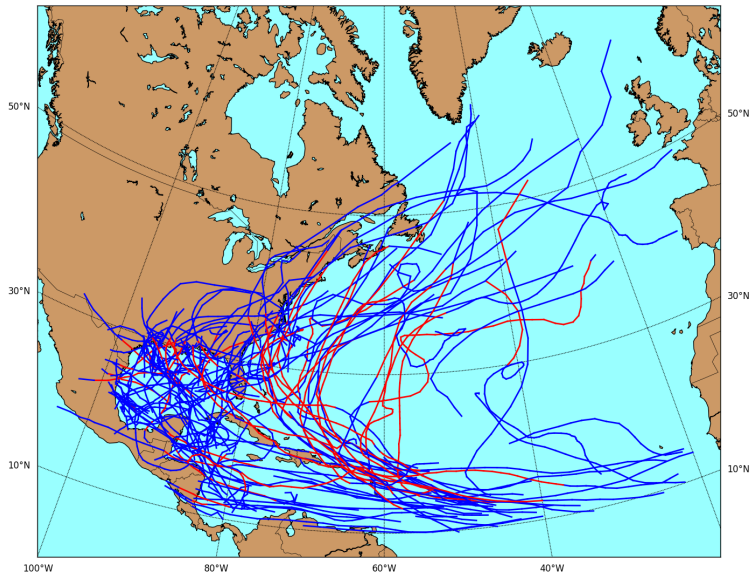


Fig. 10: A plot of all hurricanes not successfully predicted by the dimension-15 model.