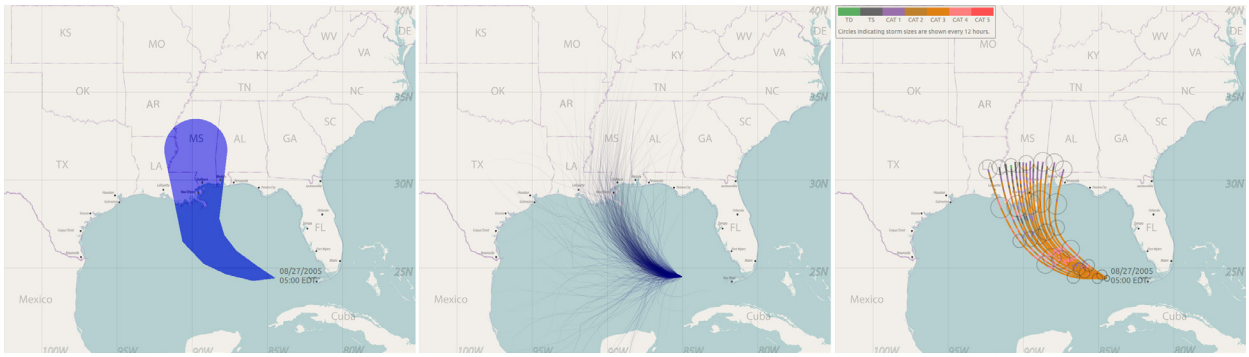
Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks

Le Liu, *Member, IEEE*, Lace Padilla, Sarah H. Creem-Regehr, and
Donald H. House, *Member, IEEE Computer Society*



(a) explicit: uncertainty cone (b) implicit: ensemble of tracks (c) selectively sampled annotated tracks

Fig. 1: Three approaches to visualizing a tropical cyclone prediction with uncertainty.

Abstract—A common approach to sampling the space of a prediction is the generation of an ensemble of potential outcomes, where the ensemble's distribution reveals the statistical structure of the prediction space. For example, the US National Hurricane Center generates multiple day predictions for a storm's path, size, and wind speed, and then uses a Monte Carlo approach to sample this prediction into a large ensemble of potential storm outcomes. Various forms of summary visualizations are generated from such an ensemble, often using spatial spread to indicate its statistical characteristics. However, studies have shown that changes in the size of such summary glyphs, representing changes in the uncertainty of the prediction, are frequently confounded with other attributes of the phenomenon, such as its size or strength. In addition, simulation ensembles typically encode multivariate information, which can be difficult or confusing to include in a summary display. This problem can be overcome by directly displaying the ensemble as a set of annotated trajectories, however this solution will not be effective if ensembles are densely overdrawn or structurally disorganized. We propose to overcome these difficulties by selectively sampling the original ensemble, constructing a smaller representative and spatially well organized ensemble. This can be drawn directly as a set of paths that implicitly reveals the underlying spatial uncertainty distribution of the prediction. Since this approach does not use a visual channel to encode uncertainty, additional information can more easily be encoded in the display without leading to visual confusion. To demonstrate our argument, we describe the development of a visualization for ensembles of tropical cyclone forecast tracks, explaining how their spatial and temporal predictions, as well as other crucial storm characteristics such as size and intensity, can be clearly revealed. We verify the effectiveness of this visualization approach through a cognitive study exploring how storm damage estimates are affected by the density of tracks drawn, and by the presence or absence of annotating information on storm size and intensity.

Index Terms—uncertainty visualization, hurricane forecasts, ensemble visualization, ensemble sampling, implicit uncertainty

1 Introduction

Much of the work on uncertainty visualization has focused on producing representations of the uncertainty in a prediction or estimation using one of the visual or spatial channels in the display such as size, opacity, color, or glyph style [23, 24]. While this approach has the appeal of explicitly presenting uncertainty along with data, it reduces the number of information channels available for conveying multivariate information, and often results in a display that visually confounds the uncertainty information with the prediction or estimation itself. A case in point is in tropical cyclone prediction, where the well known uncertainty cone, illustrated in Figure 1a, has been experimentally shown to mislead viewers into the misunderstanding that the cone's increase in width over time indicates that the storm is predicted to increase in size or intensity [14,27]. Indeed, in recent years the National Hurricane Center has added a text warning to the cone: *Note: The cone contains the probable path of the storm center but does not show the size of the storm* [20]. In fact, the unadorned cone visualization conveys no information concerning either the potential size of the storm or its predicted intensity.

Curve Boxplots [17] provide a promising alternative form of summary visualization. They have a number of advantages over the uncertainty cone, including the ability to portray levels of uncertainty and to portray outliers, but they have the same inherent problem: as the uncertainty increases, the spatial area subsumed by the visualiza-

- *L. Liu is with Magic Weaver Inc., 2755 Great America Way Ste 135, Santa Clara, CA 95054, E-mail: leliu@magicweaver.com.*
- *L. Padilla is with the Department of Psychology, Northwestern University, 2029 Sheridan Road, Evanston, IL 60208, E-mail: lace.m.k.padilla@gmail.com.*
- *S.H. Creem-Regehr is with the Department of Psychology, University of Utah, 380 S 1530 E Beh S 502, Salt Lake City, UT 84112, E-mail: sarah.creem@psych.utah.edu.*
- *D.H. House is with the School of Computing, Clemson University, 100 McAdams Hall, Clemson, SC 29672, and the Bahá'í Institute of Higher Education, Tehran, Iran, E-mail: dhouse@clemson.edu.*

Manuscript received 31 Mar. 2018; accepted 1 Aug. 2018.
Date of publication 16 Aug. 2018; date of current version 21 Oct. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2865193

tion increases. While there have been no cognitive studies on Curve Boxplots to test this hypothesis, our experience testing a variety of summary displays [27] would indicate that this increase in spatial area will induce a cognitive bias towards seeing the storm as increasing in intensity over the time frame of the prediction.

The one approach that our experiments have shown to overcome this cognitive bias is to portray the prediction as a representative set of potential storm paths. Unlike summary displays that provide an explicit representation of uncertainty, this approach provides an implicit representation [5]. Figure 1b, from Cox et al. [6], demonstrates such a display, that allows the viewer to infer the uncertainty from its spatial layout. They showed experimentally that this visualization has small but measurable advantages over the uncertainty cone in conveying the spatial spread of storm path possibilities. More importantly, recent studies have shown that it does not appear to convey the impression that the storm is predicted to grow bigger and stronger over time [22, 27]. Additional studies in different domains have shown the superiority of discrete implicit representations in place of continuous curves in conveying an accurate representation of the distribution of random variables [10–12]. While implicit uncertainty representations appear to have advantages over explicit representations, when dealing with large or complex data sets the purely implicit approach can suffer from severe overdrawing or clutter, making the visualization difficult to interpret. This drawback is apparent in Figure 1b.

Our work on tropical storm predictions has been guided over the past five years by regular consultation with key personnel at the National Hurricane Center. Of greatest relevance to the current work, were our personal communications with the Federal Emergency Management Agency (FEMA) representative at the NHC. He made it clear to us that supporting FEMA's emergency decision-making responsibilities, when a tropical cyclone is approaching, requires having clear information on *where* the storm is likely to strike, *when* it is likely to strike, and what its *intensity* is likely to be when it strikes. In addition, forecasters at the NHC wanted to see a representation of storm size along with any representation of the storm's spatial position. No single visualization currently published by the NHC provides all of this information, nor does the implicit visualization of Figure 1b.

In this paper we attempt to convey the predicted where-when-intensity-size information in an implicit uncertainty visualization. To do this, we are proposing a sampling approach that reconstructs a small representative and coherent set of paths from a larger ensemble of incoherent paths. We demonstrate this method using US National Hurricane Center tropical cyclone prediction ensembles, showing how it can be used to produce visualizations like that shown in Figure 1c. Because we display only a small number of well laid out tracks, it is possible to annotate these tracks with glyphs providing time and storm size information, and with a color coding for predicted storm intensity. These can be seen in Figure 1c, and in more detail in Figure 11. Interestingly, this visualization is structurally identical to one suggested independently in a proposal by designers in a recent *Scientific American Visual* blog article [4], but we have provided an algorithmic foundation to support its construction, and annotation with additional storm information.

In previous work we have explored ensemble sampling approaches for specific points in time but not for entire paths [14]. While this approach could be used as the basis for an interactive application for viewing a storm prediction, there remained the need for a visualization providing an overview of the prediction over its full time range (three to five days in the case of tropical storm advisories). This paper extends our earlier approach to consider complete paths but keeping the same underlying notion of what we mean by an easily visualizable representative sample from an ensemble:

- its members accurately preserve the original spatial distribution of ensemble members, and
- its members maintain a coherent spatial layout, so that overdrawing and clutter are minimized.

Primarily, the technique reported here for achieving these goals begins with extracting a median track of an ensemble by calculating the median track position at a sequence of time points, partitioning the ensemble into two equivalent groups, then recursively adding to the representative set by computing additional median tracks from successive groups. The resulting tracks are then drawn over a map of the region expected to be affected by the storm. Secondarily, our approach seeks to preserve auxiliary information contained within the ensemble. Each path generated by the National Hurricane Center algorithm carries time sampled information about storm size and expected maximum wind speed. We handle this auxiliary information by interpolating continuous storm strength and storm size fields from the ensemble, and then resampling from these continuous fields onto the set of representative paths. This information is used to annotate the tracks as they are drawn.

The following sections of this paper explain the algorithmic techniques underlying our path sampling approach, describe our approach to the design of path annotations, and show results for a number of different US Gulf Coast storm predictions. The paper concludes with a description of and results from a cognitive study that we conducted to explore this new visualization.

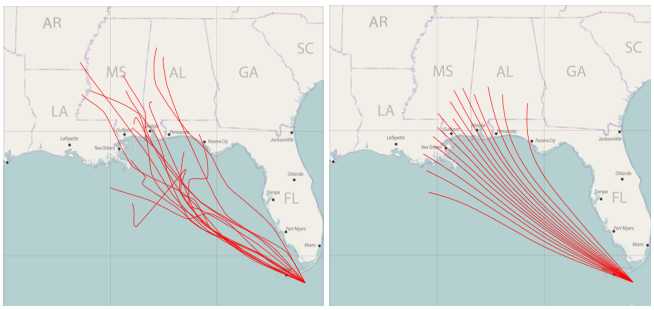
2 METHODOLOGY

Our path sampling algorithm takes as input a prediction ensemble of 1,000 storm tracks, as produced by an algorithm used by the US National Hurricane Center (NHC) [7, 8]. This ensemble is not simply a collection of outputs of storm predictions models, like those often presented in weather broadcasts as "spaghetti plots". Instead, the outputs of a number of different weather models are analyzed to produce an official storm path prediction. Taken together with a model of its uncertainty, this prediction is sampled via a Monte Carlo approach to produce a set of potential storm paths, sampled over 120 hours at one hour increments. Each sample point on a path indicates the latitude and longitude of the storm center, along with a number of predicted storm variables, which can be summarized as indicating the spatial extent of the storm, and its maximum wind intensity.

Figure 2a shows a random sampling of 15 paths from the National Hurricane Center ensemble for the 11 AM EDT prediction advisory for Hurricane Isaac, on August 26, 2012. It is obvious that these paths are quite irregular in shape, often crossing each other and sometimes doubling back on themselves. Since our goal is to create a comprehensible visualization, this argues against simply selecting a representative sampling of these paths. This is also an argument against using Curve Boxplots [17] with this data, as well as simpler direct ensemble displays such as spaghetti plots with uncertainty glyphs [28], since its incoherent structure would not lend itself well to either of these approaches. As these forecast ensembles are generated by Monte Carlo sampling of statistical models, such irregularities of individual ensemble members do not convey meaningful forecast information, and thus preserving their shapes is not important. What is important is preserving the overall spatial distribution inherent in the ensemble of storm paths. Figure 2b shows 15 paths reconstructed from the same ensemble but using our approach. Instead of selecting a confusing set of complex paths, we reconstruct a well structured set that maintains this distribution by assuring that it is preserved at each of the 120 sampled time points over the prediction.

The following describes the details of the method developed for computing a representative subset from an ensemble of forecast tracks. Formally, we can define an ensemble $\mathbf{E}$ of m tracks as $\mathbf{E} = \{\mathbf{D}_1, ..., \mathbf{D}_m\}$ where the member tracks $\mathbf{D}_i$ ($1 \leq i \leq m$) are described by a set of data points $\mathbf{D}_i = \{\mathbf{d}_{i,0}, ..., \mathbf{d}_{i,T-1}\}$, where T defines the number of samples along a track, e.g. in the NHC ensemble sampling 120-hours prediction at every hour, T equals 121; a data point $\mathbf{d}_{i,j} \in \mathbf{D}_i$ is a collection of attributes: $\mathbf{d}_{i,j} = \{t_j, p_{i,j}, A^1_{i,j}, ..., A^S_{i,j}\}$, where t_j is the time parameter of this point; $p_{i,j}$ denotes its spatial coordinates, e.g. latitude and longitude; $\{A^1_{i,j}, ..., A^S_{i,j}\}$ represents numerical values of S quantitative attributes of the data point, e.g. in a hurricane forecast ensemble these include storm size and intensity.

The steps of the framework for constructing a sample set $\tilde{\mathbf{E}}$ of $n < m$ tracks from ensemble $\mathbf{E}$ is outlined below following the pseudocode given by Algorithms 1 and 2, and the C++ implementation of

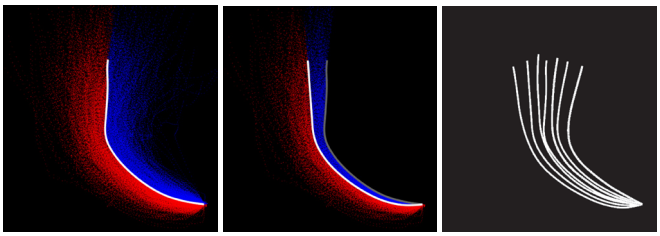


(a) randomly selected (b) our reconstruction

Fig. 2: 15 paths from an NHC ensemble.

this framework is available at `https://bitbucket.org/leliu712/hpes.git`. Briefly, Algorithm 1 uses Algorithm 2 to construct a smaller representative resampling of the tracks in the original ensemble. It then constructs a set of time sampled interpolated fields giving the spatial distributions of storm size and intensity at each time step of the original prediction ensemble, and uses these fields to annotate the resampled tracks with storm size and intensity.

Algorithm 2 constructs the representative resampling by extracting a smoothed median track from the ensemble (as shown by the white line in Figure 3a), using this median track at each time step to divide the remaining points (i.e. those not on the median track) into left and right regions (as shown by the red and blue points in Figure 3a), and then applying this same process recursively on the left region (as shown in Figure 3b), and right region. Thus, the algorithm recursively extracts new tracks in descending order of their representativeness within the original ensemble, and the merged set of median tracks determines the final resampling $\tilde{\mathbf{E}}$ (as shown in Figure 3c).



(a) level 1, median track (b) level 2, median on left (c) level 3, 7 tracks

Fig. 3: Steps in the recursive sampling process for extracting median tracks from an ensemble.

Algorithm 1: Pseudocode of our sampling framework

```
input  : Forecast ensemble E, and desired recursion level r
output : An annotated representative subset Ẽ with n = 2^r − 1 paths
Construct representative path resampling Ẽ from E via Algorithm 2 ;
foreach timestep j ∈ T do
    Compute sizeField_j using RBF interpolation [15] ;
    Compute intensityField_j using RBF interpolation ;
    foreach point p ∈ Ẽ at timestep j do
        Annotate p with storm intensity and size from intensityField_j
          and sizeField_j, sampled at p's location ;
```

2.1 Constructing a Median Track from an Ensemble

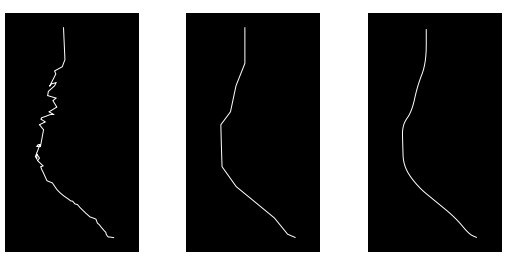
Our algorithm begins by extracting collections of spatial positions at each individual point in time from ensemble $\mathbf{E}$, e.g. the point set at time t_j is $\{p_{1,j},...,p_{m,j}\}$. Within each point set, for each time, we assign a data depth to each point using an estimate of simplicial depth [16].

Algorithm 2: Construct path resampling $\tilde{\mathbf{E}}$ from ensemble $\mathbf{E}$

```
input  : Ensemble E, and recursion level r > 0
output : Path sampling Ẽ
Initialize median track d̃ as empty ;
foreach timestep j ∈ T do
    P_j contains all points in E at time j ;
    Compute simplicial depths of points in P_j ;
    Append element of P_j with largest simplicial depth to the end of d̃ ;
Remove short-term loops in d̃ by connecting start and end points ;
Remove zigzags in d̃ using the Lang algorithm ;
Resample d̃ at uniform time steps using BSpline curve fitting ;
Ẽ = {d̃} ;
if r > 1 then
    E_l = E_r = {} ;
    foreach timestep j ∈ T do
        P_j contains all points in E at time j ;
        Determine left and right sides of median track d̃ (see
          Section 2.2);
        foreach point p_i ∈ P_j, p_i ∉ d̃ do
            if p_i is enclosed by the left region then
                Append p_i onto E_l^j;
            else
                Append p_i onto E_r^j ;
    Ẽ += call Algorithm 2 with inputs E_l, and r − 1;
    Ẽ += call Algorithm 2 with inputs E_r, and r − 1;
return Ẽ ;
```

Briefly, simplicial depth in 2D measures the centrality of a point within a dataset, by counting how many triangles formed from the points in the data set enclose that point. Thus, in a dataset with unimodal density a point near the periphery of the dataset will be enclosed by very few triangles, while one near the center will be enclosed by a much larger number. While it appears to be computationally expensive to compute, fast estimation methods have been shown to be quite robust [16]. Since simplicial depth provides a center-outward ordering, the deepest point indicates the spatial center, or median, of these points. Therefore, we choose the point with largest simplicial depth as the most significant representation of the entire set.

Subsequently, a median track of the ensemble over time is constructed by connecting each of these median points. Figure 4(a) shows a typical median path extracted in this way from an NHC tropical cyclone forecast track ensemble. As this figure shows, this raw path is highly irregular with self-intersections and zigzags, making it difficult to determine its left and right sides, but this determination is essential for the following recursive stages of the sampling algorithm. We make the assumption that these irregularities have no crucial statistical significance, as they are merely high frequency artifacts generated by the NHC Monte Carlo algorithm. Therefore, we can justifiably smooth this raw median path to remove these irregularities.



(a) raw median path (b) simplified path (c) b-spline fit

Fig. 4: Process of smoothing a median track.

Our first step in smoothing a path is to remove small scale self intersections, which in the context of tropical storm forecasting we take to be those occurring over a few hours of forecast time. These can

justifiably be eliminated because an actual storm rarely exhibits such a large variation in bearing and speed that it crosses its previous track in only a few hours. However, self intersections over long periods of time should be preserved because this phenomenon can often occur in long-term predictions. (For instance, the prediction of Hurricane Matthew at 5PM, October 6, 2016, forecast that it would reverse and hit Florida twice over a period of five days [19].) Therefore, for our loop detection algorithm we slide a small time window over each median track. If a loop is detected within the window, it is eliminated by directly reconnecting its starting and ending points, ignoring all intermediate ones. We have experimentally found that a time window of 24 hours works well over all of the ensembles that we have used.

Second, the zigzags of a track are removed by using polyline simplification. Even though there exist a number of popular techniques such as the Douglas-Peucker [9] and the Visvalingam [31] algorithms, these can produce new self intersections in the simplified results. Therefore, we chose the Lang algorithm [13], since it maintains the original points, avoids new self-intersections, and retains the positional accuracy of the original line [30]. This algorithm begins by defining a search region including a fixed number of consecutive points of the original line and then constructing a tentative simplified segment by connecting the first and last points of this region. The perpendicular distances from the interior points to the tentative simplified segment are computed. If a distance is greater than a predefined threshold, the search region is reduced by excluding the last point, and the corresponding point-to-segment perpendicular distances are computed again. This process repeats until no distance exceeds the threshold. Then, all remaining interior points of the resulting search region are eliminated, and a new search region is formed starting with the last point of the former search region. The algorithm continues until reaching the last point of the original line. Figure 4b shows the simplified version of the raw median track.

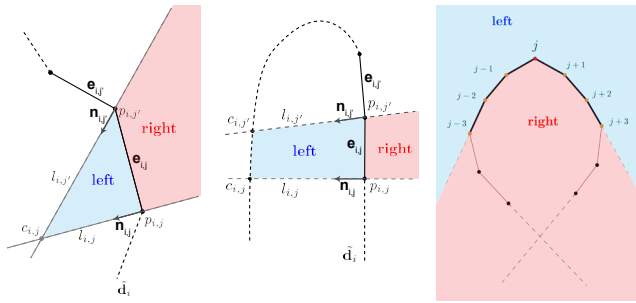
However, since the simplified track is obtained by eliminating points, it may be missing several time steps in the original forecast. To address this issue, and to obtain some additional smoothing, we construct a cubic B-spline curve [29] using the remaining points along the simplified path as control points and their associated times to construct a knot vector. We then resample this curve at the original prediction time points. A final version of the median track exemplified here is shown in Figure 4c.

2.2 Partitioning an Ensemble

The next step in the sampling approach is to divide $\mathbf{E}$ into two nearly equal smaller ensembles. To do so, we need to classify all points as being on the left or right side of the median track.

We start by defining a normal direction at each spatial point along the median track. As shown in Figure 5a, given two adjacent time points j and $j' = j+1$ on a median track $\tilde{\mathbf{d}}_i$ and their corresponding spatial locations along the track, $p_{i,j}$ and $p_{i,j'}$, a directed line segment $\mathbf{e}_{i,j}$ from $p_{i,j}$ to $p_{i,j'}$ is constructed, and its normal direction $\mathbf{n}_{i,j}$ is computed. We assign the normal direction at the point $p_{i,j}$ to be $\mathbf{n}_{i,j}$. We do this for all segments, assigning the normal direction $\mathbf{n}_{i,T-1}$ of the last point of the track to be identical to $\mathbf{n}_{i,T-2}$.

Once the normal directions are defined, the left and right sides of individual directed segments can be determined. First, the median track is extended infinitely along its starting and ending directions. Then for each segment $\mathbf{e}_{i,j}$, two infinite lines $l_{i,j}$ and $l_{i,j'}$ rooted at $p_{i,j}$ and $p_{i,j'}$ are extended along their normal directions $\mathbf{n}_{i,j}$ and $\mathbf{n}_{i,j'}$. If $l_{i,j}$ and $l_{i,j'}$ intersect at a point $c_{i,j}$ before intersecting any other part of the track, as in Figure 5a, the region enclosed by the triangle $\Delta_{p_{i,j},p_{i,j'},c_{i,j}}$ is taken to be on the left of the segment, if the dot product $\mathbf{n}_{i,j} \cdot (c_{i,j} - p_{i,j})$ is positive, or on the right if it is negative. The region enclosed by $l_{i,j}$ and $l_{i,j'}$ on the other side of $\mathbf{e}_{i,j}$ is the opposite region. If, on the other hand, $l_{i,j}$ and $l_{i,j'}$ intersect with the extended median track at points $c_{i,j}$ and $c_{i,j'}$ before intersecting with each other this determines a quadrilateral $\overline{p_{i,j}c_{i,j}c_{i,j'}p_{i,j'}}$ shown in Figure 5b. A similar left-right test can then be applied to determine whether or not this region is on the left or right, again leaving the region enclosed by $l_{i,j}$ and $l_{i,j'}$ on the opposite side of $\mathbf{e}_{i,j}$ as the opposite region. The left and right regions of the median track are then defined as the union of corresponding left and right regions of all directed line segments along the track.



(a) perpendiculars intersect (b) perpendiculars intersect path (c) time window disambiguates loops

Fig. 5: Determination of left and right regions of a directed line segment of a median path. (a) $l_{i,j}$ and $l_{i,j'}$ intersect each other, (b) $l_{i,j}$ and l_i, j' intersect $\tilde{\mathbf{d}}_i$ before intersecting each other, (c) a time window is used to disambiguate situations where a median path has a long term loop.

One complication is that a median track can potentially include loops indicating a long-term reverse of the storm, resulting in the problem that left regions of some directed line segments overlap with right regions of others, resulting in ambiguity in determining the left and right of the track. We address this issue by using the same time window introduced in 2.1. As demonstrated in Figure 5c, a sample point at time j, shown by the red point, is centered in a time window indicated by the orange points. The determination of the left and right regions of the path at time j is constrained to consider only the left and right regions of directed line segments within this time window, outlined by the bold black lines, with the first and last segments extended infinitely. Subsequently, individual points at time step j of $\mathbf{E}$ can be grouped into the left or right partition based on their time-window-constrained left or right area of the median track.

2.3 Constructing a Representative Set of Tracks

To acquire n median tracks, the previous two steps are recursively executed for the original full ensemble $\mathbf{E}$ and its following subdivisions $\mathbf{E}_L^k$ and $\mathbf{E}_R^k$, where $k \geq 1$ denotes the level of recursion. Since the total number of median tracks generated at the kth level of recursion is $2^k - 1$, the process is terminated when $k = \lceil \log_2(n+1) \rceil$.

Because any specific median track divides $\mathbf{E}$ or any succeeding portion of it into two equal halves and it is the most representative of its portion spatially and statistically, a collection of these median tracks represents $100\% \times \sum_{i=2}^{k} \frac{1}{2^{i-1}}$ of the statistical spatial distribution conveyed by the full ensemble. For instance, 7 tracks generated from 3 levels of recursion cover approximately 75% of the original spatial distribution.

One problem is that pushing the algorithm to produce a number of median tracks near to the number of the original ensemble tracks will typically produce a number of invalid median tracks. This is because the number of points remaining in partitions being subdivided decreases to the point where they are no longer sampled densely enough to represent the original distribution. This problem is exacerbated by our line smoothing algorithm, which can lead to an uneven partition of points to the left and right sides when the number of points is small. Since our method is designed to greatly reduce the number of tracks, so that $n \ll m$, this is rarely an issue. However, if it is desired to very densely sample the ensemble, we have found that filtering out median tracks that contain less than 60% of the total number of time samples in the original tracks alleviates this problem, typically leaving almost 90% of the original number of tracks, all having smoothly curved shapes.

2.4 Incorporating Ancillary Ensemble Variables

The ultimate goal of our approach is to incorporate accurately all available ensemble variables in a smaller and well spatially organized

subset of tracks. While our median track sampling approach preserves the spatial distribution of the original ensemble, and provides a good spatial layout for visualization purposes, it does not attempt to preserve the distribution of variables other than spatial position. For example, in our tropical cyclone example we are especially interested in preserving storm size and intensity. We do this using an entirely different approach.

At each time point in the original prediction ensemble we interpolate a spatial field for each variable that we wish to preserve. We do this using a Radial Basis Function (RBF) interpolation approach that we have described fully in a previous paper [15]. Briefly, for each time sample in the prediction we determine a set of weights for Gaussian kernels centered at each of the data points, whose radii are dependent on local sample density. To avoid the over-fitting problem due to the large number of original spatial locations, the RBF calculation is done on a small subset of the original points chosen to minimize the squared error of the interpolation at the known data points. This is done using a version of the Orthogonal Least Squares method [2, 3]. Then, at this time, for any point in space, the value of a variable is determined by the weighted sum of the original sample values taken at that point. For instance, to include predicted storm intensity at a specific point in time j, we compute a smooth, continuous storm intensity field using RBF interpolation over a subset of input samples chosen from all ensemble members at time point j; and the values being interpolated are their associated predicted intensities. To interpolate storm intensities over the complete time period of the forecast, RBF interpolation systems for all points in time (i.e. every hour) are constructed. Other storm characteristics such as storm size can be handled using the same strategy.

3 Validation of the Path Sampling Approach

In order to validate our algorithmic framework for path sampling, we applied it to a number of historical tropical storm prediction ensembles supplied to us by the NHC. In particular, we looked at predictions of Hurricane Katrina at 5:00 AM EDT on August 27, 2005, Hurricane Rita at 4:00 AM CDT on September 22, 2005, Hurricane Ike at 4:00 PM CDT on September 10, 2008, Hurricane Ida at 3:00 AM CST on November 8, 2009, and Hurricane Issac at 11:00 AM EDT on August 26, 2012. Each of these ensembles comprised 1,000 member tracks, with hourly time-sampled predicted locations (latitude and longitude), storm sizes, and storm intensities over the subsequent 120 hours (5 days) from the beginning of the predictions. For demonstration purposes we used only the first 72-hours (3 days) of the predictions to follow the convention of the traditional cone of uncertainty, however our approach can be applied to complete data sets without any modification.

To demonstrate that the representative tracks extracted using our approach accurately preserve the spatial distribution included in the original predictions, the original ensembles of two of these hurricanes and their corresponding subsets of representative tracks were drawn as single pixel wide red polylines over a map of the US Gulf Coast, on a 1080p HD display. Two examples are shown in a side by side comparison in Figure 6. The other storm predictions gave similar results, but are not shown due to space considerations. The left column of this figure shows the plots of the NHC ensembles, while the right shows nearly full subsets of median tracks, filtered as described in the previous section, to avoid inaccurate under sampled paths. This comparison reveals that these subsets preserve the most significant trends of these hurricanes because the high densities of tracks covered by the subsets seen in the center regions are nearly identical to those covered by the original ensembles. In addition, the individual subsets exhibit spatial spreads nearly identical, along the direction orthogonal to the predicted forward motion of the storms, to those in the original ensembles.

However, Figure 6 also shows that the tracks of these representative subsets are shorter than some of those in the original ensembles. This is because, although we preserve the directional uncertainty of the predicted storm tracks, we do not preserve the uncertainty in the predicted speed of the storm. This issue is more clearly seen when details of the coverages and their distributions are compared at specific points in time. For instance, Figure 7 shows all of the predicted storm

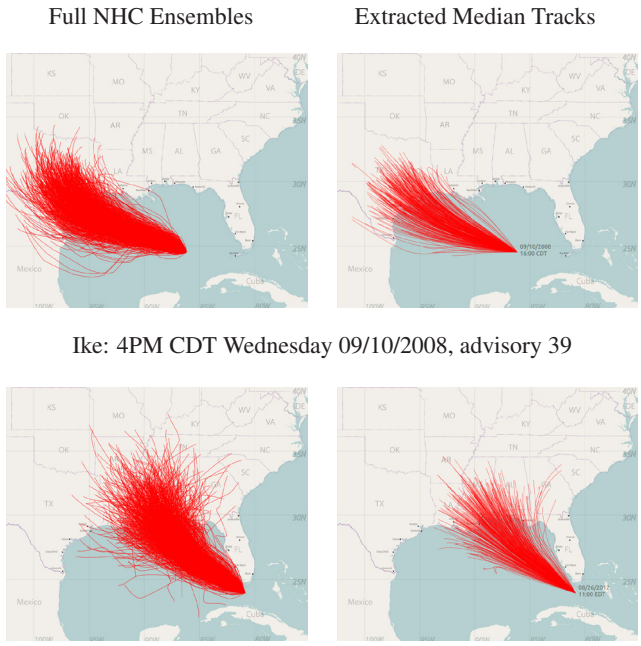


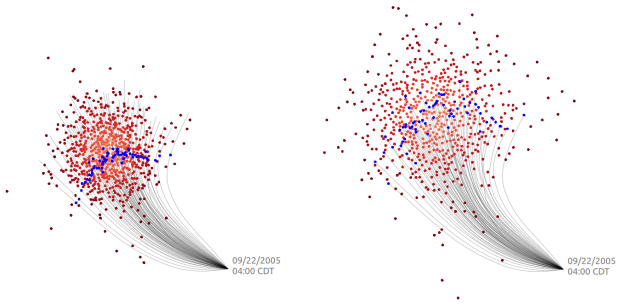
Fig. 6: A comparison between the original forecast ensembles of hurricanes Ike and Isaac (left column) and their corresponding subsets of representative tracks extracted using our algorithm (right column).

positions in the ensembles for Hurricane Ida at 60 and 72 hours from the beginning of the prediction. The red points represent the predicted locations extracted from the original ensemble, the simplicial depths of which are color-encoded using luminance, the gray polylines represent representative tracks extracted from the ensemble using our approach, and the blue points represent locations along the representative tracks at these specific points in time. As these figures show, the blue points are located at the center of the distribution (i.e. they cover the area with the brightest red points), and they faithfully represent the spatial uncertainty across the distribution (i.e. the blue outliers closely match those of the red along the direction perpendicular to the tracks). However, the blue sample points do not maintain the distribution along the forward direction of the tracks, which is a result of variation in the predicted speed of the storm.

Representing speed uncertainty using a small set of tracks that are parameterized by time is inherently difficult, because of the sparse sampling. Unless a very large number of tracks of different lengths are drawn, it will be impossible to show this aspect of the uncertainty in the prediction. It could be encoded in various ways, such as using glyphs or overdrawing alternative tracks, as shown in Figure 8. However, glyphs will inevitably overlap with those portraying possible predicted locations of adjacent hours as represented by the red and blue circles in Figure 8a. Likewise, increasing the number of tracks, as in Figure 8b, will result in visual clutter. In our opinion, both of these approaches will result in a confusing presentation, and we settle for accurately representing the median storm speed at each time step. However, this limitation can be avoided using point-based visualization approaches such as those proposed by Liu et al. [14].

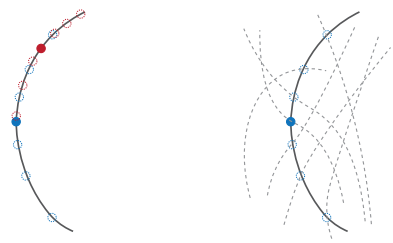
4 Visualization Design

In designing a visualization based on our path sampling approach, we had two primary considerations. First, we needed to decide on the right number of tracks to faithfully represent the uncertainty in the prediction. Second, we needed to decide on an effective visual encoding of storm intensity and size to be applied to the paths. We describe below our design choices, and in Section 5 we summarize the results of a cognitive



(a) at 60 hours (b) at 72 hours

Fig. 7: Detailed illustrations of representative tracks of Hurricane Ida at two different times from the beginning of the prediction.



(a) dashed circles at alternative positions (b) dashed alternative tracks

Fig. 8: Two ways of visualizing speed uncertainty in storm tracks.

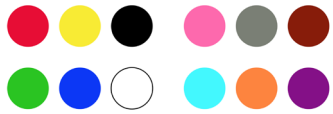


Fig. 9: Ware's catalog of 12 discriminable colors.

We began our design using the color palette recommended by Ware [32], reproduced in Figure 9, which consists of a set of colors that are easily distinguishable, and remembered. We intentionally avoided using blue, cyan, white and black from this palette as they were too close to the colors used in our background map. We also avoided yellow as it is so bright that it can be difficult to distinguish when drawn on a white background. Therefore, we chose the remaining seven colors for our storm intensity encoding: red, green, pink, gray, orange, brown, and purple. We then slightly modified the specific colors suggested by Ware to eliminate large changes in value across colors, which can cause an individual color to be over-emphasized.

The resulting storm intensity encoding scheme is illustrated in Figure 10 along with the sRGB values for the colors given on a 0-255 scale. These colors are sorted in ascending order based on their red channel, placing red after pink, so that green represents the least dangerous intensity (tropical depression) while red depicts the most dangerous intensity (category 5 hurricane).

Storm Intensity	R	G	B
Tropical Depression	90	173	97
Tropical Storm	102	102	102
Category 1	153	112	171
Category 2	191	129	45
Category 3	224	130	20
Category 4	255	127	127
Category 5	255	80	80

Fig. 10: Our storm intensity color palette.

study meant to determine the efficacy of our choices.

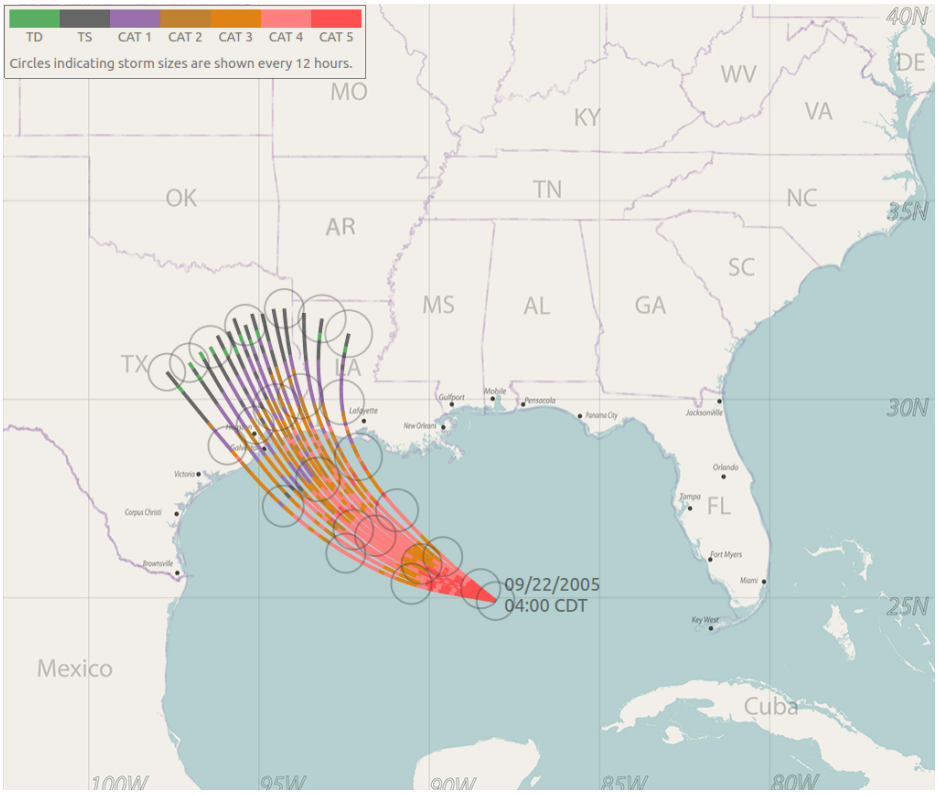
To determine the number of tracks to display we considered two factors. First, we felt that the tracks should cover as much as possible of the spatial spread of the full ensemble, to provide viewers with enough information about the full range of likely storm paths. Second, since an NHC track ensemble is usually quite dense near its spatial center, adding too many members to the extracted subset would lead to overdrawing, making it difficult to apply effective annotations to the tracks as they are drawn. Considering both of these factors, we decided on using subsets comprised of 15 tracks. These convey approximately 87.5% of the spatial spread of the full ensembles, and still provide adequate spacing between tracks. These tracks are drawn as polylines over a map of the region expected to be affected by the storm.

The predicted storm intensities are encoded in this visualization by rendering each segment of a median path using a color determined by the predicted intensity attached to the starting location of this segment. The NHC ensembles describe storm intensity as the maximum predicted wind speeds measured in nautical miles per hour. To communicate intensity more effectively to the general public, these wind speeds are converted to the Saffir-Simpson Hurricane Wind Scale [21], a widely adopted storm taxonomy system. This system divides hurricane intensities into five categories, with category 1 being the least severe (wind speeds between 64 and 82 kts), and category 5 (wind speeds above 137 kts) being the most severe. The NHC tracks tropical cyclones even when they are not of hurricane strength, using the two additional categories of tropical depression (less than 34 kts), and tropical storm (34-63 kts) to identify their intensities. Since our goal was to make the resulting seven storm categories in an NHC advisory highly visually distinguishable, we felt that these could be most appropriately portrayed using a color scheme with seven distinctly recognizable labeling colors, rather than a set of steps in hue, saturation, or value.

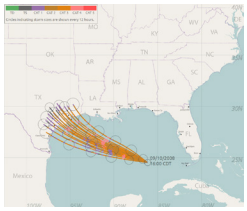
The colored storm tracks are drawn as polylines over a regional map. We experimentally determined that a line width of 4 pixels on a 1080p HD display provided good color perception, while maintaining spatial separation across most of the prediction. Lines are drawn from the start of the prediction, advancing in time by one hour for each line segment. When drawing colors, we use a *deadband* of 10% around the wind speed range of the current storm category when determining the category to be used for the color of the next segment. This helps to minimize frequent anomalous changes of color along the line that would otherwise occur if the predicted wind speed were close to one of the category boundaries.

Finally, the predicted storm sizes are included in our visualization by drawing circles centered at locations along the tracks. The radii of these circles are determined by the predicted radii of 64-kt (i.e. hurricane force) winds at these locations, thus the areas encompassed by these circles indicate the region around the storm center predicted to experience hurricane force winds. Since drawing storm size circles at every hour of a prediction would lead to severe overdrawing, the storm size circles are shown only every 12 hours. This has the added advantage that the positions of the circles indicate 1/2 day increments along that prediction. To further avoid the overdrawing problem the storm size circles are drawn on the most extreme left and right tracks, and then spaced across internal tracks to maintain spatial separation. In this way, the area potentially affected by hurricane force winds, to a certain level of confidence, is depicted using a few easily visible and interpreted circles.

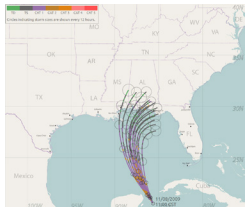
Using these design guidelines, visualizations for five NHC forecast track ensembles were produced. These are depicted in Figure 11, with the forecast for Hurricane Rita shown in detail in Figure 11a. Space limitations require that the other four storm predictions be shown at reduced resolution. As seen in these figures, we include a color



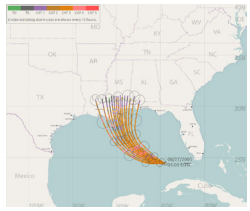
(a) Rita: 4AM CDT Thurs. 09/22/2005, advisory 17



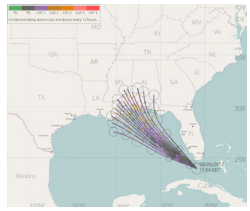
(b) Ike: 4PM CDT Wed. 09/10/2008, advisory 39



(c) Ida: 3AM CST Sun. 11/08/2009, advisory 17



(d) Katrina: 5AM EDT Sat. 08/27/2005, advisory 16



(e) Isaac: 11 AM EDT Sun. 08/26/2012, advisory 22

Fig. 11: Example visualizations of the NHC forecast track ensembles of five historical hurricanes.

key in the upper left-hand corner of each map to assist viewers in understanding the storm intensity color encoding.

An important benefit of this visualization is that crucial time-based information included in a forecast is clearly depicted. For instance Figure 11a shows that at the time of the storm advisory, Hurricane Rita was as a Category 5 storm, but predicted to drop to a Category 4 storm over the next 24 hours, and become Category 3 approximately 36 hours from the beginning of the initial prediction. Landfall is predicted near to 48 hours, and the storm is seen as weakening to Categories 2, 1, and below as it passes over land.

5 Experimental Evaluation

To evaluate the effectiveness of our visualizations in enhancing viewers' estimates of hurricane risk, we conducted an extensive cognitive study. Our study adopted the experimental design used in other recent studies exploring how participants' judgments of potential storm damage, and their confidence in these judgments, varied as a function of visualization type [22, 27]. These previous studies indicated that an ensemble visualization led to damage estimates that conform to a correct interpretation of the uncertainty in the prediction, whereas all visualizations based on the cone produced estimates indicating that increasing uncertainty was being confounded with an increase in storm size and intensity [22, 27]. Having this strong indication of the superiority of the ensemble approach in conveying uncertainty, in our new study we focused on how the ensemble can be best presented, and not on comparisons with other approaches as has already been done in prior work [14, 22, 27].

The two primary factors that we wanted to explore were how judgments of the potential for storm damage are affected by 1) the number of tracks drawn, and 2) the presence of annotations indicating storm size and intensity. We hypothesized that the number of tracks drawn would influence damage ratings (H1). We also predicted that including annotations of storm size and intensity would help viewers more accurately calibrate their damage judgments (H2). To test H1, we examined viewers' damage judgments when viewing the tracks-only displays drawn using red polylines showing either 7, 15, or 63 storm forecast tracks as between-subjects variables (n = 50, n = 51, n = 50 respectively). The number of participants was selected based on a power analysis conducted in prior work using damage ratings [27]. To test H2, we compared viewers' damage judgments while viewing annotated displays showing 15 tracks drawn using encodings of storm intensity and size (n = 50) to the displays showing 15 tracks without annotations utilized to test H1, as between-subjects variables. Figures 12(a), (b), and (c) show examples of the three levels of tracks-only displays, and (d) shows an example of the annotated display. Participants were undergraduate psychology students (male = 49, female = 152) who

participated in the experiment for course credit, with a mean age of 23.5 years ($SD = 6.6$). This naive population was desirable to examine the effects of the visual properties that we manipulated and not viewers' prior knowledge about hurricanes, as less than 3% of this population has lived in a hurricane effected region or has experienced hurricanes.

To test both H1 and H2, the participants' task was to estimate potential storm damage to a simulated offshore oil rig platform, whose placement constituted the within-subjects factors in the experiment. The location of the oil rig was indicated by a blue dot, such as the one that can be seen on each of the maps in Figure 12. In line with prior work showing that the influence of visualization characteristics varied as a function of time into the forecast and oil rig distance from the center of the storm [14, 27], we tested viewers' damage judgments for two time points (24 hours and 48 hours) at four distances from the center of the storm. Matching prior work [14, 27], distances from the central track were placed along a line perpendicular to the central track direction, at the selected forecast time, at distances of $\pm 0.20, 0.80, 1.10$, and 1.78 of the NHC 67% prediction error estimates of 186 km at 24 hours, and 347 km at 48 hours. These distances are those used by the NHC for setting the width of the uncertainty cone visualization, and are determined by tracking the NHC's historical forecast error over a time window of five years [20].

After consenting to participate and being given a description of their task, participants in each between-subjects group (7, 15, 63, and 15-annotated) completed 80 trials in a randomized order, consisting of locations of the platform relative to the central track (8 conditions), the forecasting time points (2 conditions), and the storms (5 conditions). In each trial, on a 1-7 Likert scale, the participants were instructed to estimate the level of damage that might be incurred at the simulated oil rig based on the likelihood of the storm affecting it and the storm's strength in the affected region. They were also instructed to enumerate their level of confidence in their estimations, also on a 1-7 scale.

5.1 Experimental Results

All of the details described in the analysis below, along with additional analyses, figures and tables that space did not permit can be found at the website `http://lacepadilla.com/exp/ensemble2018/Analysis.html`. To allow for comparison with prior work [14, 22, 27], multilevel models (MLM) were fit to the data using *R* with the *lme4* package [1] and using restricted maximum likelihood estimation procedures [25, 26]. A multilevel is model is appropriate for modeling these data as it uses robust estimation procedures appropriate for partitioning variance and error structures in mixed designs, as in the current study. To test H1, two MLMs were computed using the interaction between visualization type, distance, and time point, along with the corresponding lower level interactions and main effects as fixed effects and participants as random effects to account for variance in damage judgments. The same analysis was used to examine variance in confidence judgments. The variables were coded such that the 7 track display, 24-hr time point, and closest location to the center of the storm were the referents by which the other variables were compared to. Similar models were used for testing H2, where visualization type compared the 7 track unannotated display (referent) to the annotated version.

5.2 Results: H1

The following section will detail the interactions intended to test H1 and then utilize post hoc MLMs to clarify the relationships. The fixed and random effects explained 56% of the variance in damage ratings (conditional R^2). The results of this analysis revealed a significant three-way interaction between distance, time point, and the 7 vs. 63 track display, $b = 0.59$, $t = 5.81$, $p < 0.001$, 95 % CI [0.39, 0.78].

This interaction can be seen in Figure 13. Here the orange lines represent the results at 24 hours into the prediction and the blue lines at 48 hours. For the 63 track display, the 24 hour time point has an overall slope of -2.45, meaning that for every one unit change in distance, damage ratings decrease by 2.45 on average on the Likert scale from 1-7, conditional $R^2 = .55$. Whereas, for the 48 hour time point, the average slope is -1.77, which is a significant difference between the 24- and 48-hr slopes of .68 ($b = 0.67$, $t = 9.18$, $p < 0.001$, 95 % CI [0.53, 0.82], conditional R^2 for the model = 0.53). For the 63 track display, it is apparent that the damage to distance relationship is flatter at 48-hr than at 24-hr, which is consistent with the increased spatial spread of predicted paths. This suggests that participants' attention is being drawn away from the center of the prediction as the spatial spread of the prediction increases, an effect that we do not see for either the 7 or 15 track displays. For the 7 track display, there is not a significant difference in slope between the orange and blue lines (mean 24-hr slope = -2.10, mean 48-hr slope = -2.01, $b = 0.08$, $t = 1.34$, $p = 0.17$). It should be noted that while there was not a three-way interaction between distance, time point, and the 15 vs. 63 track displays, the slopes for the 15 track display were not significantly different as a function of time point, (mean 24-hr slope = -2.1, mean 48-hr slope = -2.01, $b = 0.12$, $t = 1.62$, $p = 0.10$). In summary, our conclusion is that when using a tracks-only display, more than 15 tracks must be drawn in order for viewers to understand the uncertainty information in the display, which is in line with H1 suggesting that the number of lines drawn influences viewers' judgments of damage.

The analysis of confidence levels showed that for the tracks-only displays, participants were more confident about their estimations when shown the 63 track display than with the 15 and 7 track displays ($p = 0.02$). There was no significant difference in confidence between the 15 and 7 track displays.

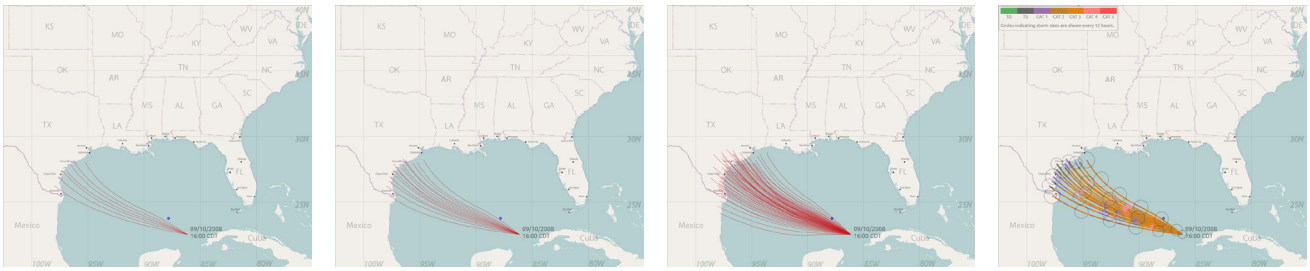
5.3 Results: H2

In order to explore the effect of adding annotations of storm intensity and size to the display, we first conducted an analysis comparing 15 track tracks-only displays to 15 track annotated displays, using similar MLMs as previously described. Figure 13 shows the corresponding damage to distance relationships. Use of annotations significantly influences the relationship between distance-to-center and damage rating ($b = 0.68$, $t = 9.10$, $p < 0.001$, 95 % CI [0.53, 0.82], conditional R^2 for the model = 0.49). It is apparent in the figure that including annotations results in lower damage ratings at the spatial center of the prediction, and higher ones far from the center, indicating that the annotated visualization is conveying the uncertainty in the prediction.

Further, we tested our prediction that the annotations of storm size and intensity would help viewers more accurately calibrate their damage judgments by examining if the specific intensity and storm size at the time point of the oil rig accounted for a significant proportion of variance in viewers' damage ratings. Focusing on just the participants that viewed the 15 track annotated display, we conducted an MLM that included the interaction between distance and time, along with size and intensity of the storm at a given time point for each hurricane. This analysis revealed that both size ($b = 0.013$, $t = 4.32$, $p < 0.001$, 95 % CI [0.007, 0.019]), and intensity ($b = 0.01$, $t = 10.06$, $p < 0.001$, 95 % CI [0.008, 0.011]) accounted for a significant proportion of variance in damage ratings over and beyond effects of distance and time (conditional R^2 for the model = 0.49). In line with our prediction, viewers' damage ratings increased as both size and intensity increased, which is evidence for H2. Additionally, a post hoc analysis revealed that neither accounted for a significantly larger proportion of variance, meaning that they equally influenced damage ratings.

For confidence ratings, there was no significant difference in confidence between the 15 track annotated display and the 15 track tracks-only display. This last result is somewhat surprising, given the overall better performance with the annotated display.

Since the 63 track display was superior to the 15 track display, it was important also to compare it to the 15 track annotated display. This comparison can be made visually by comparing the 2nd and 4th columns of Figure 13. Our analysis revealed a significant three-way interaction between distance, time and the two visualization types ($b = -0.66$, $t = -6.25$, $p < 0.001$, 95 % CI [-0.86, -0.45], conditional R^2 for the model = 0.49). Comparing the damage to distance trend lines in the figure reveals an overall flatter relationship when the annotation is used, indicating that participants became more aware of the spatial spread of uncertainty in the prediction, even though the annotated display uses less than 1/4 of the number of tracks in the unannotated display. However, unlike the 63 track trend lines, the 15 track annotated



(a) 7 tracks, unannotated (b) 15 tracks, unannotated (c) 63 tracks, unannotated (d) 15 tracks, annotated

Fig. 12: Examples of visualizations compared in the cognitive study. The blue dots indicate the position of a simulated oil rig platform.

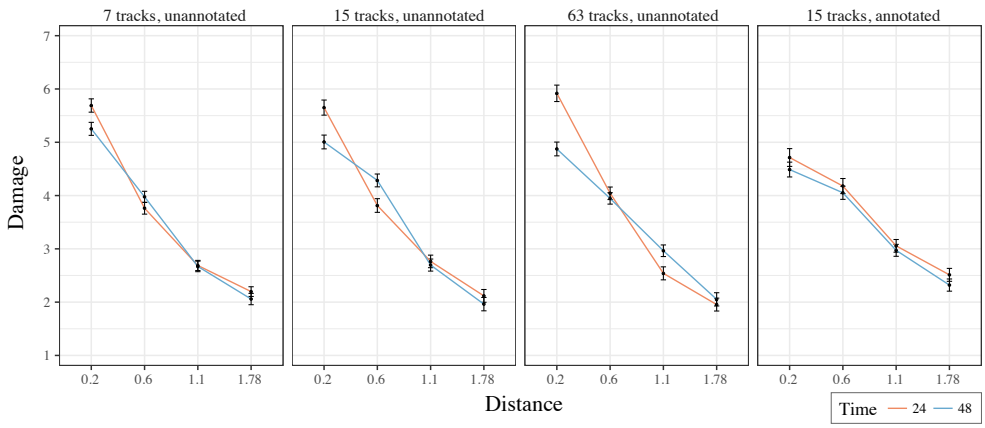


Fig. 13: Mean damage judgments and 95% confidence intervals for 7, 15, and 63 unannotated, and 15 annotated track displays at 24 and 48 hours. Confidence intervals computed using the 2008 Morey correction [18].

trend lines stay roughly parallel, meaning they do not have significantly different slopes ($b = 0.00$, $t = 0.22$, $p = 0.82$). In contrast, the 24-hr slope for the 63 track display is significantly steeper than the 48-hr slope, as revealed in the analysis for H1. This suggests that the ideal annotated display should contain more than 15 tracks so that the effect of the increased spatial spread of predicted paths with increasing time is better recognized. Optimizing the number of tracks will be the subject of a future study.

In summary, the key findings from the experiment are the following. If drawing unannotated tracks, more than 15 tracks are required before a fair reading of the uncertainty in the prediction is indicated. We specifically showed that 63 tracks work well. Further, with annotation, there are strong positive relationships between predicted storm size and intensity and assigned damage estimates. On the other hand, when shown 15 tracks annotated with our representations of storm size and intensity, participants were more cognizant of the spatial spread of uncertainty than with 63 unannotated tracks, although there is an indication that using somewhat more than 15 annotated tracks would improve sensitivity to increasing spatial spread with increasing time point. Thus, the annotation is helpful both in reducing visual clutter by lowering the number of tracks that need to be drawn to convey uncertainty, and in conveying significant elements of storm risk.

6 Conclusion

We have shown how an ensemble of paths describing an uncertain process can be reduced to a representative set and used to produce a visualization implicitly representing the uncertainty by its spatial distribution. Our approach is applicable when it is not necessary to preserve the original tracks, but it is satisfactory to construct a new set that maintains the original set's spatial distribution. Our algorithm constructs a set of representative tracks by recursively finding median paths that divide the set into nearly equal sized subsets. The algorithm preserves non-spatial variables associated with the original tracks by interpolation and resampling onto the newly constructed paths. We demonstrated this approach as applied to tropical cyclone prediction ensembles produced by the US National Hurricane Center (NHC), and developed a visualization based on our algorithm that is able to encode storm timing, size, and intensity information in the form of annotated tracks drawn over a map. Finally, we described a cognitive experiment providing guidelines for design. It demonstrates how the number of paths used in the visualization affects its ability to convey the spatial uncertainty in the storm prediction, and indicates that the visualization's annotations work well in conveying storm prediction attributes.

However, there is still work to be done to perfect the method if it were to be used during a live event. In a personal communication from Mark DeMaria, Branch Chief of the Technology and Science Branch of the NHC, he said that we show some interesting results on new ways to display the underlying data from their Monte Carlo wind speed probability model, and that the user tests clearly show benefit from the size and intensity information annotations. However, he is concerned that since we do not maintain the storm's speed uncertainty we risk underestimating or overestimating the strike time of a storm. In addition, he points out that our interpolation technique will tend to underestimate the effect of landfall on reducing the storm's strength, and that interpolation can decouple the storm size and intensity information so that a storm below hurricane strength can still show a non-zero hurricane force wind radius. We are aware of these limitations, and plan to address them in our future work to improve the approach.

Acknowledgments

The authors would like to thank Mark DeMaria and Edward Rappaport of the NHC, and Matthew Green of FEMA for key advice and critique. We also thank the InfoVis anonymous reviewers who helped us to greatly improve our presentation. The prediction ensembles used in our study were supplied by Mark DeMaria. Our work was supported in part by the NSF under Grant Nos. IIS-1212501, and IIS-1212806.

REFERENCES

[1] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01

[2] S. Chen, E. Chng, and K. Alkadhimi. Regularized orthogonal least squares algorithm for constructing radial basis function networks. *International Journal of Control*, 64(5):829–837, 1996.

[3] S. Chen, C. Cowan, and P. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.

[4] J. Christiansen. Visualizing uncertain weather. *Scientific American Visual* `http://blogs.scientificamerican.com/sa-visual/visualizing-uncertain-weather`, 2017.

[5] M. Correll and M. Gleicher. Implicit uncertainty visualization: Aligning perception and statistics. *IEEE VIS 2015 Workshop "Visualization for decision making under uncertainty"*, 2015.

[6] J. Cox, D. House, and M. Lindell. Visualizing uncertainty in predicted hurricane tracks. *International Journal of Uncertainty Quantification*, 3(2):143–156, 2013.

[7] M. Demaria, John, A. Knaff, M. Brennan, D. Brown, R. Knabb, R. T. Demaria, A. Schumacher, C. Lauer, D. Roberts, C. Sampson, P. Santos, D. Sharp, and K. Winters. Improvements to the operational tropical cyclone wind speed probability model. *Wea. Forecasting*, 28(3):586–602, 2013.

[8] M. Demaria, John, A. Knaff, R. Knabb, C. Lauer, C. R. Sampson, and R. T. Demaria. A new method for estimating tropical cyclone wind speed probabilities. *Wea. Forecasting*, 24:1573–1591, 2009.

[9] D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10(2):112–122, 1973. doi: 10.3138/FM57-6770-U75U-7727

[10] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18. ACM, New York, NY, USA, to appear.

[11] J. Hullman, M. Kay, Y. S. Kim, and S. Shrestha. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):446–456, 2018. doi: 10.1109/TVCG.2017.2743898

[12] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 5092–5103. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858558

[13] T. Lang. Rules for robot draughtsmen. *Geographical Magazine*, 42(1):50–51, 1969.

[14] L. Liu, A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. Uncertainty visualization by representative sampling from prediction ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2165–2178, 2017.

[15] L. Liu, M. Mirzargar, R. Kirby, R. Whitaker, and D. House. Visualizing time-specific hurricane predictions, with uncertainty, from storm path ensembles. *Computer Graphics Forum (Proceedings of EUROVIS '15)*, 34(3):371–380, 2015.

[16] R. Y. Liu. On a notion of data depth based on random simplices. *The Annals of Statistics*, 18(1):405–414, 1990.

[17] M. Mirzargar, R. Whitaker, and R. Kirby. Curve boxplot: Generalization of boxplot for ensembles of curves. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2654–2663, Dec 2014.

[18] R. D. Morey. Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, 4:61–64, 2008.

[19] NOAA. Matthew graphics archive. *National Hurricane Center* `http://www.nhc.noaa.gov/aboutsshws.shtml`, 2016.

[20] NOAA. Definition of the nhc track forecast cone. *National Hurricane Center* `http://www.nhc.noaa.gov/aboutsshws.shtml`, 2017.

[21] NOAA. Saffir-simpson hurricane wind scale. *National Hurricane Center* `http://www.nhc.noaa.gov/aboutsshws.shtml`, 2017.

[22] L. M. Padilla, I. T. Ruginski, and S. H. Creem-Regehr. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive research: principles and implications*, 2(1):40, 2017.

[23] A. Pang. Visualizing uncertainty in natural hazards. In *Risk Assessment, Modeling and Decision Support*, vol. 14 of *Risk, Governance and Society*, pp. 261–294. Springer Berlin Heidelberg, 2008. doi: 10.1007/978-3-540-71158-2_12

[24] K. Potter. *The visualization of uncertainty*. PhD thesis, The University of Utah, 2010.

[25] S. Raudenbush, Y. C. A.S. Bryk, R. Congdon, and M. du Toit. *Hierarchical Linear Modeling 7*, 2011.

[26] S. Raudenbush and A. Bryk. Hierarchical linear models: Applications and data analysis methods. *Sage*, 1, 2002.

[27] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016.

[28] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1421–1430, 2010.

[29] I. J. Schoenberg. Contributions to the problem of approximation of equidistant data by analytic functions. part a. on the problem of smoothing or graduation. a first class of analytic approximation formulae. *Quarterly of Applied Mathematics*, 4(1):45–99, 1946.

[30] W. Shi and C. Cheung. Performance evaluation of line simplification algorithms for vector generalization. *The Cartographic Journal*, 43(1):27–44, 2006.

[31] M. Visvalingam and J. D. Whyatt. Line generalisation by repeated elimination of the smallest area. Technical report, Cartographic Information Systems Research Group (CISRG), The University of Hull, Discussion Paper, 1992.

[32] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.