The influence of graphical format on judgmental forecasting accuracy: Lines versus points

Zoe Theocharis[1], Leonard A. Smith[2] and Nigel Harvey[1]

1. University College London

2. London School of Economics and Political Science

Nigel Harvey

Department of Experimental Psychology

University College London

Gower Street

London WC1E 6BT

UK

Tel: +44 207 679 5387

Fax: +44 207 436 4276

Email: n.harvey@ucl.ac.uk

Word Count: 9998

Running head: Graphical format effects in forecasting

Acknowledgement: This work was funded by an SAS/IIF Grant to Support Research on Forecasting

Abstract

People made forecasts from real data series. The points in the series were un-trended and independent. Hence forecasts should have been on the mean value. However, consistent with previous research on forecasting biases, forecasts were too close to the last data point. It appears that forecasters see positive sequential dependence where none exists. In three experiments, we examined this bias in different types of forecasting task: point forecasting, probability density forecasting, interval forecasting. In all cases, we found that it was greater when the data series were displayed using continuous line graphs than when it was displayed using discrete point graphs. Consistent with arguments made by Zacks and Tversky (1999), we suggest that people are more likely to group data together and to see patterns in them when those data are presented in a continuous than in a discrete format. These findings have implications for forecasting practice.

Key words: Forecasting, Judgment, Format effects, Graphical displays

The influence of graphical format on judgmental forecasting accuracy: Lines versus points

In most forecasting tasks, predictions for the future values of some variable are based on a record of previous values of that variable (the data series). For example, in demand forecasting for supply chain management, forecasters predict future sales of products from the past sales of those products. Forecasts are produced in one of three ways. In computer-based forecasting, a formal (e.g. statistical) procedure realised on a computer is used to produce the forecasts by processing the data series; judgment plays no role. In judgmental forecasting, unaided human judgment is used to analyse the data series and produce the forecasts from it. In computer-aided forecasting, a formal computer-based procedure and human judgment are combined in some way. For example, the average of the two forecasts independently produced by these methods may be used as the final forecast. Alternatively, the forecaster may use judgment to make an adjustment to the formal forecast.

Fildes and Goodwin's (2007) large scale survey of company forecasting indicated that computer-based forecasting was employed in about a quarter of cases, judgmental forecasting was used in a further quarter of cases, and computer-aided forecasting was adopted in the remaining cases. A more recent but smaller scale survey by Fildes and Petropoulos (2015) showed no change in the frequency of computer-based forecasting but a lower level (16%) of judgmental forecasting (and a correspondingly higher level of computer-aided forecasting). Both surveys concur that judgment is involved in producing about three-quarters of company forecasts. Here we describe studies of judgmental forecasting. However, our findings are likely to generalize to the judgmental component of computer-aided forecasting.

Judgmental forecasting

There is now a large body of work on judgmental forecasting. Findings have been thoroughly reviewed (Goodwin and Wright, 1993, 1994; Lawrence, Goodwin, O'Connor and Önkal, 2006; Webby and O'Connor, 1996). It has become clear that, relative to the forecasts produced by statistical

means, judgmental forecasts are biased in various ways. These biases include trend damping (Eggleton, 1982), elevation of desirable variables and depression of undesirable ones (Lawrence and Makridakis, 1989), addition of noise to forecasts (Harvey, 1995), and misperception of sequential dependence (Eggleton, 1982). Here we focus on this last effect: misperception of sequential dependence.

In an un-trended series of independent data points, a forecast should lie on the mean value of those points. However, it is typically found to be too close to the last point in the data series. This is where it ought to be when the series contains positive first-order autocorrelation (i.e., sequential dependence). For example, for a series with an autocorrelation of 0.5, the forecast should be half way between the last data point and the mean of the series. Thus, it appears that people perceive independent series as if they were sequentially dependent. A process account of this phenomenon based on use of the anchor-and-adjust heuristic can be proposed. People anchor their judgment on the last data point and then adjust towards the mean of the series. Because adjustment is insufficient when this heuristic is used (Tversky and Kahneman, 1974), the forecast is not as close to the mean as it should be.

Format effects

Data series can be presented to forecasters in a tabular or in a graphical format. Within each of these broad categories, there are sub-divisions: tables of data can be presented in a horizontal row or in a vertical column; graphs of data can be provided as line graphs, as point graphs, or as bars.

There is some consensus that graphical presentation produces better performance than tabular presentation when people are required to use their judgment to analyse trends and to make forecasts (Coll, Thyagarajan and Chopra, 1991; Dickson, DeSanctis and McBride, 1986; Tullis, 1988). Harvey and Bolger (1996) confirmed that graphical presentation is superior for this purpose when data contain trends and showed that this was because trend damping was much greater with tabular presentation.

Research on the effects of using different types of graphical presentation is more limited. Newman and Scholl (2012) presented people with bars representing the mean values of the data set and asked them to judge the likelihood that a point placed above or below the top of the bar was part of the underlying distribution. Participants gave higher values for this likelihood for a point below the top of the bar (i.e., within the bar) than for a point the same distance above the top of the bar (i.e., outside the bar). Okan, Garcia-Retamero, Cokely and Maldonado (2018) showed that this bias could be markedly reduced by presenting the mean values as points rather than as bars. This was so even though participants rated the bar graphs more positively than the point graphs.

Harvey and Reimers (2012) required people to make forecasts from data series presented as bars, line graphs, or point graphs. Forecasts were systematically lower (and error in them was correspondingly higher) when series were presented as bars than when they were presented in the other formats. This bias was present for series containing upward trends and for those containing downward ones. It would counteract the effect of damping when data contain downward trends but would reinforce it when they contain upward trends. The effect was reversed with hanging bars: forecasts were then systematically higher with bars than with the other two formats. It therefore appears that bars draw people's attention towards them in a manner that other formats do not and that the way that they position their forecasts is affected by this attentional displacement.

In summary, the above research indicates that presentation of data as bar graphs is associated with certain biases and that these biases can be reduced by using line graphs or point graphs instead. Here we ask whether a difference between presenting data as line graphs and presenting them as point graphs also results in a difference in the way that forecasts are made. To date, no such differences have been reported but research on graphical perception leads us to expect that this difference in graphical format will affect the degree to which sequential dependence is misperceived.

Bar graphs are recommended for use when the horizontal axis refers to discrete categories, such as male versus female. Line graphs are more appropriate when it refers to a continuum, such as age. In practice, these recommendations are not always followed. Zacks and Tversky (1999) presented participants with the same set of data displayed either in a bar graph or in a line graph. They were more likely to describe the relationship between x and y variables as continuous when a line graph was used. For example, some participants presented with line graphs showing height on the y-axis plotted against sex on the x-axis, described the relationship as "The more male a person is, the taller he/she is". In contrast, bar graphs merely led to the observation that, on average, men are taller than women. These findings suggest that people are more likely to group data together and to see patterns in them when those data are presented in a continuous than in a discrete format. Conversely, the discrete format emphasises the frequency and range of each category rather than the relationship between those categories.

We have seen that people over-emphasise the relation between successive points in a time series: they anchor their forecasts too strongly on the last data point. Zacks and Tversky's (1999) findings show that use of a discrete format serves to de-emphasise the relation between successive data points. As a result, forecasts should be less strongly anchored on the last data point. Thus, when there is no autocorrelation in the data series, this should lead to forecast being more accurate when a discrete format is used to present data than when a continuous format is used.

Hypotheses

Here we compare forecasting from continuous line graphs with forecasting from discrete point graphs. We test the hypothesis (H_1) that people make forecasts closer to the last data point when data series are presented as continuous line graphs than when they are presented as discrete point graphs. We also test the hypothesis (H_2) that, in a series with no significant autocorrelation, this will result in more accurate forecasts with the discrete point graphs.

To ensure that our findings are generalisable, we test these hypotheses in the three different types of forecasting task that are commonly used by practitioners: point forecasting (Experiment 1), probability density function forecasting (Experiment 2), and prediction interval forecasting (Experiment 3).

Experiment 1: Point forecasting

In point forecasting, predictions for the most likely value of a variable are made. Forecasts may be made just for the immediately upcoming period or for more distant forecast horizons as well.

Method

Participants made predictions for the next five values of a 30-point time series. Once they had done that, the time series rolled forward by one time-period and this process was repeated for 13 trials. The time series were presented to participants as continuous line graphs or as discrete point graphs.

Participants In total, 60 students (46 females) from University College London acted as participants. Their mean age was 20 years. They were not paid for their participation.

Design Thirty participants were randomly allocated to each group. The first group produced point forecasts from continuous line graphs while the second group made their predictions from unconnected point graphs.

Stimulus materials The data series comprised a real-life series from which forecasting practitioners have made predictions. The series described the annual number of hurricanes hitting the Atlantic coast of the USA from 1966 to 2012. All data were drawn from official sources provided by the USA National Oceanic and Atmospheric Administration (NOAA) Office of Oceanic and Atmospheric Research (http://www.aoml.noaa.gov/general/lib/lib1/nhclib/mwreviews/mwreviews.html) [1]. In the current work, only a subset of this hurricane occurrences database was displayed (1966 to 2007) because satellite technology was available to accurately monitor hurricane activity only from this

period onwards. Neither autocorrelation (AR1 = 0.04) nor global trends in the series reached statistical significance.

On each trial, participants saw 30 years of this series. Thus, on the first trial, they saw the series of the number of hurricanes striking the Atlantic coast of the USA between 1966 and 1995 and made predictions for 1996 to 2000. On the next trial, they saw the series for 1967 to 1996 and made predictions for 1997 to 2001. This rolling procedure continued until the 13th trial when they saw the series for 1978 to 2007 and made predictions for 2008 to 2012.

In all displays, the y-axis showed the number of hurricane occurrences while the x-axis represented time in years. In continuous line graph displays, the data points were connected by a continuous line (Figure 1, upper panel); in the discrete point graph displays, they were not (Figure 1, lower panel).

_____

Figure 1 about here

_____

Procedure Each participant performed the task individually on a computer. They read a short introduction and then entered their demographic details (age, sex). Instructions were as follows:

"In this experiment, you will take the role of an advisor to a top-level insurance company that specialises in home insurance pricing based on hurricane time-series data. As part of the induction process, you will be shown 13 hurricane time series, corresponding to real data from the Atlantic coast area. The time series represent annual numbers of hurricanes hitting the specific regions. Each time series contains 30 years of historical data for you to gain some knowledge of the time series' characteristics. Your task is to produce forecasts for the next 5 years. To indicate your forecasts of hurricane numbers, click at the punctuated lines at the end of the graph. A dot will appear where you forecast. Further instructions will be provided at the top of the screen at each stage to prompt you for any actions required."

The experiment was coded in Javascript and performed as an online task. Each of the time series was displayed individually. The participants' task was to indicate their judgmental forecasts on the

hurricane occurrences for the next five years on the five dotted lines presented at the end of each series. Once the five judgments had been made, participants clicked the "continue" button to proceed to the next trial. Each participant made five predictions on 13 trials and so produced a total of 65 forecasts. After completing the task, a question was displayed that asked participants about the strategy that they used to make their predictions; they typed their answers in a textbox.

For participants in the continuous "Lines" group, time series were presented as line graphs and, as forecasts were made, a blue line linked each new forecast with the last data point (forecast for horizon 1) or with the immediately preceding forecast (remaining forecasts). For participants in the discrete "Points" group, time series were presented as disconnected points and, as forecasts were made, no connection linked forecasts with the previous points.

Results

To test H_1, we extracted the Mean Absolute Distance (MAD) of forecasts from the last displayed point and then compared the size of this measure in Group 1 (continuous format) and in Group 2 (discrete format). For the first horizon forecast, we took the difference between the forecast and the last data point. For later horizons, we took the difference between the forecast for step t + 1 and the forecast for step t (i.e. the anchor). MAD scores for the five horizons in each of the 13 trials are shown in Table 1 for the two display conditions.

Forecasts from the hurricane time series producing the lowest error lie on its mean value because the series contained no significant trends or autocorrelation. Hence, to test H_2, we measured forecast accuracy by extracting the Absolute Difference from the Mean (ADFM) and compared the value of this measure in the two conditions to determine whether it was smaller in the group that saw the discrete graphical format. ADFM scores for the five horizons in each of the 13 trials are shown in Table 2 for the two display conditions.

_____

Tables 1 and 2 about here

_____

Mean absolute distance scores Data were analysed with a three-way mixed model ANOVA, using display condition (continuous versus discrete) as a between-participants factor and forecast horizon (1 – 5) and trial (13 sets of forecasts) as within-participant factors. Here and later, we adjust degrees of freedom in our analysis of variance (ANOVA) according to the recommendations of Greenhouse and Geisser (1959) when Mauchly's test indicated violation of sphericity.

This analysis revealed a main effect of forecast horizon ($F(4, 232) = 26.64$; $p < .001$): MAD scores decreased from the first to the second horizon. Furthermore, an interaction between forecast horizon and display condition ($F(4, 232) = 3.06$; $p = .017$) showed that this decrease in MAD was greater and started from a higher initial value in the continuous display condition. The simple effect of display was significant only at the first forecast horizon ($F(1, 58) = 10.27$; $p = .002$). This indicates that anchoring was greater with the continuous format and is consistent with H_1. Figure 2 shows these effects: it depicts the MAD scores for the five horizons in the two display conditions. Also shown for comparison are MAD scores derived from forecasts obtained via exponential smoothing, the statistical forecasting approach most favoured by practitioners (e.g., Weller and Crone, 2012). (The α parameter used to produce these forecasts was the one that minimised the absolute error in the forecasts.)

The analysis also showed a main effect of trial ($F(12, 696) = 16.11$; $p < .001$) and an interaction between that variable and horizon ($F(48, 2784) = 6.48$; $p < .001$). These effects arose because effects of anchoring varied systematically across trials only for forecasts for the first horizon (Table 1).

_____

Figures 2 and 3 about here

_____

Absolute difference from the mean Forecasting performance was better when participants saw data in the discrete data format. In an ANOVA using the same factors as before, the overall effect of display condition was marginally significant in a two-tailed test ($F(1, 58) = 3.15$; $p = 0.85$) but significant in a one-tailed test ($p < .05$) more appropriate for our directional hypothesis. The simple effect of display was significant for the first ($F(1, 58) = 6.79$; $p = .012$) and second forecast horizons ($F(1, 58) = 4.19$; $p = .045$). Figure 3 shows these effects: it depicts the ADFM scores for the five horizons in the two display conditions. Also shown for comparison are ADFM scores derived from forecasts obtained via exponential smoothing.

The analysis also showed effects of trial ($F(6.17, 357.75) = 12.46$; $p < .001$) and an interaction between trial and horizon ($F(21.63, 1254.35) = 3.15$; $p < .001$). As with the MAD scores, this interaction arose because effects of anchoring varied systematically across trials only for forecasts for the first horizon (Table 2).

Discussion

Graphical presentation of the time series had an impact on the forecasters' performance: forecasts for the first and second horizons were significantly inferior when data were presented in the continuous format. In line with Zacks and Tversky's (1999) arguments, the discrete format served to de-emphasise the relation between successive points. As overall autocorrelation was close to zero in the hurricane series, this de-emphasis was beneficial.

The first hypothesis was partially supported: format primarily influenced forecasting for the first horizon. This implies that the effect of format identified by Zacks and Tversky (1999) resulted in the continuous display emphasizing the relation between successive points in the data series; this affected forecasts for the first horizon. However, the type of display had little effect on how people made forecasts for later horizons: this implies that it did not influence on how they perceived the relation between the last data point and the first forecast or the relations between successive

forecasts. This is consistent with Bolger and Harvey's (1993) results that indicated that a forecast for the first horizon and those for later horizons are made in different ways: forecasts for the first horizon are influenced by points in the data series whereas those for later horizons are influenced primarily by the position of the immediately preceding forecast. Here we found that the format affected how people used previous points in the data series and so its beneficial effect was specific to the first forecast (Figure 2). As a result, its effect on performance was maintained and increased only slightly for the second forecast and did not increase further over the remaining horizons. Had there been a beneficial effect of format on degree of anchoring for every forecast horizon, the relative performance advantage of that format over the continuous one would have accumulated systematically over horizons.

Zacks and Tversky (1999) suggest that only continuous formats encourage people to impose patterns on the data, even where none exist. This proposal is in line with previous findings indicating that forecasters are prone to see non-existent patterns in noisy (line display) series and emulate them in their forecast sequence (O'Connor, Remus & Griggs, 1993). If such pattern imposition accounts for the difference between formats that we obtained, it is reasonable to expect that participants would mention it in response to the final question about their forecasting strategy. Indeed, 18 out of 30 participants in the continuous condition mentioned they followed the last segment pattern while only 8 out of 30 participants mentioned following a pattern from the last segment in the discrete condition ($\chi^2 = 7.69$; $p < .01$).

Finally, we should address a methodological issue. In the experiment, the observed time series was shifted forward by one period on each trial and participants were asked to make forecasts for five horizons. We adopted this procedure to ensure that our experiment closely matched the way in which practitioners re-forecast from the same series after new data from the most recent time periods have been obtained. However, it meant that most values that were predicted by participants were forecast on five successive trials. It is possible that (some) people (sometimes)

remembered the forecast that they gave for a given year on an earlier trial and used it again (or were influenced by it) when making a forecast for the same year on a later trial. However, to make the within-participant comparisons reported above, we assumed independence of successive forecasts for the same year. It is possible that this assumption was not justified. Many other researchers into judgmental forecasting used a similar scroll-forward procedure to ours (e.g., Angus-Leppan and Fatseas, 1986; Kusev, van Schaik, Tsaneva-Atanasova, Juliusson and Chater, 2018; Lawrence, 1983) but, as far as we are aware, none tested this assumption of independence.

We took all nine years (2000-2008) that were forecast five times, once at each of the five horizons. Then, for each horizon and assuming random intercepts and slopes, we fitted a multilevel linear model (Gellman and Hill, 2007), with forecasters as the level one variable and year as the level two variable. If independence holds, there should be no correlation between the 540 residuals from models for successive horizons. However, the correlations between residuals for horizons 1 and 2, 2 and 3, 3 and 4, and 4 and 5 were 0.51 ($t(538) = 13.76$; $p < .001$), 0.41 ($t(538) = 10.42$; $p < .001$), 0.43 ($t(538) = 11.05$; $p < .001$), and 0.37 ($t(538) = 9.24$; $p < .001$), respectively. This means that the independence assumption was not justified, that the error terms for the within-participant effects in our two mixed model ANOVAs are likely to have been distorted, and that any apparent significance of those effects should be treated with caution. However, tests of our hypotheses did not depend on the significance or otherwise of these effects: they relied on comparisons between two independent groups. Our conclusions with respect to these hypotheses remain valid.

Experiment 2: Forecasting Probability Density Functions

Participants were shown hurricane time series and were asked to place bets over the range of hurricane count values for the next year. This procedure enabled participants to generate probability density functions for one-step-ahead forecasts.

Based on Zacks and Tversky's (1999) findings, we expected that participants would anchor more on the last point when they saw the data series in continuous format. Hence, their PDFs and CDFs would show a greater shift away from the empirically derived functions than when participants saw data in the discrete format. We expect these shifts to be greater when the last data point is an outlier (distant from the series mean) than when it is not (H_3).

Method

Participants Eighty university students, (59 females) participated in the experiment. Their mean age was 21 years. Forty participants were randomly allocated to each of the two groups. They were not paid for their participation.

Design A 2x2 factorial design was adopted with the presentation format (continuous versus discrete) as a between-participants variable and the proximity of the last data point to the series mean as a within-participants variable. (A last data point within one standard deviation of the mean of the empirical series was classified as close whereas one outside that range was categorized as distant.) The dependent variable was participants' one-step-ahead probability density forecasts obtained by measuring the spread of their bets across twenty available bins. These 20 bins allowed only integer values for hurricane counts from a minimum of one up to a maximum of 20.

Stimulus materials The experiment was a pen-and-paper task with stimuli presented in a booklet. Stimuli consisted of two hurricane time series graphs. Graphs were similar to those for 1975-2004 and 1976-2005 in Experiment 1 but with two differences. First, the years on x-axis were replaced with numbers 1-30. Second, the five vertical, punctuated lines at the end of the x-axis were replaced by a line of 20 bins, with the bin range corresponding to hurricane counts. For example, bin 10 from bottom corresponded to 10 hurricane occurrences.

Data were presented as continuous line graphs in one condition and as discrete point graphs in the other. These two different displays are shown in Figure 4. Upper panels represent the pre-2005 series (close proximity) while lower panels represent the post-2005 series (distant proximity).

_____

Figure 4 about here

_____

These two data sets corresponding to the pre- and post-2005 exemplars (e.g. periods 1975-2004 and 1976-2005) shared similar characteristics: 29 out of 30 hurricane events were common. Only one new point (for 2005) appeared in the 1976-2005 series. Thus, any differences in bets between these two cases should be attributed to the value of the last data point in the series. This value was namely, nine hurricanes in 2004 (close to the series mean) and 15 hurricanes in 2005 (distant from the series mean).

Procedure The purpose of this experiment was to elicit density forecasts, generate probability distribution functions (PDFs), and thence cumulative distribution functions (CDFs), of judgmental forecasts.

Each participant performed the task individually in a quiet location. Participants were first given the experimental booklet and asked to write their age and gender on the first sheet of the booklet. They then turned the first sheet over and saw the first hurricane time series. Instructions for the experiment were provided as follows:

"In this experiment, you will take the role of an advisor to a top-level insurance company that specialises in home insurance pricing based on hurricane time-series data. As part of the induction process, you will be shown two hurricane time series, corresponding to real data from the Atlantic and Pacific coast areas. The time series represent annual hurricane counts hitting the specified regions. Each time series contains 30 years of historical data.

In this task you are given £100 and you should allocate those to the 20 bins appearing at the right-hand side of the given time series. Money allocation will be higher in the bins where you believe there is a greater probability for the next data point to occur and lower in bins where there is little chance for the next point to appear. To allocate your money, please enter your bets to each of the specified bins. You should allocate all £100. (If we played this for real, you would receive the money in the bin corresponding to the actual outcome)."

Thus, participants were endowed with a virtual sum of £100 and asked to allocate the whole amount to the 20 bins at the end of the time series. Both time series were presented either as continuous lines or as discrete unconnected points. To the right of historical data, a scale of 20 bins, ranging from 1 to 20 hurricanes, enabled participants to allocate their bets for the next year. Their money allocation (i.e. bet) had to be higher for a bin when they perceived the probability for the occurrence of number of hurricanes specified by that bin to be higher, and lower for a bin when they perceived the chance of the occurrence of the number of hurricanes specified by that bin to be lower. Once participants had read the instructions, they had the opportunity of asking for further clarification of the task requirements. After completing the task for one time series, they proceeded to the second one. Upon completion of both graphs, they were debriefed and thanked. The experiment took approximately 10 minutes to complete.

Results

For both the 1975-2004 (pre-2005) and the 1976-2005 (post-2005) series, bets were aggregated across participants to obtain the average bets assigned to each of the 20 bins. The probability distribution functions (PDF), and the cumulative distribution functions (CDF) of the aggregated bets across the 20 bins were then constructed for each of the two exemplars in each condition.

Empirical distribution functions of bets were also created based on the time series of the hurricane occurrences given to participants. This was achieved by simply counting the number of hurricane occurrences over the two periods (i.e. 1975-2004 and 1976-2005) and then assigning the

corresponding proportion of the endowed sum to bets to each of the 20 bins. For example, if six hurricanes occurred on three of the 30 years, there was a 10% chance of six hurricanes and so 10% of the £100 was assigned to the bin corresponding to six hurricanes. These empirical curves represented the best information available to participants for guiding their distribution of bets across the bins. The two curves for the pre-2005 and post-2005 series were very similar because they contained 29 out of 30 hurricane events in common.

_____

Figures 5 and 6 about here

_____

Continuous presentation format The PDF and CDF of the aggregated results, together with the corresponding empirical data, are shown in Figures 5 and 6, respectively. The shift of the pre-2005 functions to the right of the empirical ones indicates that the mean of the participants' bets was somewhat too high. Given that forecasters anchored on the last data point, this was to be expected because that last data point in the pre-2005 series was well above the series mean. The shift of the post-2005 functions even further to the right reinforces this interpretation because the last data point for that series was an outlier that was well above the series mean.

_____

Figures 7 and 8 about here

_____

Discrete presentation format PDFs and CDFs of the aggregated results, together with the empirically derived functions, are shown in Figures 7 and 8, respectively. The curves for both pre-2005 and post-2005 series are shifted to the right of the empirically derived functions. However, the degree of shift is the same for the two series. This implies that the shift away from the empirically derived curves does not reflect an anchoring phenomenon (anchoring would produce a greater shift for the post-2005 series). This implies that the rightward shift of both experimental curves arises for another reason.

The elicited distributions appear to have a higher variance and to be less skewed than the empirical ones. These differences are likely to have arisen because the upward shift in the mean of the elicited distributions meant that the left tail of those distributions was not influenced by an end-point effect (i.e., number of hurricanes could not be less than zero). This, in turn, implies that people have a tendency to produce symmetrical distributions when free of constraints imposed by end-points.

Participants' forecasts for the number of hurricanes were systematically too high. One possible reason for systematic over-forecasting is that the scenario led participants to assume asymmetric pay-offs. They were told that they were to assume that they were working for an insurance company: as a result, they may have assumed that that under-forecasting would cause the firm to lose money whereas over-forecasting would provide the firm with excess profits at the expense of householders who would have to pay higher premiums. Another possibility is that participants put their largest bets close to the mean (rather than the mode) of the empirical distribution and then imposed symmetry on their distribution of bets (as suggested in the previous paragraph).

The absence of a difference between the pre-2005 series and the post-2005 series with the discrete format but the presence of such a difference with the continuous format is consistent with H_3. It indicates that presenting the data series using a discrete graphical format serves to de-emphasise the relation between successive points and, hence, reduces anchoring effects that are found when a continuous graphical format is used to present the data series.

To confirm these results, we carried out two analyses. First, we averaged the value of the bets that each participant allocated to bins with hurricane occurrence numbers 10 - 14 (i.e. extreme hurricane activity range, greater than one standard deviation from the mean). Bets were averaged separately for the pre-2005 series (1975-2004) and post-2005 series (1976-2005) in both continuous and discrete displays conditions. We then ran a mixed model ANOVA on these data using display condition (continuous versus discrete) as a between-participants factor and series (1975-2004 versus 1976-2005) as a within-participants factor. This revealed a main effects of display condition ($F(1, 78)$

= 10.84; p < .001) and series (F (1, 78) = 19.35; p < .001), together with an interaction between these factors (F (1, 78) = 9.28; p = .003). This interaction occurred because the value of the bets in bins 10 - 14 was similar for both series with the discrete display and for the 1975-2004 series with the continuous display; however, it was very much higher for the 1976-2005 series with continuous display (Figures 5 and 7). Independent samples t-tests showed no significant difference between the two display conditions for the 1975-2004 series (t (78) = 1.90; p = .17) but did show one for the 1976-2005 series (t (78) = 10.48; p = .002).

Second, we averaged the value of the bets that each participant allocated to bins with hurricane occurrence numbers 5 – 9 (i.e., average hurricane activity range within one standard deviation from the mean). Given that there was greater betting on extreme bins in the continuous display condition with the 1976-2005 series (compared with the other display and series), we would expect *less* betting on average bins with that display and series. An ANOVA with the same factors as before confirmed this. It again revealed main effects of display condition (F (1, 78) = 4.18; p = .044) and series (F (1, 78) = 40.84; p < .001), together with an interaction between these factors (F (1, 78) = 26.85; p < .001). This interaction arose because the value of the bets in bins 5 - 9 was similar for both series with the discrete display and for the 1975-2004 series with the continuous display; however, it was very much *less* for the 1976-2005 series with continuous display (Figures 5 and 7). Independent samples t-tests showed no significant difference between the two display conditions for the 1975-2004 series (t (78) = 0.25; p = .62) but did show one for the 1976-2005 series (t (78) = 6.86; p = .01).

The results of these analyses reinforce the interpretation that we provided above. Consistent with H_3, anchoring effects are reduced by using a discrete presentation format for data series.

Discussion

Participants showed significantly greater anchoring on extreme values of the last data point when series were presented using a continuous graphical format than when they were presented using a

discrete graphical format. This result serves to validate the conclusions of the first experiment within the context of a completely different forecasting task.

The fact that density forecasts are strongly affected by display format, especially when recent data points are more than one standard deviation from the series mean, has implications for real-world hurricane forecasting where probability density forecasting is often used.

Experiment 3: Forecasting using prediction intervals

Prediction intervals are important in hurricane forecasting by practitioners[2]. These intervals specify upper and lower forecast boundaries within which the future value of the predicted variable is expected to lie with a specific probability. In hurricane forecasting, this probability is usually set at 70% but in other applications, such as demand forecasting, it is often set at 90% or 95%. Prediction intervals are known to be too narrow (Lawrence & Makridakis, 1989; Lawrence & O'Connor, 1993; O'Connor & Lawrence, 1989, 1992), suggesting overconfidence. It is likely that this phenomenon arises because participants anchor on the last data point and then adjust away from it in each direction to produce the required interval (Harvey, 1997). Intervals are too narrow because adjustment is typically insufficient (Tversky & Kahneman, 1974).

Method

Participants were presented with the same historical hurricane time series data that were used in Experiment 1 but, in this experiment, they were requested to provide 70% prediction interval forecasts for the next five years. Based on Zacks and Tversky's (1999) findings and following the results obtained in Experiments 1 and 2, we expected participants would be more overconfident in the continuous display condition (H_4). This is because, in that condition, greater anchoring on the last data point to produce prediction intervals would produce less adjustment away from that point and hence result in narrower intervals.

Participants Sixty students (40 females) from University College London acted as participants. Their mean age was 20 years. They were not paid for their participation.

Design Participants were randomly allocated to two groups, with the constraint that there were 30 participants in each group. The first group (continuous representation) produced prediction intervals from continuous line graphs while the second group (discrete representation) made their predictions from unconnected point graphs.

Stimulus materials The time series that were used were the same as those in Experiment 1. At the end of the x-axis of each one five vertical, punctuated lines were displayed at horizontal positions representing the next five years in the series. Participants marked their 70% prediction interval forecasts on these lines.

Procedure Participants performed the task individually on computers. They read a short introduction and then entered their demographic details (age, sex). Instructions were the same as in Experiment 1 except that, this time, instead of point forecasts, 70% prediction intervals were required. Thus, acting as insurance advisors, participants were requested to provide 70% prediction intervals of hurricane counts for the next five years based on 30 years of historical data. It was explained to them that 70% prediction intervals meant that each future observation would fall into the corresponding forecasted interval with a 70% probability. The prediction intervals were marked by clicking twice on each of the five punctuated lines at the end of the graph to indicate the interval's upper and lower boundaries. After completing the forecasts for all 13 data series, participants were debriefed and thanked.

Results

We compared the mean width of prediction intervals across the two conditions. The size of the intervals was calculated by taking the difference between the upper and lower values of participants' responses.

According to H_4, participants show more overconfidence (i.e. narrower prediction intervals) when given the continuous display than when given the discrete one. The data were consistent with this for all horizons (Figure 9).

We calculated the actual size of the 70% prediction intervals for each horizon in each series. We then subtracted these values from the corresponding ones estimated by each participant. This difference score, averaged across horizon and series, provides a measure of the degree to which each participant tends to mis-estimate the width of prediction intervals. When it is negative, participants tend to underestimate prediction interval widths; when it is positive, they tend to overestimate them.

The mean value of the difference score in the continuous display group was -1.25 with a standard deviation of 0.98. This mean value was significantly different from zero ($t(29) = 7.02$; $p < .001$), thereby indicating significant underestimation of prediction interval widths in this group. The mean value of the difference score in the discrete display group was -0.12 with a standard deviation of 1.74. This mean value was not significantly different from zero ($t(29) = 0.37$; $p = .71$) and so there was no evidence of mis-estimation of prediction interval widths in this group. An independent samples t-test (not assuming equality of variances) showed that the difference scores in the two display conditions were significantly different from one another ($t(45.56) = 3.11$; $p = .003$).

Figure 9 shows the degree of underestimation of the prediction interval widths in the two display conditions over each of the five horizons. A two-way mixed model ANOVA with display condition as a between-participants factor and forecast horizon as a within-participants factor on these measures of underestimation of prediction intervals revealed only a main effect of display condition ($F(1, 58) = 9.13$; $p = .004$).

_____

Figure 9 about here

_____

Discussion

Prediction intervals were narrower than the empirically derived ones with the continuous format but not with the discrete format. The difference in performance with different display formats can again be explained in terms excessive anchoring and insufficient adjustment in the continuous format condition (Harvey, 1997) and amelioration of these problems by use of the discrete display.

Our findings replicate previous results obtained with continuous display formats (Lawrence & Makridakis, 1989; Lawrence & O'Connor, 1993; O'Connor & Lawrence, 1989, 1992). Prediction intervals were too narrow. In the past, this has been taken as evidence that people are overconfident in their forecasts. However, simply by presenting data series in a discrete format, we can ensure that forecasters' judged intervals are well-calibrated relative to empirically derived intervals. It seems unlikely that this change in format acts to reduce people's confidence in their forecasts. It is more likely that, consistent with Zacks and Tversky (1999), it acts to de-emphasise the relation between successive points in the series and so reduces excessive anchoring.

General Discussion

Human judgments contribute a great deal to the accuracy of forecasting but they are sometimes subject to certain systematic errors. Using uncorrelated and un-trended real hurricane time series, the objective of the present study was to investigate judgmental biases in point forecasts (Experiment 1), density functions (Experiment 2) and prediction intervals (Experiment 3), and to study whether they can be ameliorated by changing the graphical format used to present the data series.

Biases in judgmental forecasting

Lawrence and O'Connor (1995) found that the sort of under-adjustment to be expected if judges use anchoring was not evident when judgmental forecasts were made from many real series (Makridakis et al, 1982). However, Reimers and Harvey (2011) argued that this does not mean that forecasts

from real series are not subject to biases. Instead it indicates that people are well-adapted to series that are broadly representative of their real-world environment. Moderate degrees of positive autocorrelation are typical of our environment (Gilden, 2009) and when people forecast from such series, they are unbiased. However, not all real series are typical. Some show higher levels of autocorrelation: they are forecast in a biased way that suggests that people perceive their autocorrelation as lower than it is. Other real series, such as the hurricane series used here, show very little autocorrelation[3]: they are forecast in a biased way that implies that people perceive their autocorrelation as higher than it really is. However, when we average over a whole set of real series with many different levels of autocorrelation, biases in different directions largely cancel each other out.

Thus, the anchoring effects that we have demonstrated with real series in our experiments are important. They show that the previous research with simulated series that has been used to argue that judgmental forecasts are biased is indeed relevant to forecasting from real series. Biases appear with real series when those series are not typical of the series that people encounter in their environment. For example, some real series may contain atypically high or atypically low levels of autocorrelation: we can expect judgmental forecasting from those series to be biased. In other words, it is possible to be broadly well-adapted to series encountered across the environment as a whole, but to still show some systematic biases when dealing with particular series.

Why do biases occur with series that have atypical levels of autocorrelation? People exposed to many series in the environment will gain some impression of the overall level of autocorrelation that they contain. When they encounter a new series, this average environmental autocorrelation can be regarded as an initial estimate for the autocorrelation in the new series. By processing the patterns in that series, they make an adjustment away from their initial estimate. However, because the data series is limited in length and noisy, their adjustment is only partial. Because it is only partial, the residual influence of the environmental autocorrelation still has some effect and this effect is what

we label as a bias. Consistent with this account, biases are larger in noisier data (Harvey and Reimers, 2013; Reimers and Harvey, 2011). However, as this account makes clear, biases are not to be regarded as signs that judgment is irrational: they can be produced by a process that can be characterised as close to a Bayesian one.

Reducing forecasting biases

We know that various factors can influence the degree of bias that people exhibit. For example, Reimers and Harvey (2011) argued that people are constantly updating their estimates of the level of autocorrelation that is typical of their environment. They first presented people either with many series with low levels of autocorrelation or many series with high levels of autocorrelation. Then they required people to make forecasts from target series with moderate levels of autocorrelation. People who had previously seen many series with low levels of autocorrelation produced forecasts that indicated that they perceived a lower autocorrelation in the target series than people who had previously seen many series with high levels of autocorrelation.

Thus, the degree of autocorrelation that people perceive in a given series is labile. It can be influenced by previous experience. The three experiments reported here demonstrate that it is also influenced by the way series are presented. Lines linking successive points serve to imply that there is a relation between those points that is inconsistent with their independence. To improve judgmental forecasting from independent points, we should present data series as unconnected points. Conversely, we would expect (though we have not shown it) that forecasting from points that are strongly sequentially dependent would be improved by presenting data series as line graphs rather than as unconnected points.

Many studies have shown that judgmental prediction intervals are too narrow. This can be explained in terms of anchoring: people anchor on the last data point and adjust away from it in both directions to produce the upper and lower bounds of the interval. Again, it appears that the degree

to which they are 'attracted' to the last data point is influenced by the graphical format in which the data series are presented. Line graphs emphasise connections (even when they are not logically or statistically present) between successive points, between the last data point and the first forecast (Experiment 1), and, apparently, between the last data point and the bounds of a prediction interval (Experiment 3). Simply by changing the data presentation format from continuous to discrete, it is possible to eliminate this effect and thereby enable people to produce well-calibrated intervals.

Limitations and future work

First, it remains unclear whether the advantage of discrete graphs for forecasting purposes extends to other domains where series show higher autocorrelation. There are reasons to suspect that they will not. With series showing a very high autocorrelation, the autocorrelation that people perceive (as implied by their forecasts) has been found to be less than it should be (Reimers & Harvey, 2011). For such series, continuous graphical formats that emphasise the relation between successive points are likely to produce better performance than discrete ones. Hence, future experiments should test real time series that have high levels of autocorrelation.

We also suspect that series with trends may not show the same advantage of discrete over continuous presentation format. Trends also depend on a relation between successive points and continuous presentation formats may serve to emphasise that relation. Harvey and Bolger (1996) have already shown that graphical presentation (via line graphs) reduces trend damping relative to tabular presentation (where, presumably, the relation between successive points is less salient).

We have studied only one element of the hurricane forecasting process. In future work, it would be useful to study how model-based forecasts are integrated with judgmental forecasts. In particular, is the weighting given to model-based forecasts influenced by the data format? Also, how is the integration process influenced by presenting model-based forecasts not just for the future horizons that require forecasts but also for past time points for which the outcomes are known and

displayed? With such a display, is the integration affected not just by the format in which the data are presented but also by the format in which past model forecasts are presented?

Our participants were not paid. Camerer and Hogarth's (1999) reviewed 74 studies that manipulated incentives: they were either absent, low, or high. Of the 59 studies in which performance accuracy could be assessed, incentives had no effect in 27 of them, facilitated performance in 23 of them, and impaired performance in nine of them. For accuracy to be improved by incentives, the task had to be one in which increased mental effort would improve performance. Tasks in this category typically involve memory encoding and recall (Kahneman and Peavler, 1969); tasks that involve pattern perception are not included in it. Hence, it is unlikely that addition of incentives would have influenced our results.

References

Angus-Leppan, P. & Fatseas, V. (1986). The forecasting accuracy of trainee accounts using judgmental and statistical techniques. *Accounting and Business Research, 16*, 179-188.

Bernard, V. & Thomas, J. (1990). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings*. Journal of Accounting and Economics, 13*, 305-340.

Bolger, F. & Harvey, N. (1993). Context-sensitive heuristics in statistical reasoning. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 46,* 779-811.

Coll, R., A. Thyagarajan & Chopra, S. (1991). An experimental study comparing the effectiveness of computer graphics data versus computer tabular data. *IEEE Transactions on Systems, Man and Cybernetics, 21,* 897-900.

Dickson, G.W., DeSanctis, G. & McBride, D.J. (1986). Understanding the effectiveness of computer graphics for decision support: A cumulative experimental approach. *Communications of the ACM, 29,* 40-47

Eggleton, I. R. C. (1982). Intuitive time-series extrapolation. *Journal of Accounting Research, 20,* 68-102.

Fildes, R., & Goodwin, P. (2007). Good and bad judgment in forecasting: Lessons from four companies. *Foresight: The International Journal of Applied Forecasting, Fall 2007*, 5-10.

Fildes, R., & Petropoulos, F. (2015). Improving forecast quality in practice. *Foresight: The International Journal of Applied Forecasting, Winter 2015,* 5-12.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge: Cambridge University Press.

Gilden, D. L. (2009). Global model analysis of cognitive variability. *Cognitive Science, 33,* 1441–1467.

Goodwin, P., & Wright, G. (1993). Improving judgmental time series forecasting: A review of the

guidance provided by research. *International Journal of Forecasting, 9*, 147-161.

Goodwin, P., & Wright, G. (1994). Heuristics, biases and improvement strategies in judgmental time-series forecasting. *Omega: International Journal of Management Science, 22,* 553-568.

Greenhouse, S. W. & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 24*, 95-112.

Harvey, N. (1995). Why are judgments less consistent in less predictable task situations? *Organizational Behavior and Human Decision Processes, 63*, 247 – 263.

Harvey (1997). Use of heuristics: Insights from forecasting research. *Thinking and Reasoning, 13,* 5-24.

Harvey, N., & Bolger, F. (1996). Graphs versus tables: Effects of data presentation format on judgmental forecasting. *International Journal of Forecasting, 12,* 119-137.

Harvey, N., & Reimers, S. (2012). Bars, lines, and points: the effect of graph format on judgmental forecasting. *Proceedings of the 32nd Annual International Symposium on Forecasting*, Boston, MA, 2011, IIF.

Harvey, N., & Reimers, S. (2013). Trend damping: Under-adjustment, experimental artifact, or adaptation to features of the natural environment? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 589-607.

Kahneman, D., & Peavler, W. S. (1969). Incentive effects and pupillary changes in association learning. *Journal of Experimental Psychology, 79,* 312–318.

Kusev, P., van Schaik, P., Tsaneva-Atanasova, K., Juliusson, A. & Chater, N. (2018). Adaptive anchoring model: How static and dynamic presentations of time series influence judgments and predictions. *Cognitive Science, 42,* 77-102.

Lawrence, M. J. (1983). An exploration of some practical issues in the use of quantitative forecasting models, *Journal of Forecasting, 2,* 169-179.

Lawrence, M., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence-intervals. *Organizational Behavior and Human Decision Processes, 43,* 172-187.

Lawrence, M., & O'Connor, M. (1993). Scale, variability and the calibration of judgmental prediction intervals. *Organizational Behavior and Human Decision Processes, 56,* 441-458.

Lawrence, M., & O'Connor, M. (1995). The anchoring and adjustment heuristic in time series forecasting. *Journal of Forecasting, 14,* 443– 451.

Lawrence, M., Edmundson, R., & O'Connor, M. (1985). An examination of the accuracy of judgmental extrapolation of time series. *International Journal of Forecasting, 1*, 25-35.

Lawrence, M., Goodwin, P., O'Connor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22,* 493-518.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., & Winkler, R. (1982). The accuracy of extrapolation (time-series) methods: Results of a forecasting competition. *Journal of Forecasting, 1,* 111-153.

Newman, G. E. & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar effect bias. *Psychonomic Bulletin and Review, 19,* 601-607.

O'Connor, M., & Lawrence, M. (1989). An examination of the accuracy of judgmental confidence intervals in time series forecasting. *Journal of Forecasting, 8,* 141 – 155.

O'Connor, M., & Lawrence, M. (1992). Time series characteristics and the widths of judgmental confidence intervals. *International Journal of Forecasting, 7*, 413 – 420.

O'Connor, M., Remus, W., & Griggs, K. (1993). Judgmental forecasting in times of change. *International Journal of Forecasting, 9*, 163-172.

Okan, Y., Garcia-Retamero, R., Cokely, E.T. & Maldonado, A. (2018). Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy. *Quarterly Journal of Experimental Psychology.* Published online 1 January 2018.

Reimers, S., & Harvey, N. (2011). Sensitivity to autocorrelation in judgmental time series forecasting. *International Journal of Forecasting, 27*, 1196-1214.

Tullis, T.S. (1988) Screen design, In M. Helander, (Ed.), *Human-computer Interaction* (North-Holland, Amsterdam), 367-411.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, I85, 1127-1131.

Webby, R. & O'Connor, M. (1996). Judgemental and statistical time series forecasting: A review of the literature. *International Journal of Forecasting, 12,* 91-118.

Weller, M., & Crone, S. F. (2012). Supply chain forecasting: Best practices & benchmarking study. Lancaster: Lancaster University. http://www.lancaster.ac.uk/ lums/forecasting/material/

Zacks, J., & Tversky, B. (1999). Bars and lines: A study of graphic communication. *Memory and Cognition, 27*, 1073–1079.

Endnotes

1. Hurricane forecasting is vital for ensuring that sufficient preparations and emergency procedures are in place in anticipation of hurricanes. One such preparation relates to the adjustment of pricings in the insurance and reinsurance sector. Every year, the NOAA Climate Prediction Centre provides a formal, model-derived seasonal outlook of the overall expected activity for the year's hurricane season. This information, together with historical hurricane time series data, serves as the basis for the judgmental forecasts of the number of hurricanes in future years that are made by lay people and by practitioners, such as those working in the insurance industry.

2. Within its formal, model-derived seasonal outlook, the NOAA's Climate Prediction Centre provides overall expected activity for the year's hurricane season in the form of prediction intervals. Statistical input from such formal models, along with the historic time-series data that serve as a basis for forecasting the number of hurricanes in future years, are reviewed annually by insurers. They use their judgment to integrate all available information to set insurance prices.

3. Price changes in ideal markets are independent but those in real markets do not always fit this model (e.g. Bernard and Thomas, 1990).

Table 1. Experiment 1: Mean Absolute Deviation (MAD) scores of forecasts across the five horizons in the 13 trials in both display conditions

Display	Trial	Horizon 1	Horizon 2	Horizon 3	Horizon 4	Horizon 5
Lines	1	3.66	2.15	1.79	1.49	1.73
	2	2.68	1.56	1.68	1.00	1.62
	3	1.58	1.40	1.99	1.81	2.07
	4	3.59	2.09	1.39	1.60	1.72
	5	2.17	1.85	2.11	1.59	1.49
	6	1.06	1.77	1.82	1.71	1.63
	7	1.79	1.96	1.44	2.09	2.24
	8	1.93	1.74	2.17	1.97	1.80
	9	2.15	2.25	1.55	1.64	1.91
	10	2.03	1.60	1.92	1.44	2.03
	11	4.46	2.52	2.60	2.08	2.07
	12	2.86	2.52	2.61	2.67	2.65
	13	2.24	2.33	1.93	2.78	2.43
Points	1	4.56	2.07	1.88	1.57	2.29
	2	4.10	1.64	1.60	1.54	2.21
	3	1.93	1.29	1.60	1.94	2.12
	4	3.82	1.97	2.18	1.74	2.05
	5	2.61	1.69	2.02	1.46	2.09
	6	2.08	1.89	1.68	1.98	1.72
	7	2.35	1.61	1.35	1.26	1.61
	8	2.39	1.39	2.56	1.77	1.98
	9	1.57	1.71	1.93	2.01	1.78
	10	2.44	1.92	1.63	1.87	2.06
	11	5.93	2.40	2.23	2.34	2.42
	12	3.06	1.95	2.43	2.50	2.71
	13	2.63	1.99	2.31	2.67	2.95

Table 2. Experiment 1: Absolute Distance **f**rom the Mean (ADFM) scores of forecasts across the five horizons in the 13 trials in both display conditions

Display	Trial	Horizon 1	Horizon 2	Horizon 3	Horizon 4	Horizon 5
Lines	1	3.04	2.16	2.13	1.50	1.77
	2	1.71	1.72	1.90	2.38	2.42
	3	1.31	1.34	1.82	1.91	2.06
	4	2.50	1.76	1.42	1.57	1.70
	5	1.31	1.47	1.72	2.07	2.12
	6	2.26	2.10	1.77	2.60	2.42
	7	3.00	2.57	2.18	1.88	2.59
	8	1.62	1.92	2.14	1.94	2.19
	9	2.58	2.23	1.65	1.59	1.76
	10	2.26	2.13	1.92	2.01	2.64
	11	4.96	3.15	2.80	2.82	3.23
	12	2.25	2.95	2.83	3.04	2.34
	13	2.02	2.81	2.66	2.72	2.04
Points	1	1.62	1.46	1.48	1.36	1.82
	2	1.72	1.54	1.62	1.76	1.96
	3	1.21	1.04	1.45	1.66	1.80
	4	1.62	1.44	1.46	1.92	2.06
	5	1.62	1.42	1.66	2.04	2.23
	6	1.64	1.79	2.15	2.08	1.84
	7	2.33	1.95	1.99	1.85	1.92
	8	2.01	1.86	2.15	1.73	2.10
	9	1.43	1.73	1.82	1.90	2.04
	10	1.76	1.76	1.84	1.50	1.87
	11	3.35	2.13	2.90	2.55	2.48
	12	2.13	2.53	2.35	2.55	2.02
	13	2.46	2.58	2.32	2.25	2.10

Figure legends

Figure 1: Screenshot of hurricane series presented in the continuous line graph display (upper panel) and discrete point graph display (lower panel).

Figure 2: Mean MAD scores for five forecasts in the two display conditions. Also shown are scores associated with forecasts obtained via exponential smoothing.

Figure 3: Mean ADFM scores for five forecasts in two conditions. Also shown are scores associated with forecasts obtained via exponential smoothing.

Figure 4: Hurricane series presented to the participants: continuous (left) and discrete (right) presentation formats with last data point close to (upper) and distant from (lower) the series mean.

Figure 5: PDFs derived from observed data (upper panel), from the continuous format group for pre-2005 series (centre panel), and from the continuous format group for post-2005 series (lower panel).

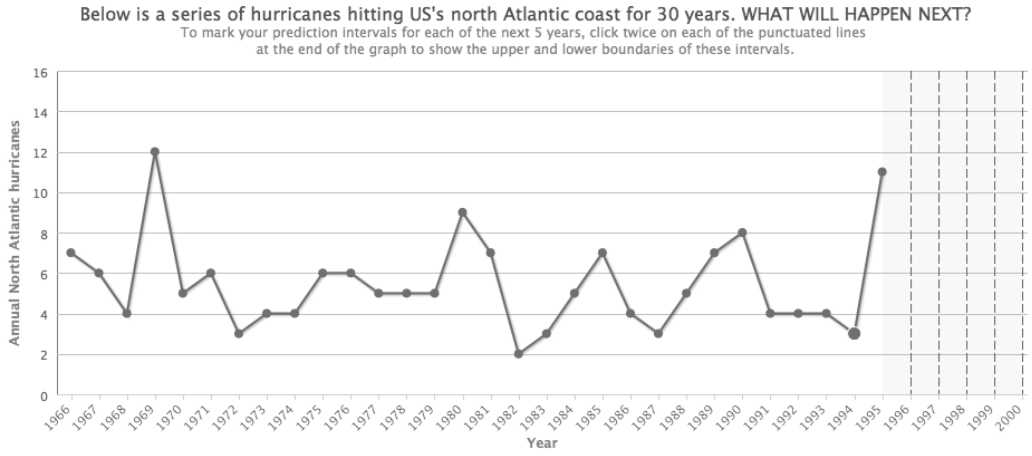
Figure 6: CDFs derived from the continuous format group and from the observed data.

Figure 7: PDFs derived from observed data (upper panel), from the discrete format group for pre-2005 series (centre panel), and from the discrete format group for post-2005 series (lower panel).

Figure 8: CDFs derived from the discrete format group and from the observed data.

Figure 9: Underestimation of prediction interval widths in the two display conditions.

Figure 1


Below is a series of hurricanes hitting US's north Atlantic coast for 30 years. WHAT WILL HAPPEN NEXT?
To mark your prediction intervals for each of the next 5 years, click twice on each of the punctuated lines
at the end of the graph to show the upper and lower boundaries of these intervals.
Annual North Atlantic hurricanes
Year

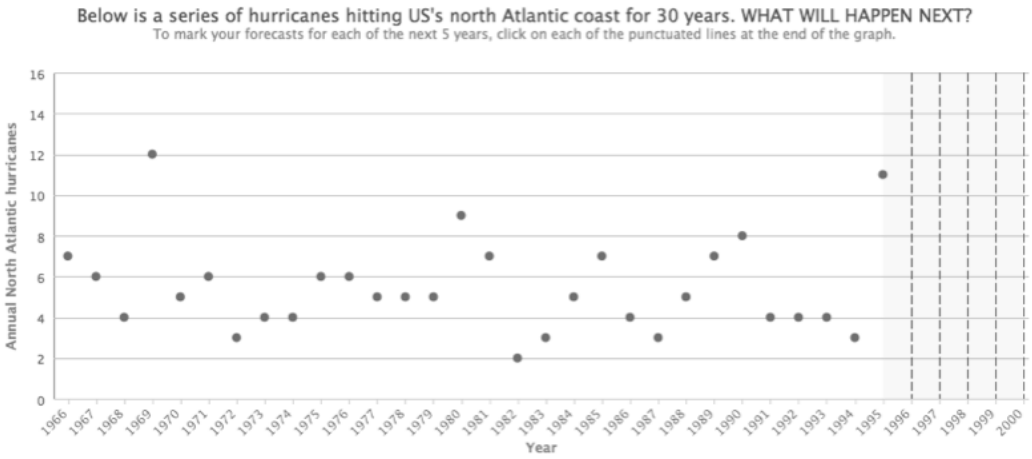

Below is a series of hurricanes hitting US's north Atlantic coast for 30 years. WHAT WILL HAPPEN NEXT?
To mark your forecasts for each of the next 5 years, click on each of the punctuated lines at the end of the graph.
Annual North Atlantic hurricanes
Year

Figure 2

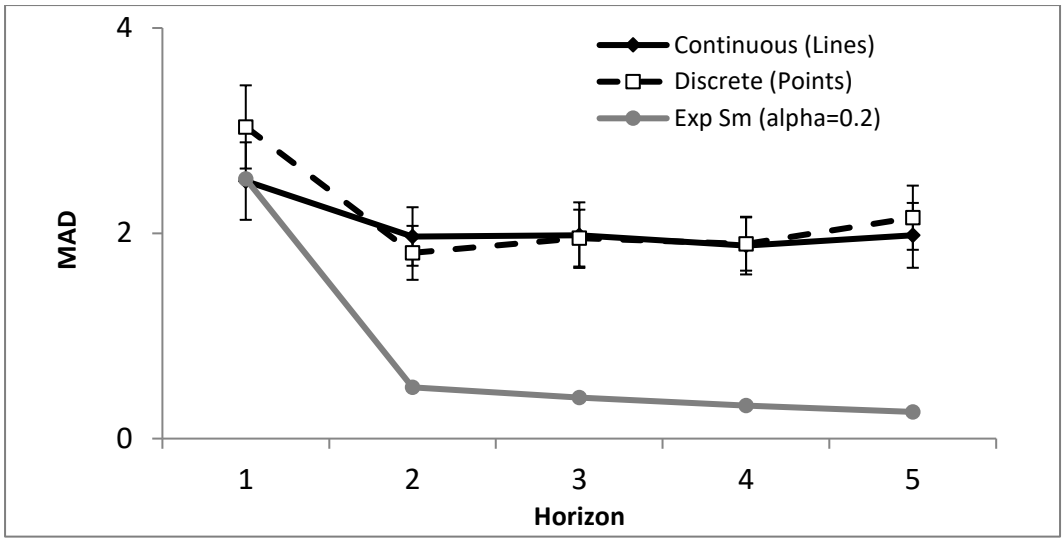

4
2
0
MAD
1
2
3
4
5
Horizon
Continuous (Lines)
Discrete (Points)
Exp Sm (alpha=0.2)

Figure 3

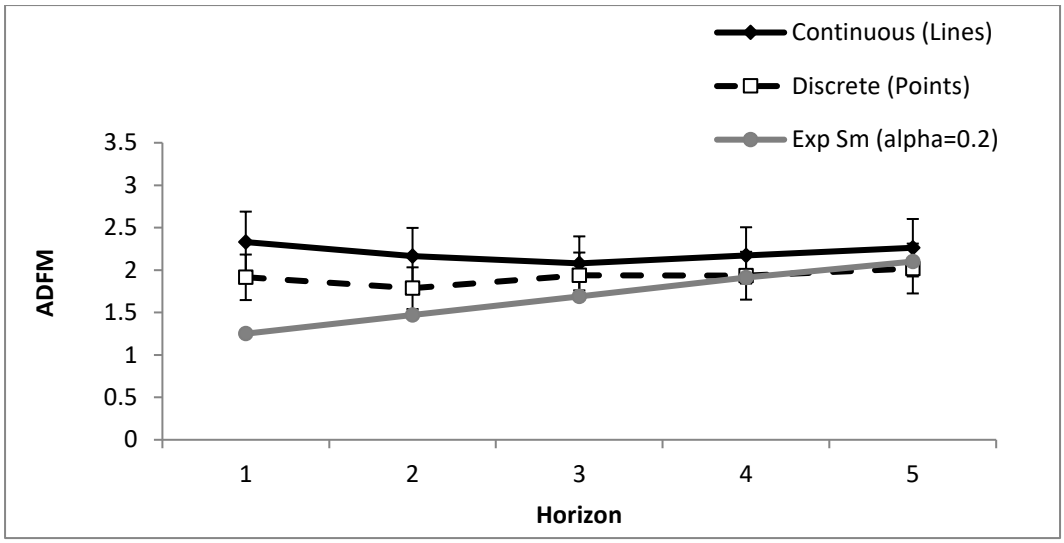

Continuous (Lines)
Discrete (Points)
Exp Sm (alpha=0.2)
ADFM
3.5
3
2.5
2
1.5
1
0.5
0
1
2
3
4
5
Horizon

Figure 4

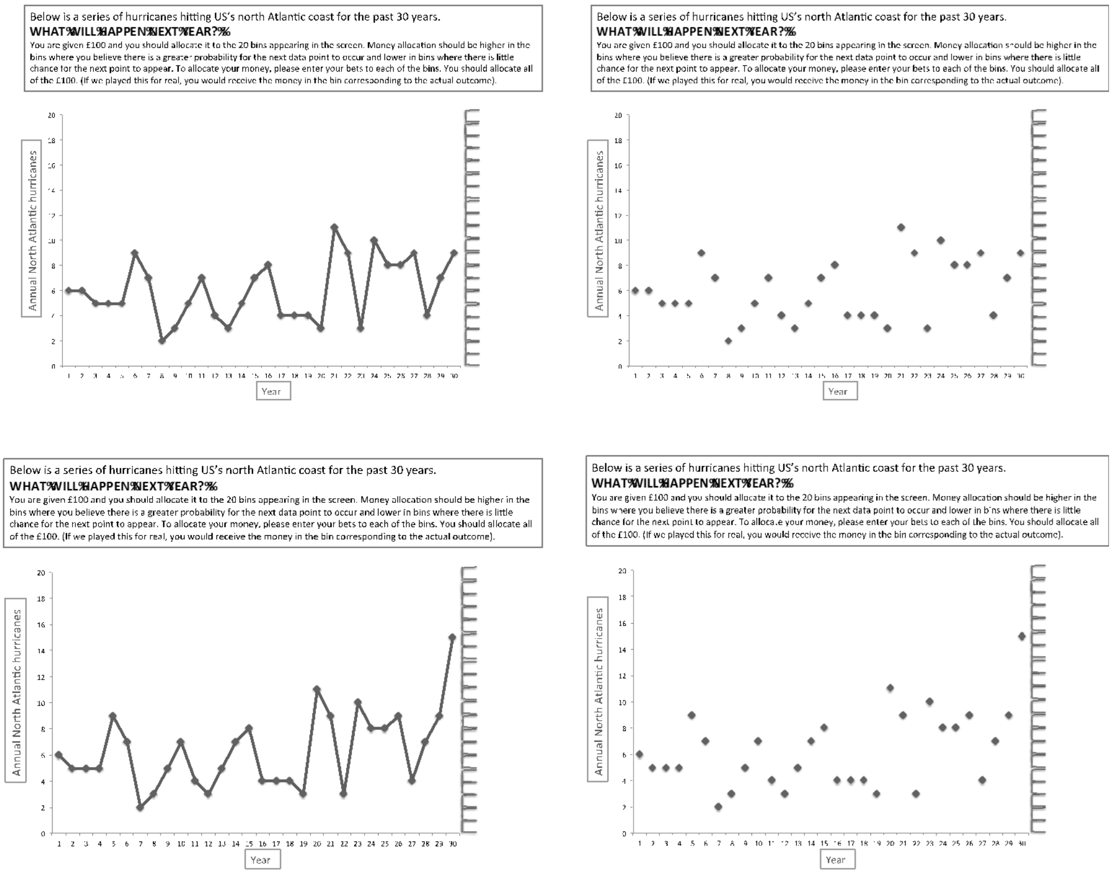
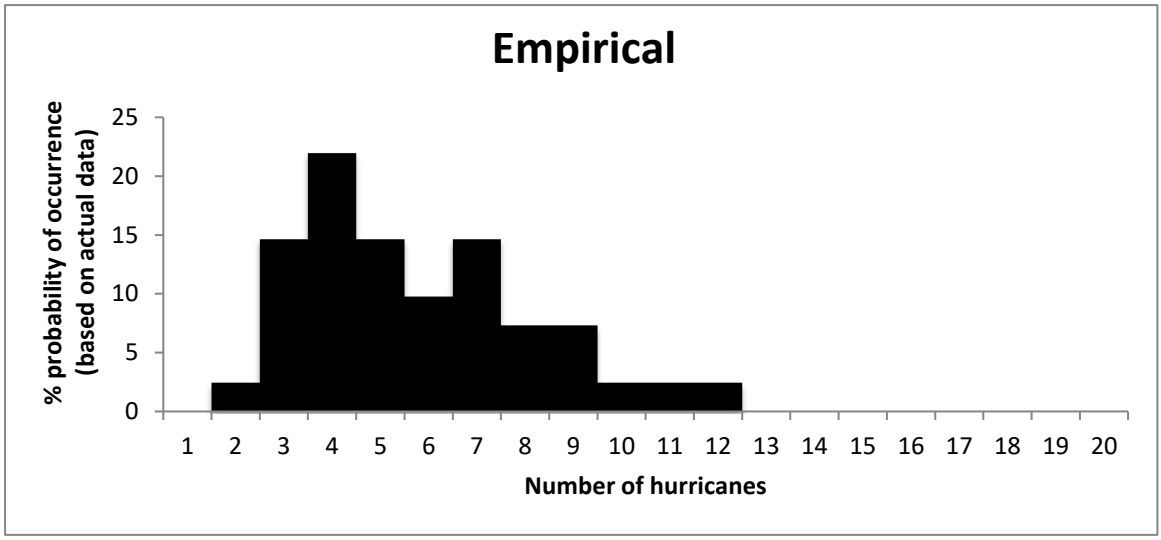
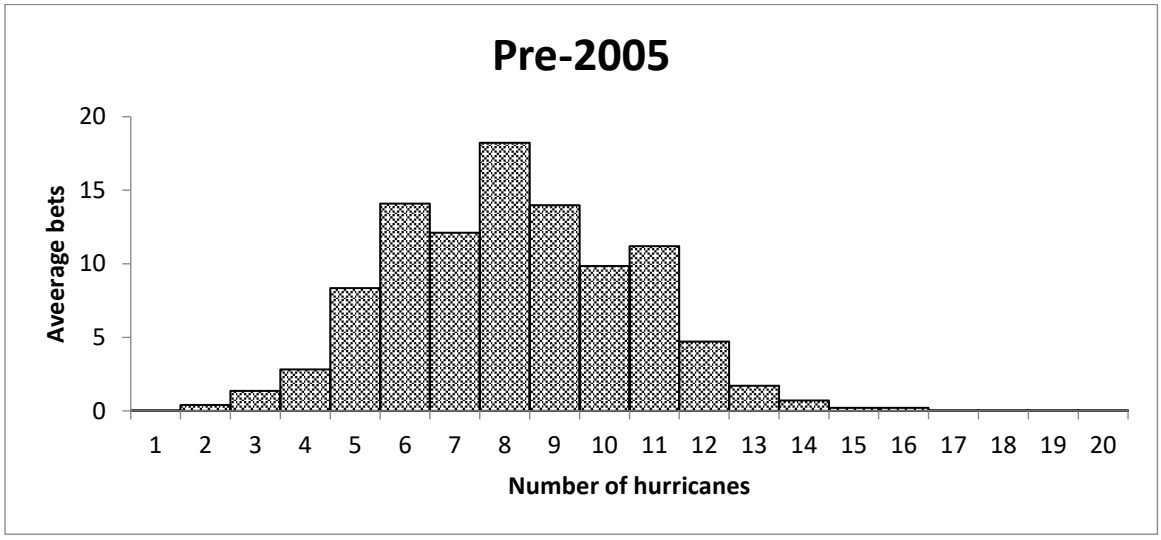

Below is a series of hurricanes hitting US's north Atlantic coast for the past 30 years.
WHAT%WILL%HAPPEN%NEXT%YEAR?%
You are given £100 and you should allocate it to the 20 bins appearing in the screen. Money allocation should be higher in the bins where you believe there is a greater probability for the next data point to occur and lower in bins where there is little chance for the next point to appear. To allocate your money, please enter your bets to each of the bins. You should allocate all of the £100. (If we played this for real, you would receive the money in the bin corresponding to the actual outcome).
Annual North Atlantic hurricanes
Year

Figure 5


Empirical
% probability of occurrence (based on actual data)
0
5
10
15
20
25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes


Pre-2005
Aveerage bets
0
5
10
15
20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes

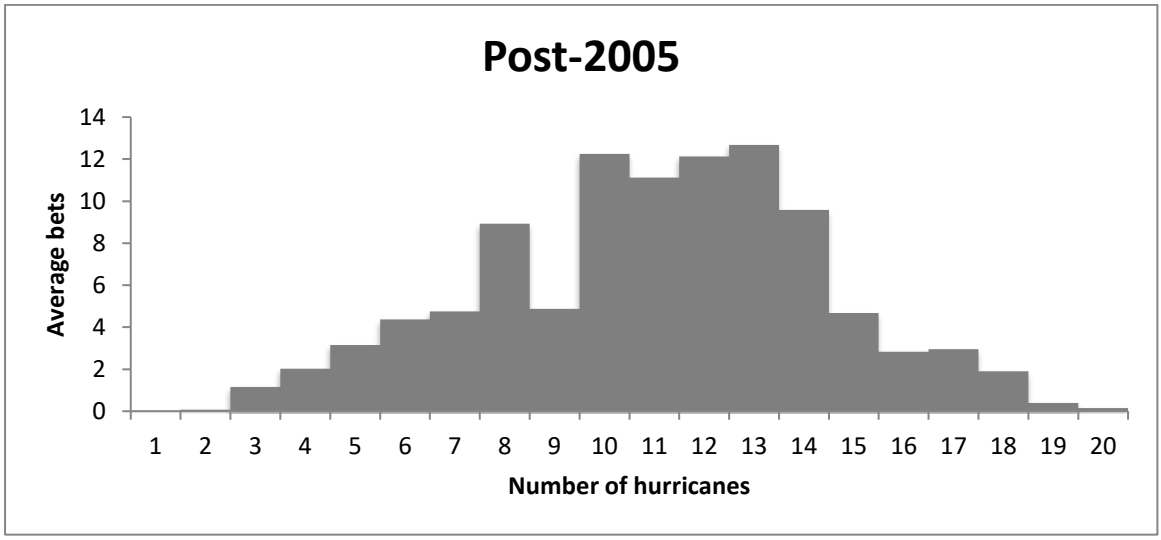

Post-2005
Average bets
0
2
4
6
8
10
12
14
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes

Figure 6

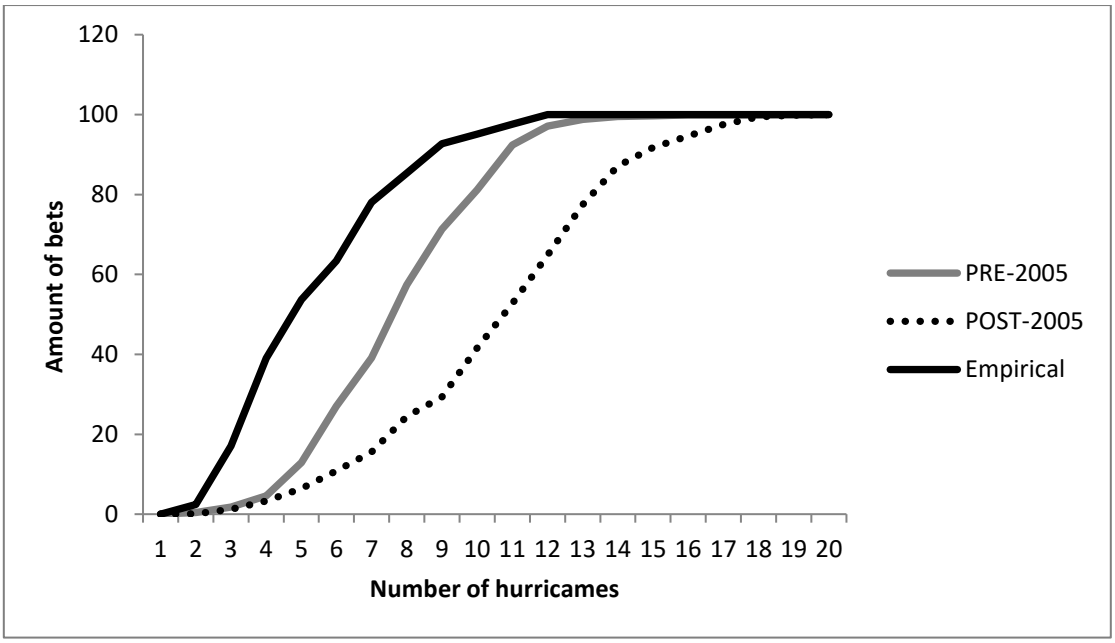
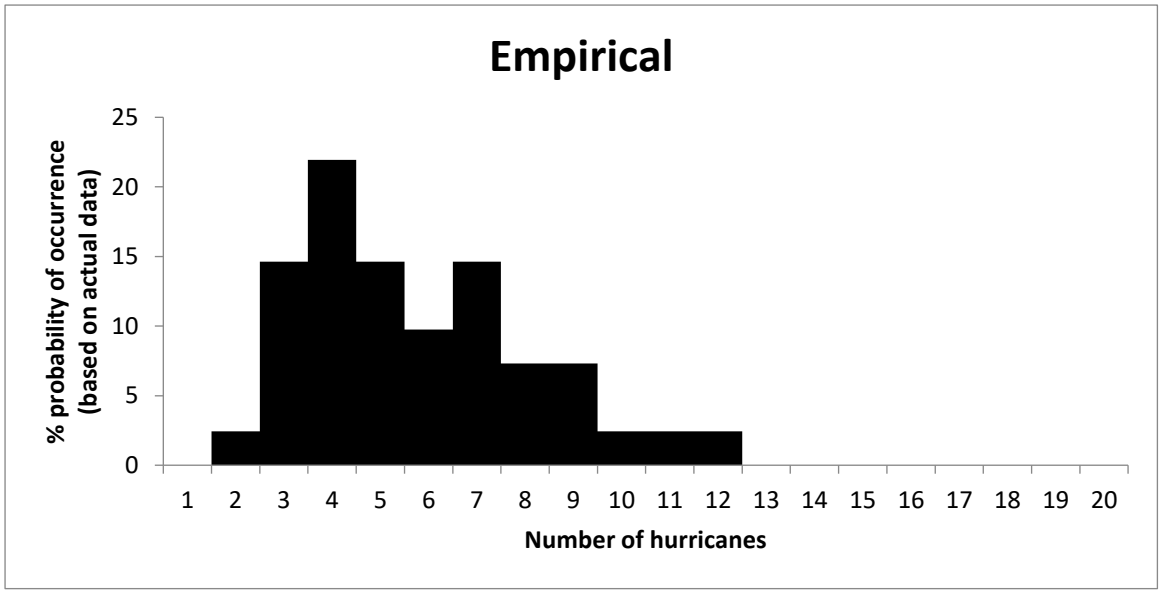
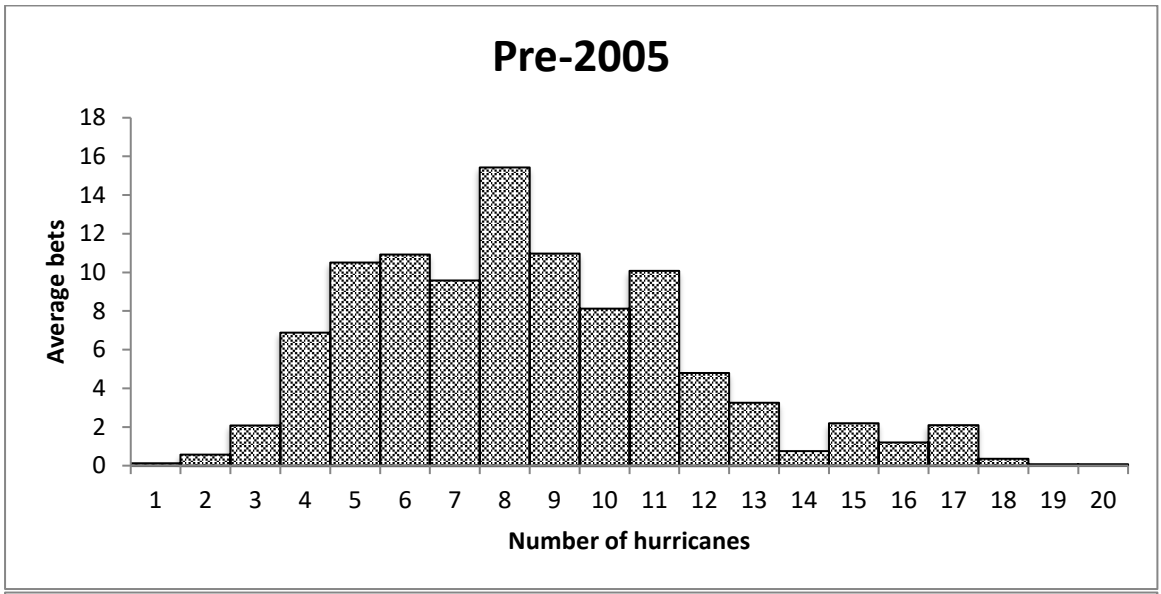

120
100
80
60
40
20
0
Amount of bets
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricames
PRE-2005
POST-2005
Empirical

Figure 7


Empirical
% probability of occurrence (based on actual data)
0
5
10
15
20
25
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes


Pre-2005
Average bets
0
2
4
6
8
10
12
14
16
18
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes

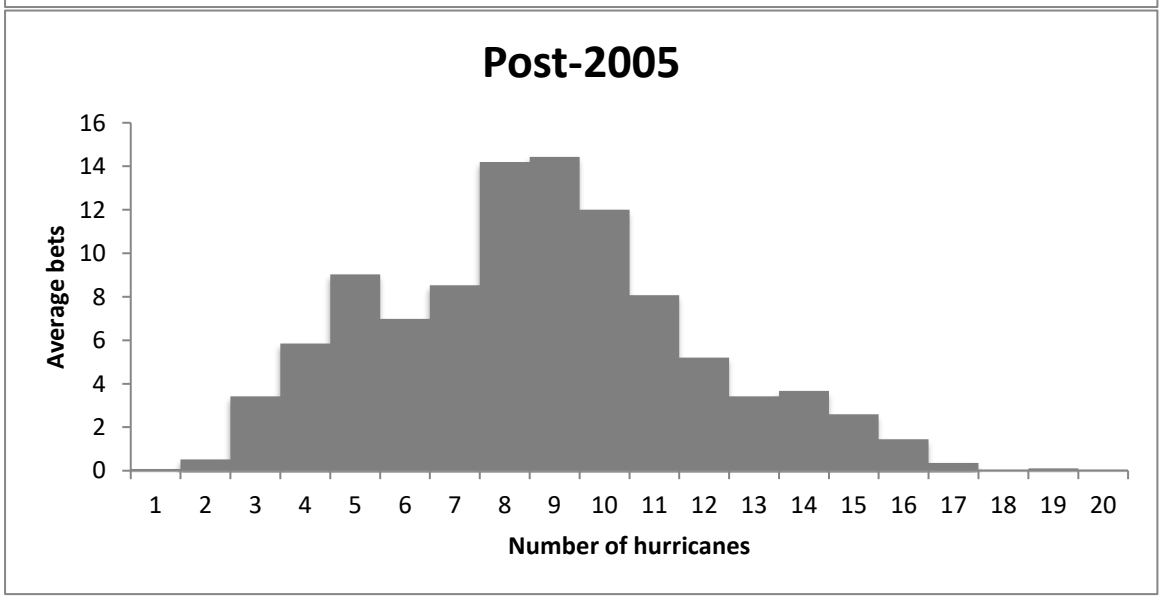

Post-2005
Average bets
0
2
4
6
8
10
12
14
16
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes

Figure 8

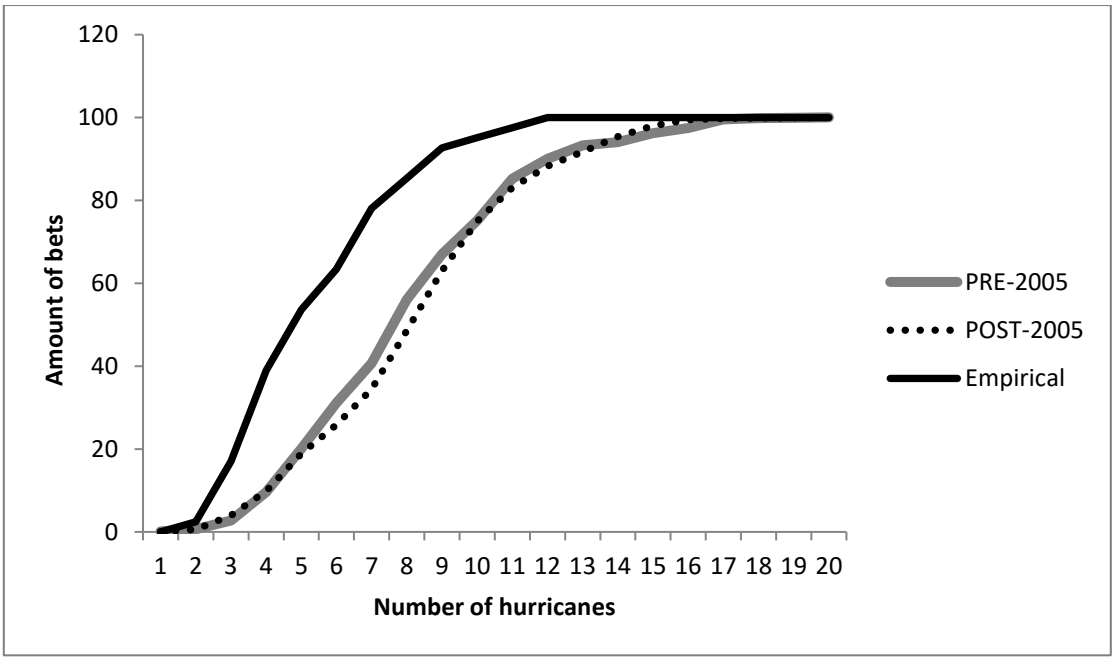
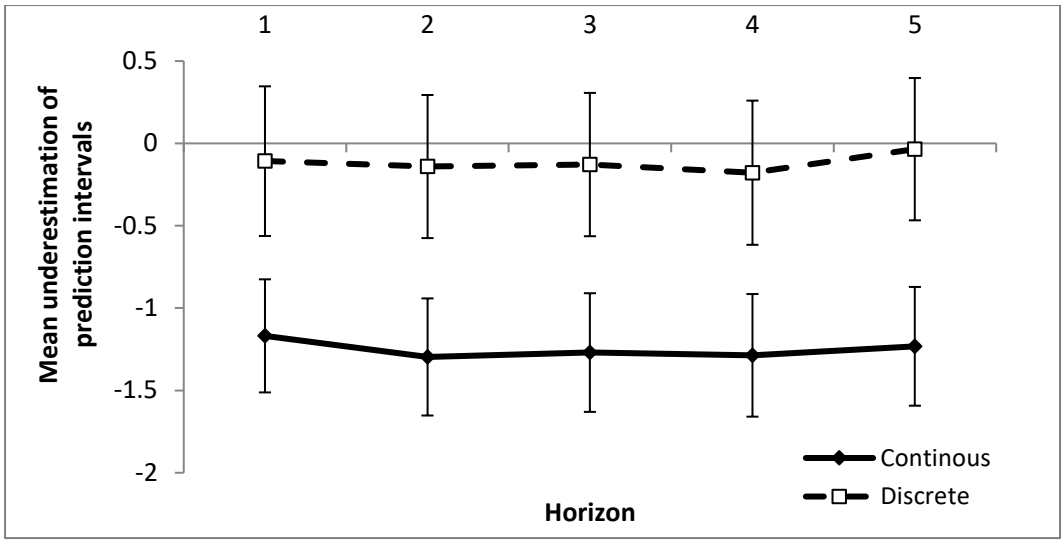

120
100
80
60
40
20
0
Amount of bets
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of hurricanes
PRE-2005
POST-2005
Empirical

Figure 9


1
2
3
4
5
0.5
0
-0.5
-1
-1.5
-2
Mean underestimation of prediction intervals
Horizon
Continous
Discrete