IIF/SAS® Annual Award for Business Applications (2012-2013)

FINAL REPORT

Improving judgmental input to hurricane forecasts in the insurance and reinsurance sector

Zoe Theocharis, Leonard Smithand Nigel Harvey

2015

Introduction

This is the final report for the eleventh year annual international award for financial support of research in the sector of forecasting and business practice, from the IIF, in collaboration with SAS®. The proposal was submitted under the title: “Improving judgmental input to hurricane forecasts in the insurance and reinsurance sector” and its aim is to investigate ways of improving the judgmental forecasting that underlies insurers’ decisions about future insurance prices.

This report comprises a brief summary of the relevant literature and a report of three experiments. We focus on a judgmental bias that arises when people make judgmental forecasts from real hurricane series and on its amelioration by changing the format in which data are graphically displayed. Our work also includes an innovative approach to eliciting judgmental probability density functions.

The findings described here were reported at the 2014 International Symposium on Forecasting. A version of this paper will be submitted to the International Journal of Forecasting.

Background

Hurricane forecasting

Hurricanes are among the most hazardous of natural disasters, with their occurrence unsettling the lives of countless people through devastation of business and homes as well as through casualties. When a hurricane arrives on land, it can lead to major power outages, intense flooding, long-term displacement as well as economic damage. Many hurricanes, with category strengths ranging from the mild to destructive (e.g., Hurricane Katrina, 2005) have devastated the North Atlantic coast of the USA in the past and will continue to do so in the future. Furthermore, the strength and severity of future storms is predicted to increase as a consequence of global warming and climate change (Inman, 2010). Hurricane forecasting is absolutely vital to ensure that sufficient preparations and emergency procedures are in place in anticipation of hurricanes. (One such preparation is related to the adjustment of pricings in the insurance and reinsurance sector).

Every year, the United States National Oceanic and Atmospheric Administration (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. Research on the judgment processes underlying these forecasts has not previously been examined. Our experiments represent an initial attempt to remedy this situation.

Biases in judgmental forecasts

Despite their proven efficiency (Lawrence, Edmundson, and O'Connor, 1985), judgmental forecasts and judgmental adjustments to model-based forecasts are affected by a number of biases: forecasters appear to overestimate the degree of autocorrelation in the series (Reimers & Harvey, 2011), underestimate trends in the series (Harvey & Reimers, 2013), include noise in their sequence of forecasts (Harvey, 1995), and make forecasts that are higher for desirable variables, such as sales and profits, than for undesirable ones, such as losses (Harvey & Reimers, 2013). Hurricane time series do not possess a significant degree of autocorrelation and so we expect that judgmental forecasts made from them will be influenced by the first of these biases. They are not trended and so will not be subject to the second bias. (They may be affected though by the third and fourth bias).

Most of the research that has identified these biases has been carried out on simulated series. The statistical characteristics of such series are known with some certainty and so optimal forecasts can be identified and used as a basis for extracting forecast error measures. However, it is important to validate findings from studies using simulated series on real series to ensure their ecological validity.

This is an important issue because Lawrence & O'Connor (1995) failed to find evidence of trend damping in a *set* of real series drawn from the M-competition (Makridakis et al, 1982). Reimers and Harvey (2011) argued that their findings do not conflict with those derived from studies using simulated series. They suggested that people are adapted to the features of their environment and that, as a result, there should be no biases when people forecast from series with features (trend, autocorrelation) that are representative of those naturally present in their environment. Of course, not all natural series will be representative of the environment in this way. Some will contain trends and autocorrelations that are greater than the average trend and autocorrelation in the environment and some will contain trends and autocorrelations that are less than the average of those in the environment. The judgmental biases in these two cases will be in opposite directions. As a result, when people are tested on series randomly (or quasi-randomly) drawn from the environment (e.g., the M-competition), biases in one direction will cancel out those in the other direction and there will be no overall bias when measures are taken across the set of series as a whole. This is what Lawrence and O'Connor (1995) found. If Reimers & Harvey's (2011) interpretation is correct, the autocorrelation bias should be found in natural series (e.g., hurricane series) in which the autocorrelation is close to zero and, hence, less than the positive autocorrelation that is typical of such natural series.

The format in which time series are presented to forecasters influences the degree to which their forecasts are biased. For example, Harvey and Bolger (1996) showed that trend damping is lower when data are presented graphically than when they are presented in tabular format. More recently, Harvey & Reimers (2012) have shown that the size of biases is affected by the particular graphical format that is used to present the time series data: they found that trend damping is greater when time series data are presented to forecasters as lines or points than when they are presented as bars. Earlier, Lawrence and O'Connor (1992) found that forecasting performance is affected by the scale of graphs used to present data to forecasters. Thus it should be possible to reduce the size of biases in hurricane forecasting by judicious selection of the graphical format used to represent the data. This, in turn, should increase forecast accuracy. What type of graphical format is most likely to have this beneficial effect? To answer this question, we must turn to research on graphical perception.

Graphical perception

Apart from the research cited above, there has been no work on the effect of different types of graphical format on judgmental forecasting. However, there has been some investigation of graphical format effects in other contexts. For example, Zacks and Tversky (1999) presented participants with the same set of data in either bar graph (i.e. discrete) or line (continuous) displays. They found that participants were more likely to describe a relationship between x and y variables as being continuous when a line graph was shown than when a bar chart was used. This could apply even to dichotomous variables. For example, some participants, who were presented with line graphs showing gender on the x-axis against height, 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 data category rather than the relation between those categories.

People over-emphasise the relation between successive data 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 graphical format serves to de-emphasise the relation between successive points. As a result, forecasts should be less strongly anchored on the last data point. When there was no autocorrelation in a data series, this, in turn, should lead to forecasts being more accurate with the discrete format than with the continuous format.

Summary

We test the hypothesis that people make forecasts closer to the last data point with continuous than with discrete graphical format (H_1). Furthermore, we test the hypothesis that, with hurricane series that have no significant autocorrelation, this will result in more accurate forecasts with the discrete graphical format (H_2). We test these hypotheses with three different types of forecasting task: point, probability density and prediction interval forecasting.

Experiment 1: Point Forecasting

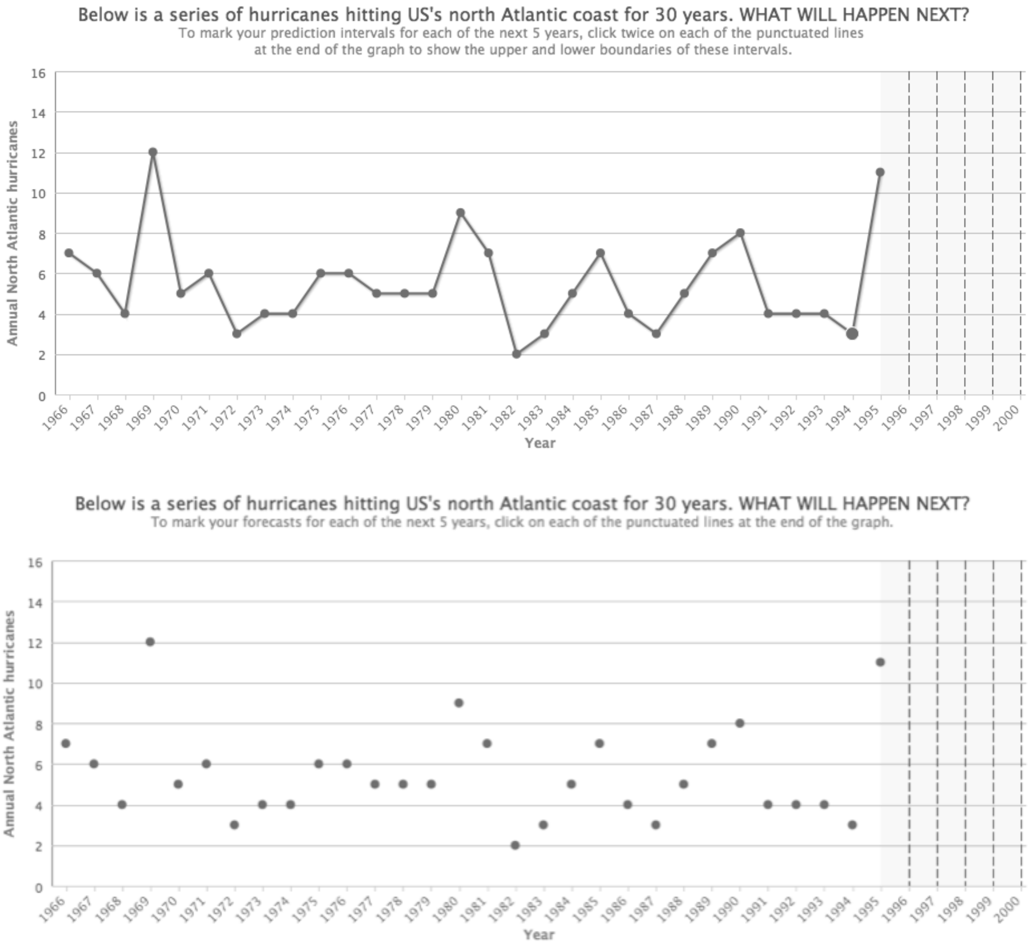
In this experiment, participants were presented with 13 time series exemplars of 30 years' historical hurricane occurrences drawn from NOAA's database and they were asked to make judgmental point forecasts for the next five years. (Forecasts for five years are officially used in practice). Hurricane time series were presented either in continuous line graphs or as discrete, unconnected point graphs in a between participants design. We compared forecasting performance and the degree to which forecasts were anchored on the last data point (a measure of the autocorrelation bias) in these two conditions.

Method

Participants In total, 60 students (46 females, 14 males) at University College London acted as participants. Their mean age was 20 years. They were not paid for their participation.

Design Participants were divided into two groups. The first group (continuous representation) produced point forecasts from continuous line graphs while the second group made their predictions from unconnected point graphs. Thirty participants were randomly assigned to each condition.

Stimulus materials The stimuli used here comprised 13 hurricane time series (see Table 1 for information displayed and requested for each exemplar and Figure 1 for an example of the discrete and continuous representation conditions) showing the annual number of hurricanes hitting the Atlantic coast area from 1966 to 2012. Each time series depicted 30 years' historical data (from 1966 onwards) of the annual numbers of hurricanes hitting the north Atlantic coast in the U.S. Thus, the y-axis showed the number of hurricane occurrences while the x-axis represented time in years. All data were drawn from official sources (NOAA, 2013). In the current work, only a subset of this hurricane occurrences database was used (1966 to 2012) because this was the only period where satellite technology was available to accurately monitor hurricane activity. In this subset, hurricane time series showed no autocorrelation or global trends. The first 30-point time series presented to participants corresponded to hurricane data for the period 1966 to 1995. The next one rolled forward by one time-step ahead, thereby presenting data from 1967 to 1996. The same rolling procedure produced the rest of the exemplars up until the 13th exemplar, where data presented corresponded to the period 1978-2007. In this last exemplar, participants had to produce forecasts for the period 2008-2012. The experiment was coded in Javascript.



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.

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.

Figure 1. Screenshot of hurricane series presented in the discrete and continuous conditions.

Table 1 Information displayed and requested for each exemplar (experimental trial)

Plot	Thirty years displayed	Five years to be forecast
1	1966-1995	1996-2000
2	1967-1996	1997-2001
3	1968-1997	1998-2002
4	1969-1998	1999-2003

5	1970-1999	2000-2004
6	1971-2000	2001-2005
7	1972-2001	2002-2006
8	1973-2002	2003-2007
9	1974-2003	2004-2008
10	1975-2004	2005-2009
11	1976-2005	2006-2010
12	1977-2006	2007-2011
13	1978-2007	2008-2012

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 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 5 years on the 5 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 predictions for 13 different time series and hence produced a total of 65 forecasts. After completing all 13 trials, 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).

Optimal forecasts from the hurricane time series lie on the mean value of the data series because it contained no trend or autocorrelation. Hence, to measure accuracy in order to test H_2, we extracted 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

Mean absolute difference scores Graphs of MAD in the two conditions are shown in Figure 2. One-way repeated-measures ANOVAs showed effects of forecast horizon for both the continuous (line) format (F (4, 1556) = 9.08, p < .001) and the discrete (point) one (F (4, 1556) = 39.52, p < .001). Post-hoc tests then showed that the MAD score for the first horizon was significantly higher than the MAD score for the second horizon (Lines: F (1, 389) = 18.40, p < .001, η^2 = 4.5%,; Points: F (1, 389) = 85.01, p < .001, η^2 = 17.9%). There were no other significant differences between the pairs 2-3, 3-4 and 4-5.

The mean MAD score for the first horizon was greater in the group that saw data in the discrete format than in the group that saw them in the continuous format (t (778) = 3.49, p < .001). This indicates that anchoring was greater with the continuous format and is consistent with H_1.

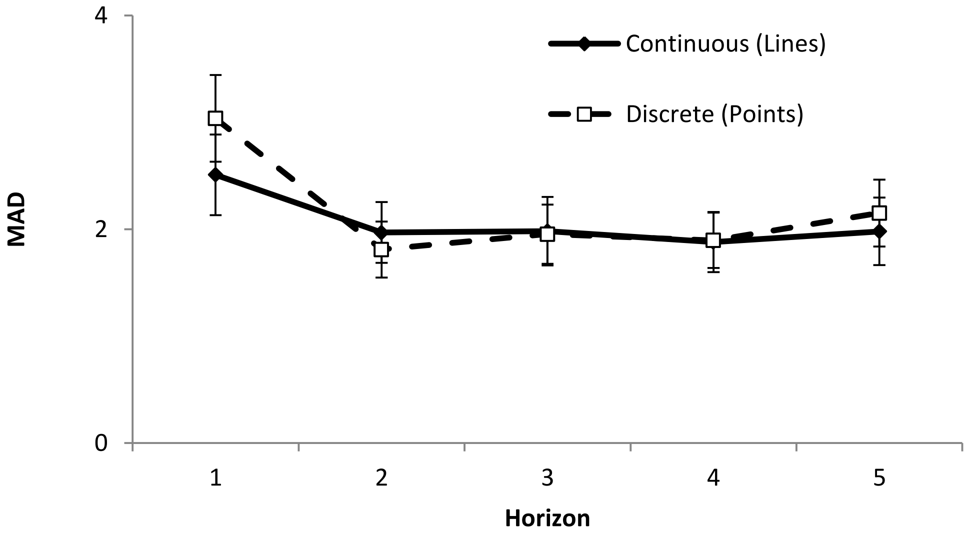


Figure 2. Mean MAD scores for five forecasts in two conditions with standard error bars.

Absolute difference from the mean Graphs of the mean ADFMs in two conditions are shown in Figure 3. Forecasting performance was better when participants saw data in the discrete data format. The difference between formats was significant for the first horizon (t (778) = 3.40, p < .001, d = .247), second horizon (t (778) = 3.35, p < .001, d = .235), fourth horizon (t

(778) = 2.02, p = .022, d = .145), and fifth horizon (t (778) = 2.16, p = .016, d = .247). These results are consistent with H_2.

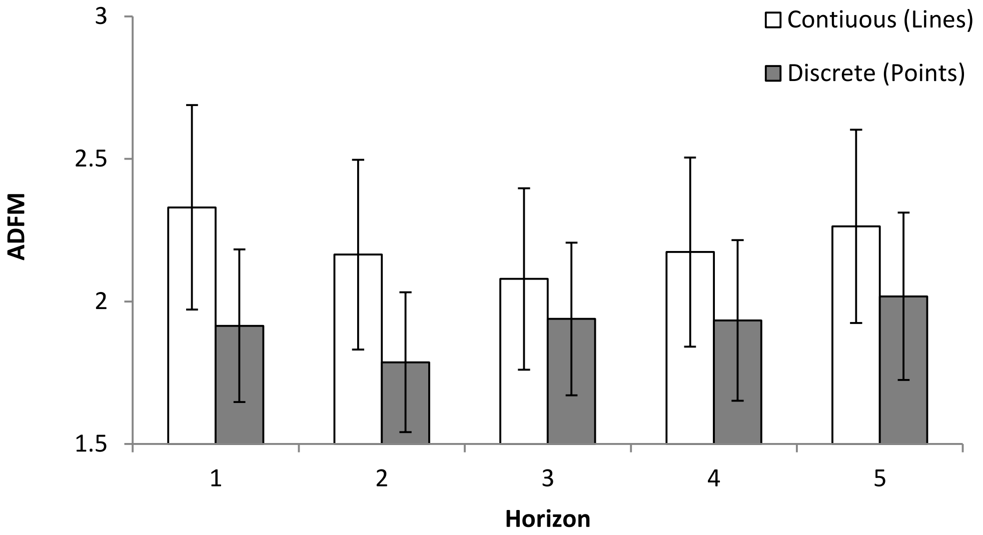


Figure 3. Mean ADFM scores for five forecasts in two conditions with standard error bars.

Discussion

Graphical presentation of the time series did have an impact on the forecasters' performance: forecasts for all horizons (except horizon three) were inferior when data were presented in the continuous format. In line with Zacks and Tversky (1999), 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. However, with series with high autocorrelation, the autocorrelation that people perceive (as implied by their forecasts) is 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.

The first hypothesis was partially supported: format influenced forecasting only for the first horizon. This implies that the effect of format identified by Zacks and Tversky (1999) serves to emphasize the relation between successive points in the data series but not the relation between the last data point and the first forecast or the relations between successive forecasts. Thus the discrete format reduced anchoring but this beneficial effect was specific to the first forecast (Figure 2). As a result, its effect on performance was maintained but did not increase over the remaining horizons (Figure 3). (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 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 ($\chi^2 = 7.69$; $p < .01$).

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

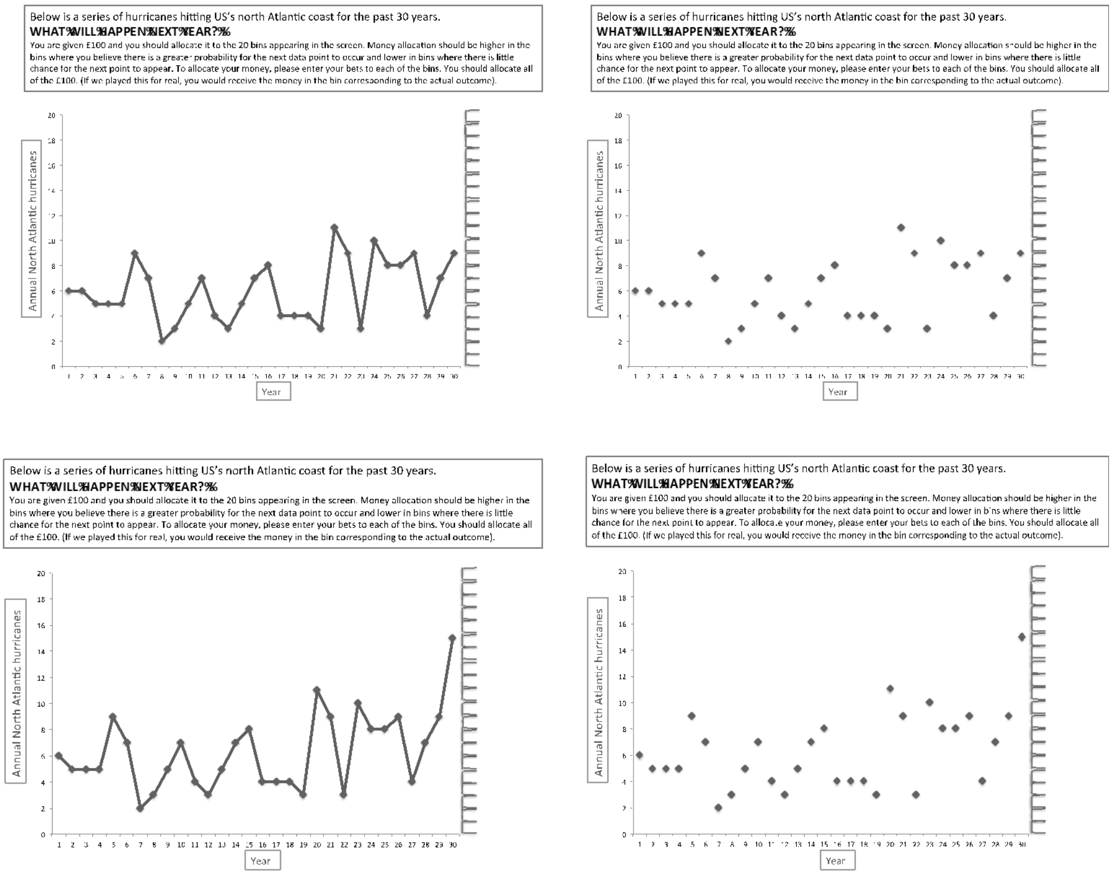
Participants Eighty university students, (21 males and 59 females), aged 18 to 26 (M = 21.27, SD = 1.77), participated in the experiment. They were randomly assigned to the continuous or discrete format conditions, with the constraint that there were 40 participants in each condition. They were not paid 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 was classified as close whereas one outside that range was categorized as distant.)   The dependent variable was participants' one-step-ahead density forecasts measured by bets' spreading into the twenty bins available.

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 are similar to plot 10 and plot 11 in experiment 1 (see Table 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 (e.g., bin 10 from bottom corresponded to 10 hurricane occurrences).

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. They only differed in one value: the number of hurricanes in year 2005, which was substituted for the number of hurricanes in 1975. Thus, any differences in bets between these two exemplars should be attributed to the value of the last data point in the series, namely, nine hurricanes in 2004 (close to the series mean) and 15 hurricanes in 2005 (distant from the series mean).

Data were presented as continuous line graphs in one condition and as discrete points 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).



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).

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).

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).

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).

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.

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 0 to 20 hurricanes, enabled participants to allocate their bets for the next year. Money allocations (i.e. bets) should be higher for bins where there is perceived higher probability for the next data point to occur, and lower for bins where there is less chance for next point to occur. After completing the task for one time series, participants 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 bets that participants should have placed based on the hurricane frequencies they were given for the past 30 years. 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.

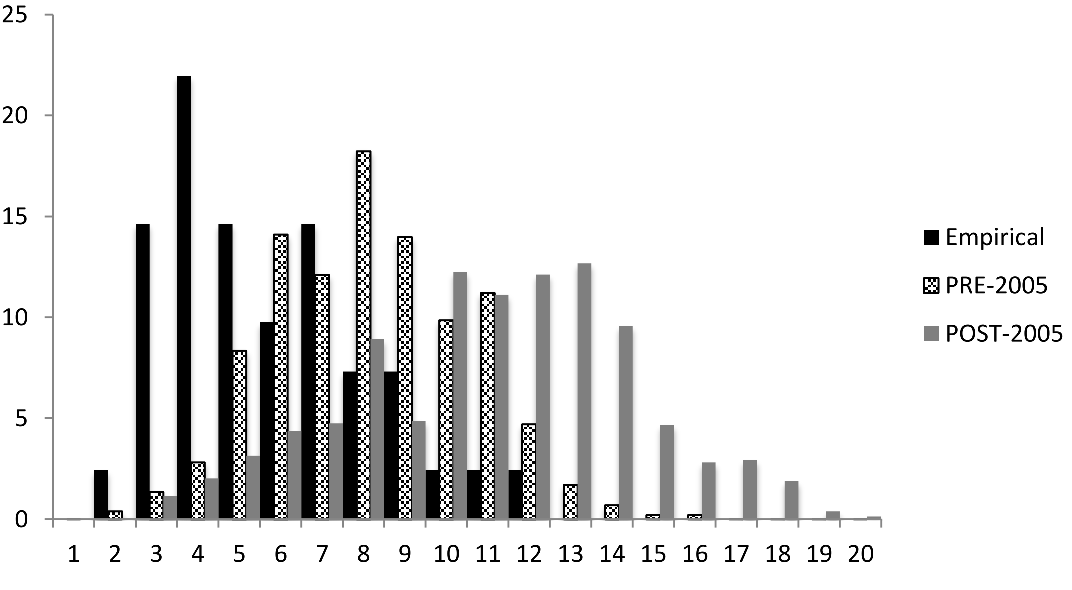


Figure 5. PDFs of the aggregated results (continuous format), together with empirical data.

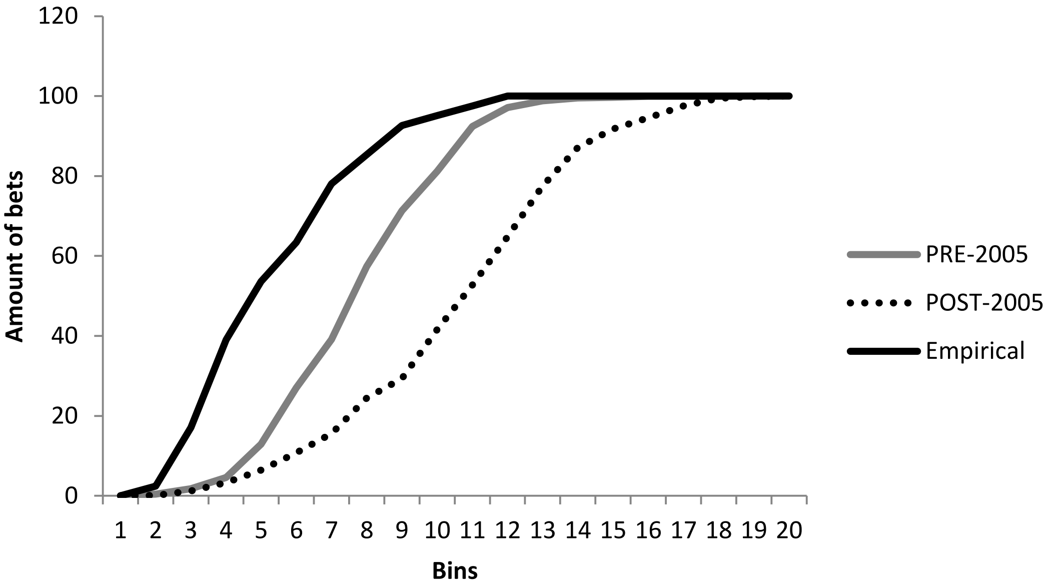


Figure 6. CDFs of aggregated results (continuous format), together with empirical data.

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. This was expected on the basis of the anchoring account because the last data point in the pre-2005 series was somewhat 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.

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.

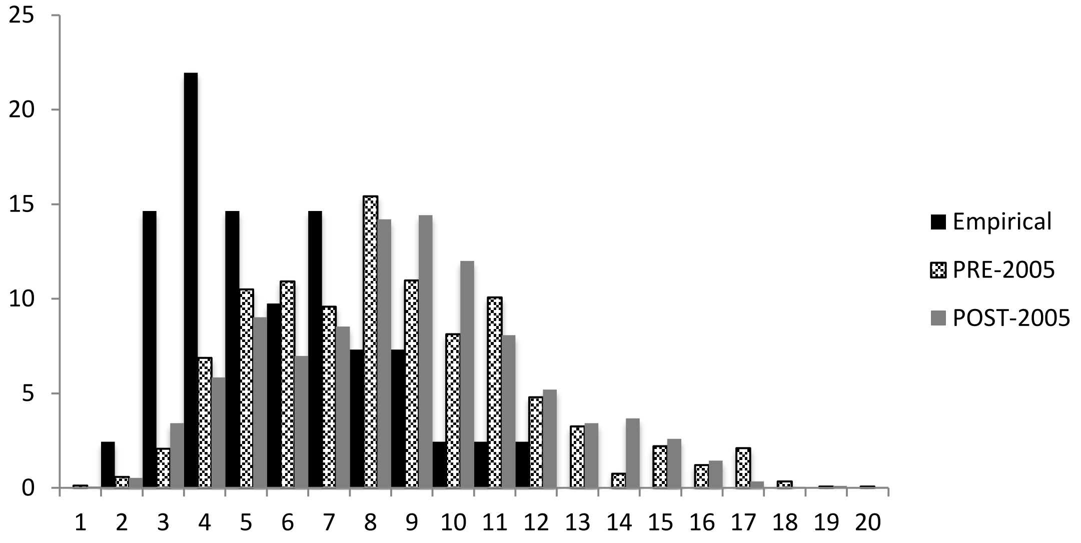


Figure 7. PDFs of the aggregated results (discrete format), together with empirical data.

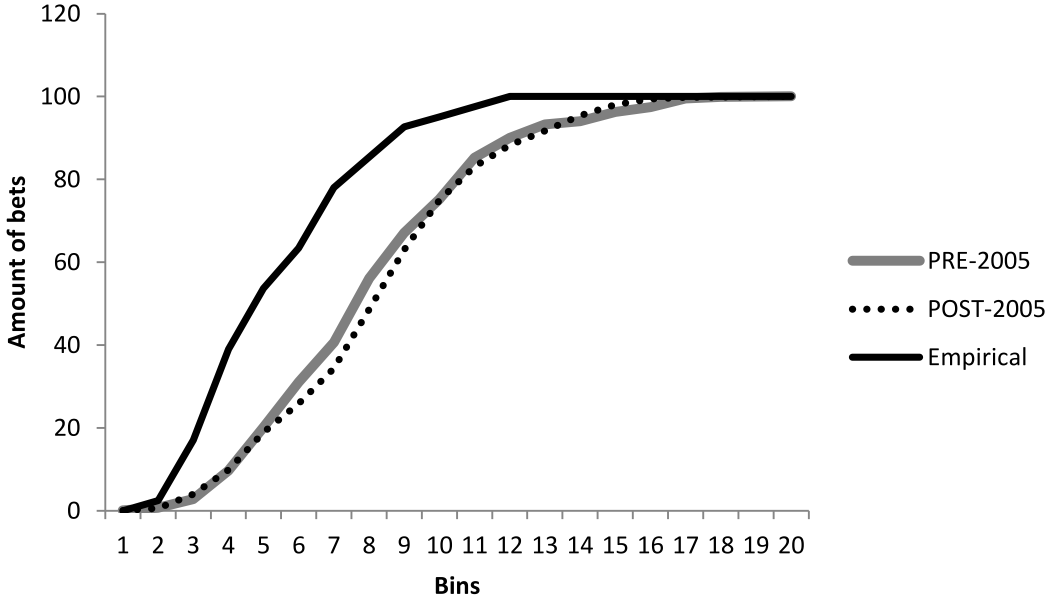


Figure 8. CDFs of aggregated results (discrete format), together with empirical data.

Participants appear to systematically over-forecast. One 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.

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 averaged the bets participants allocated in bins for hurricane occurrences 5 to 9 (i.e. average hurricane activity range of one standard deviation) and those for bins with hurricane occurrences 10-14 (i.e. extreme hurricane activity range, greater than one standard deviation). Bets were summed separately for the pre-2005 series (1975-2004) and post-2005 series (1976-2005) in both continuous and discrete displays conditions. Four independent samples t-tests were conducted to compare the amount of bets placed in average hurricane activity bins for the pre-2005 series between the two displays, in extreme hurricane activity bins for the pre-2005 series between the two displays, in average hurricane activity bins for the post-2005 series between the two displays, and, finally, in extreme hurricane activity bins for the post-2005 series between the two displays.

Results revealed significant differences in bets placed in extreme activity bins during the post-2005 period such that higher bets were observed for the continuous presentation (t (398) = 4.17, $p < .001$). Additionally, bets placed in the average activity bins during the post-2005 period were significantly higher for the discrete display (t (398) = -5.72, $p < .001$). No differences were found between the amount of bets distributed to the extreme activity bins and average activity bins in two conditions for pre-2005 series (t (398) = 2.20, p = .826, for extreme activity bins; t (398) = 1.37, p = .173, for average activity bins).

These results 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 hurricane forecasting where density forecasting is the norm.

Experiment 3: Forecasting using prediction intervals

Every year, the United States NOAA's Climate Prediction Centre provides a formal, model-derived seasonal outlook of overall expected activity for the year's hurricane season. Expected activity is provided in the form of prediction intervals. These comprise prediction bounds that specify upper and lower forecast boundaries within which the future value of the predicted variable is expected to lie with specific probability (Lawrence et al., 2006). This probability is typically set to 70%.

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. To date, there has been no research on their judgment processes: relative contributions of different pieces of information to the final forecasts and decisions are unknown. This experiment provides an initial study of the contribution provided by their judgmental prediction intervals.

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). Because adjustment is typically insufficient (Tversky & Kahneman, 1974), intervals are too narrow.

Participants were presented with the same historical hurricane time series data that we used in Experiment 1 but were now 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 will 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.

Method

Participants Sixty students (40 females, 20 males) at University College London acted as participants. Their mean age was 19.9 years. They were not paid for their participation.

Design Participants were divided into two groups. The first group (continuous representation) produced prediction intervals from continuous line graphs while the second group made their predictions from unconnected point graphs (discrete representation). Thirty participants were randomly assigned to each condition.

Stimulus materials The same time series that were used in Experiment 1 were employed here. At the end of the x-axis of each one were five vertical, punctuated lines that represented the next five years in the series. These allowed participants to mark their 70% prediction interval forecasts.

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 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 between the two conditions, with the size of the intervals 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) in the continuous display than in the discrete one. The data were consistent with this for all horizons (Figure 9).

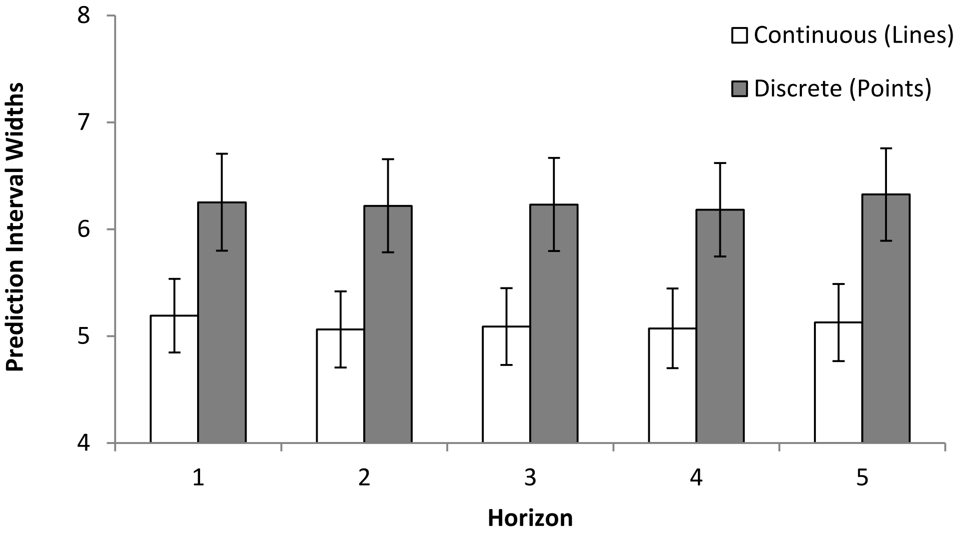


Figure 9. Mean prediction intervals for five forecasts in two conditions with standard error bars.

To examine the significance of these effects, we carried out a two-way analysis of variance (ANOVA) on the prediction interval widths, using horizon as a within-participants variable (five levels) and condition (continuous display, discrete display) as a between-participants variable. Here and later, Huynh-Feldt corrections were applied to address violations of sphericity. There was a significant main effect of condition (F (1, 58) = 5.84; p < .05). The main effect of horizon and the interaction between the two variables were not significant.

We calculated the magnitude of actual prediction intervals for each series. These were then compared to the average forecast prediction intervals in the two conditions. Participants were expected to be overconfident in their forecasts. In other words, participants' 70% prediction intervals were expected to be narrower than the actual 70% prediction intervals. This was found to be true for the continuous display condition (t (389) = 15.6, p < .001) but not for the discrete display condition (t (389) = 1.11, p = .267). Thus, prediction intervals in the continuous format condition (M = 5.10) but not in the discrete format condition (M= 6.24) were narrower than the actual prediction intervals (M = 6.36): overconfidence appeared only with the continuous display.

According to the anchoring account (Harvey, 1997), the upper and lower bounds of prediction intervals are produced by adjusting away from the last displayed point and intervals are too narrow because adjustment is insufficient (Tversky & Kahneman, 1974). The above analysis is consistent with this account when displays were continuous. However, when they were discrete, intervals were not too narrow: adjustment appears to have been appropriate.

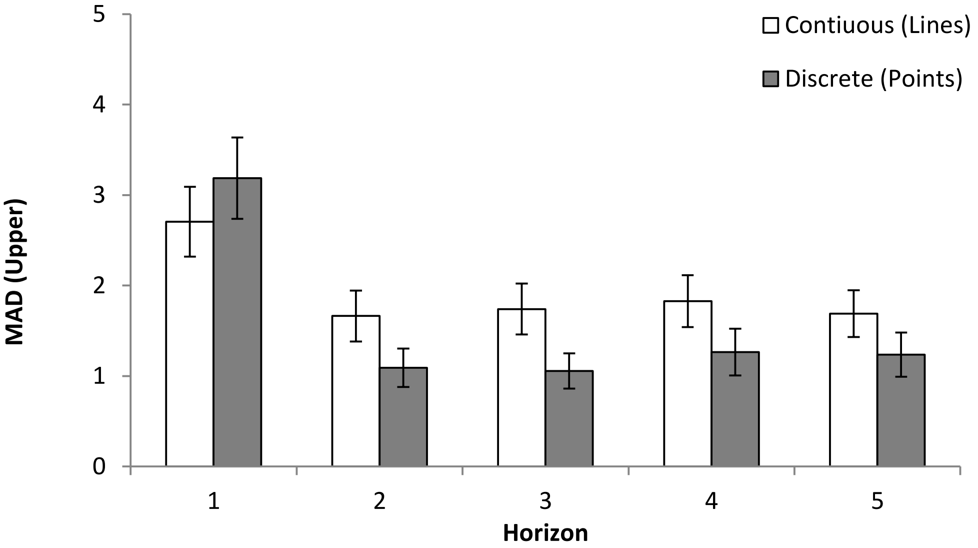


Figure 10. Mean MAD scores for upper prediction interval forecasts for the five horizons in the two conditions with standard error bars.

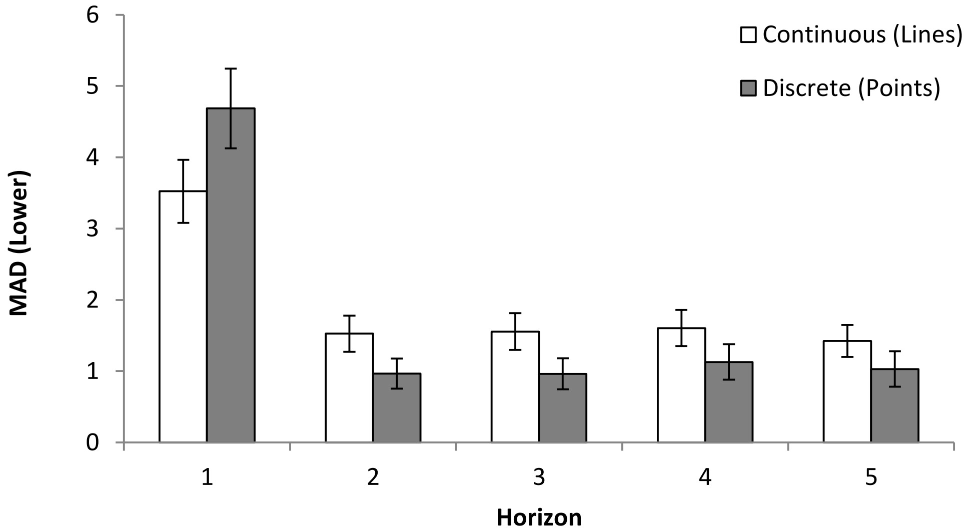


Figure 11. Mean MAD scores for lower prediction interval forecasts for the five horizons in the two conditions with standard error bars.

This analysis implies that adjustment from the last displayed point for both upper and lower bounds of the interval was greater in the discrete format condition than in the continuous display condition. Thus, we extracted MAD scores for both upper and lower bounds of the intervals (Figure 10 and Figure 11) and used two one-way ANOVAs to compare them across conditions. For the first horizon, these analyses showed that MAD scores were greater in the discrete display condition both for the upper bound of the interval (t (389) = 2.11, p < .05) and for the lower one (t (389) = 10.84, p < .001). These results are consistent with the anchoring account.

There were differences in upper and lower prediction interval MADs between the pairs 2-3, 3-4 and 4-5 but in the opposite direction. All of these differences were significant. However, it is important to emphasize again that the MAD scores for later horizons are relative to the previous forecast (i.e. upper or lower bound of the previous prediction interval) rather than relative to the last point of the data series. Thus, these findings for horizons beyond the first one indicate that people made less change to the size of the interval as horizon increased when intervals were already wide (discrete format) than when they were not (continuous format). They are not inconsistent with the anchoring account.

Discussion

Forecasting using prediction intervals was suboptimal with the continuous format but not with the discrete one. 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' intervals are well-calibrated. 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 immensely to the accuracy of forecasting, but they are sometimes subject to certain errors. Using uncorrelated and un-trended real hurricane time series, the main 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 these biases 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 for a widely varying set of 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 representative of their environment as a whole. 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: 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 cancel each other out.

Thus, the anchoring effects that we have demonstrated with real series (e.g., Experiment 2) 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 in 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, 2013). 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 manner in which 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. As mentioned above, we have reasons to suspect that they will not. Hence, future experiments should test real times 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?

References

Bolger, F., and Harvey, N. (1993). Context sensitive heuristics in statistical reasoning. *Quarterly Journal of Experimental Psychology*, *46A(4)*, 779–811.

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

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.

Inman, M. (2010). Settling the science on Himalayan glaciers. *Nature Reports Climate Change*, *19,* 28-30.

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

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. J., & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes, 42,* 172 – 187.

Lawrence, M., & O'Connor, M. (1992). Exploring judgmental forecasting. *International Journal of Forecasting, 8,* 15-26.

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., 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.

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.

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

Tversky, A., & Kahneman, D., (1974). Judgment under uncertainty: heuristic and biases, *Science, 185*, 1124-1131.

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

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