Is Ruth Chepngetich's Marathon World Record Valid? A Statistical Analysis of Elite Marathon Performances

Abigail Mabe and Richard L. Smith*
Department of Statistics and Operations Research
University of North Carolina, Chapel Hill, NC 27599-3260, USA

January 31, 2025

1 Motivation and Outline of Paper

On October 13, 2024, Kenyan runner Ruth Chepngetich ran a women's world record 2 hours, 9 minutes and 56 seconds (2:09:56) at the Chicago Marathon, the first time ever that a woman had run under 2:10, improving on the previous world record by nearly two minutes, and Chepngetich's personal best for a marathon by nearly five minutes. This performance immediately drew commentary querying the validity of the record, specifically, whether Chepngetich was taking illegal drugs. The respected running journalist and former Boston Marathon winner Amby Burfoot wrote an article[1] citing the large number of Kenyan athletes who had failed drug tests and comparing her performance with, amongst others, Rosie Ruiz's win in the Boston Marathon in 1980 (Ruiz was disqualified after the organizers determined she had not run the full course) and the performances by Chinese female track runners in 1993 (that are widely believed to have been due to illegal drug use). However, Burfoot acknowledged that Chepngetich has never failed a drug test and there is no proof that she was using illegal drugs on this occasion. A counter-argument was made by running journalist Jessy Carveth[2], who discussed the likely benefits of pacing (Chicago was a mixed-gender race so it is entirely plausible that Chepngetich's performance benefitted from the presence of male runners around her) and the improvements in running shoe technology which have resulted in substantially faster running times both among elite runners and at the recreational level [4]. Despite these disagreements, World Athletics, the international governing body of track and field, ratified the record[3], but this will not settle arguments over its validity.

In this paper, we use statistical analysis of previous performances in the Chicago Marathon to argue that a sub-2:10 performance, while still very impressive, was not in fact so implausible, in fact one could almost say that such a performance was inevitable within a few years. The statistical analysis is based on extreme value theory, a method earlier used in analyzing the performance by Chinese female runners in 1993 [10, 12], but taking account of recent trends in marathon times.

*Contact email: rls@email.unc.edu

[1] https://marathonhandbook.com/opinion-why-its-hard-to-trust-ruth-chepngetichs-marathon-world-record

[2] https://marathonhandbook.com/ruth-chepngetich-world-record-plausible

[3] https://worldathletics.org/news/press-releases/ratified-world-records-chebet-duplantis-mclaughlin-levrone-chepngetich-kawano

Whereas the 1993 analyses of Chinese runners only reinforced the suspicion that those performances were drug-assisted, here we reach a directly contrary conclusion.

2 Statistical Analysis

The initial analysis is based on the 20 best women's running times from the Chicago Marathon website[4] for 1998–2024, but excluding 2020 for which there was no race because of the Covid pandemic. The start year of 1998 was chosen to be far enough back in time to allow a sufficiently long series for analysis, but not so far back that trends in the early part of the record would be completely irrelevant to today's times. Later in this paper, we explore the sensitivity of the results to the choice of starting year for the analysis.

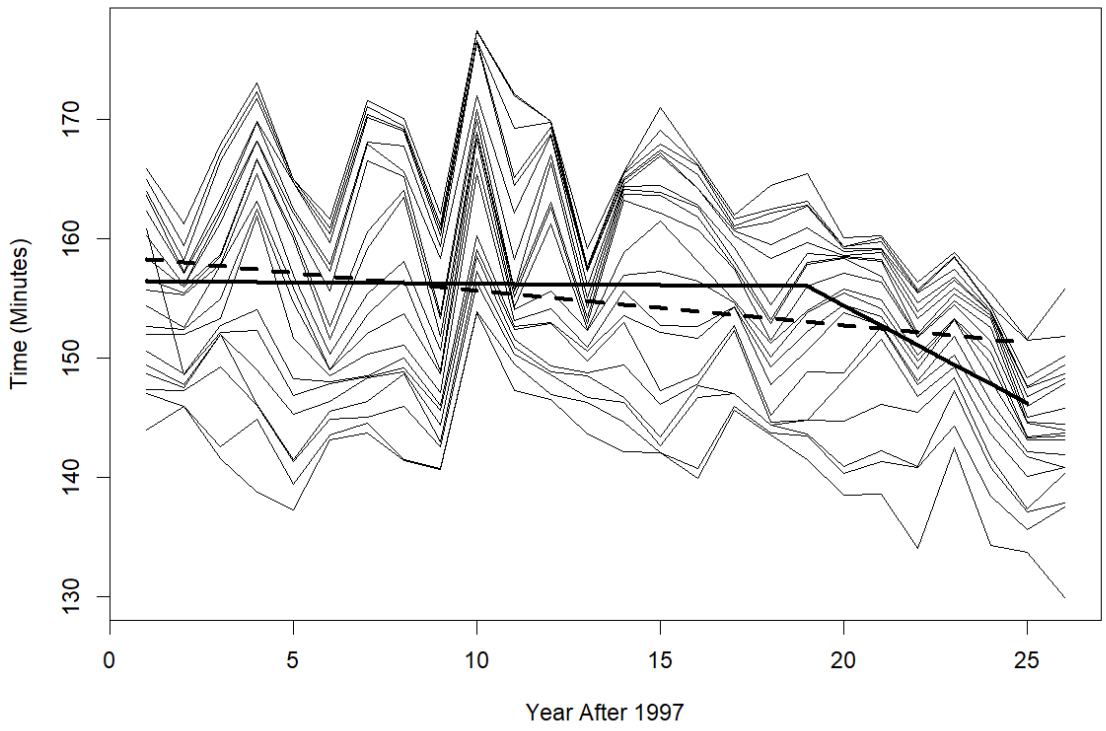

Figure 1: Best 20 Running Times in the Women's Race in the Chicago Marathon, 1998–2024 (omitting 2020)

Figure 1 plots the 20 best times for each year after 1997 (omitting 2020, so year 26 corresponds to the 2024 race). It is clear that there is some random variation from year to year — for example, the times for 2007 (year 10 in this plot) were greater than those for the years around that time, which reflects the extremely hot conditions that year (in fact, in 2007 the police suspended the race after many runners collapsed on the course). However, overall the graph shows that times

[4]https://www.chicagomarathon.com/runners/race-results/. Some apparent errors have been corrected.

have steadily decreased (i.e. improved) during the 26 years of the study, with possibly a greater rate of improvement since 2017, the year that Nike Vaporfly running shoes were introduced. Since that year, nearly all the top runners have used Vaporflys or one of the other shoes that feature a carbon plate insert, that have been attributed for the rapid improvement in running times in recent years. Also shown on the graph are a linear fit through all the datapoints up to 2023 (dashed line), and a "changepoint model" (solid lines) featuring a change of slope beginning in 2017. Visually, it appears that the changepoint model fits the data better than the linear model, and we provide quantitative evidence to back up that assertion in the following analysis. However, the lines plotted in Figure 1 use standard linear regression methods, and do not take account of the special character of extremes, which is relevant to the present discussion because these running times are extremes, i.e. the best 20 runners among tens of thousands of participants.

2.1 Statistical model

The proposed statistical analysis is based on the "r largest order statistics method", which was first proposed in the 1980s [11, 14] and is described in detail in the book by Coles [2]. This method was originally developed to study extreme environmental events and was the basis for two papers published about the performances achieved by Chinese women runners in 1993 [10, 12]. The idea is to pick a relatively small number r — typical examples are numbers like 5 or 10 — and to analyze the r most extreme events in each year. For the present study we have modified the method to focus on the r smallest events since we are talking about running times, but we still use "r largest" for the description (the mathematics is essentially the same whether one is considering largest or smallest events).

The foundation of extreme value analysis is the Generalized Extreme Value distribution, usually abbreviated to GEV, which in the present form adapted to sample minima is of the form

$$\Pr\{Y \le y\} \;=\; G(y;\mu,\sigma,\xi) \;=\; 1-\exp\left[-\left\{1+\xi\cdot\frac{\mu-y}{\sigma}\right\}_+^{-1/\xi}\right] \tag{1}$$

where Y is the variable of interest (in this case, the minimum running time in a marathon), y is some real number where the distribution is to be evaluated, and μ, σ and ξ are the parameters of the distribution. We also note the notation $\{\cdot\}_+$ where the subscript + signifies that negative values are replaced by 0. Here μ is the location parameter, i.e. it represents the center of the distribution we are studying, which in the present example, would be somewhere near the mean winning time of the race. The parameter σ is the scale parameter, and represents how widely spread out is the distribution, while ξ is the shape parameter that has a strong influence on how extreme the extreme values can get. A positive value of ξ typically represents a distribution where there is no well-defined limit to how large the process can get; for example, windspeeds in a hurricane. In contrast, $\xi < 0$ represents a short-tailed distribution. In temperature extremes, for example, ξ is typically about –0.1, reflecting that temperatures are bounded but not very tightly constrained. For running records, values in the range –0.3 to –0.5 are common, meaning a short-tailed distribution, and in the results presented here, the values of ξ are of the order of –0.5 to –0.6, reflecting a very short-tailed distribution (i.e. very low probability of a large jump in the record). Nevertheless, for the specific data analyzed here, we see improvements of several minutes over the time period of the study.

The basic model assumes that μ, σ and ξ are constant for all years of data. For the Chicago Marathon data, Figure 1 shows clearly that this is not realistic — while we might dispute the exact form of the trend it is clear that there is an overall downward trend in the results from 1998–2023. Therefore, in place of μ, we write μ_t to represent the location parameter in year t, where $t = 1$ corresponds to 1998, and we consider three models:

(a) No-trend model: $\mu_t = \mu$, the same for all t (this model is not realistic as shown by Figure 1, but we retain it in the comparison so that we can show analytically that it is not competitive with the other two models);

(b) Linear trend model: $\mu_t = \beta_0 + \beta_1 t$ where β_0 and β_1 are now the parameters to be estimated. In our analyses, we typically find β_1 about –0.5, which is interpreted as meaning that times among the first r runners have improved at a rate of about half a minute per year;

(c) Changepoint model: Fix a change year t_0 and define $\mu_t = \beta_0 + \beta_1(t - t_0) + \beta_2(t - t_0)_+$ where $(t-t_0)_+$ is 0 when $t < t_0$ and $t-t_0$ when the latter quantity is positive. Under this model, the slope of the line jumps by an amount β_2 after year t_0. The time t_0 at which the slope changes is commonly called a changepoint in statistical analysis. In our analyses we have typically taken $t_0 = 19$, corresponding to a jump after year 19 of the analysis, i.e. calendar year 2016, though we also consider alternative changepoints as part of our sensitivity analyses.

2.2 Estimating the statistical model

All the models contain unknown parameters σ and ξ which have to be estimated. Depending on the choice among models (a), (b) and (c), we also have to estimate either (a) μ, or (b) β_0 and β_1, or (c) β_0, β_1 and β_2. In addition, we would like to have a statistical criterion for deciding which of (a), (b) or (c) is the most appropriate model.

The essential tools to do this are:

(i) Method of maximum likelihood: For each of models (a), (b) and (c), we define the likelihood function, that represents how well the model fits the data with a given set of parameters. The maximum likelihood estimators are, in effect, the values of the parameters that give the greatest agreement between model and data.

(ii) Likelihood ratio test. For comparing any two of models (a), (b) and (c), we first maximize the likelihood under both models, then take one likelihood divided by the other. For example, the likelihood for model (c) will always be larger than the likelihood for model (b), because of the extra parameter β_2 in model (c). The likelihood ratio is the maximized likelihood for model (c) divided by the maximized likelihood for model (b). The likelihood ratio test is a formal statistical test to decide which of the two models is better.

(iii) A third technique (alternative to maximum likelihood) is Bayesian analysis. The textbook by Gelman and co-authors [3] is a very good introduction to Bayesian analysis. In Bayesian analysis, we do not attempt to maximize the likelihood but instead calculate the *posterior distribution* which expresses the uncertainty about the parameters in the form of a full probability distribution. In some of our analyses, this will improve the practical interpretation of the results.

3 Results

Initially we take $r = 10$ and fit the data to years 1–25 (i.e. 1998–2023, omitting 2020). The GEV analysis does not inlcude 2024 because the intent of the analysis is to use data prior to 2024 to assess the plausibility of the result actually observed in 2024. The initial GEV results are in Table 1.

Parameter	Estimate	S.E.	z-ratio	p-value
μ	140.059	0.539	259.95	0
$\log\sigma$	1.261	0.048	26.34	6.1×10^{-153}
ξ	–0.511	0.037	–13.82	1.9×10^{-43}
NLLH	393.713			

Table 1: Table of parameter values for model 0, years 1–25, $r = 10$

The last row here indicates the negative log likelihood, which is mainly useful for comparison with other models.

This table shows the point estimate, the standard error (S.E.), the z-ratio (estimate divided by the standard error) and the p-value for each of the parameters. Note that the parameter σ has been replaced by $\log\sigma$, which leads to a more stable optimization taking into account the constraint $\sigma > 0$. The p-value is calculated as the two-sided probability that a standard normal random variable is greater in magnitude than the stated z-ratio, and is relevant as a test of the hypothesis that the given parameter is 0. In the case of μ and $\log\sigma$ this is not relevant because it is not plausible that either of those parameters could be 0, but for ξ this is plausible, because when $\xi \to 0$, the expression in (2) becomes $1 - \exp[-\exp-\{(\mu - y)/\sigma\}]$ which is known as the Gumbel distribution, a well known special case of the GEV. Therefore, in this case where we observe $\hat{\xi} = -0.511$ with an associated extremely small p-value, we conclude that a value $\xi \geq 0$ is virtually impossible, i.e. we can be certain that $\xi < 0$ which confirms the short-tailed nature of running time distributions. As a side comment, it is conventional in statistical analyses to label any parameter with a p-value less than 0.05 as "statistically significant", so another way of interpreting this result is to say that ξ is (very highly) statistically significantly different from 0. Another quantity of interest is the endpoint of the distribution, represented algebraically as $\mu + \sigma/\xi$, which is the value of y for which $G(y; \mu, \sigma, \xi) = 0$ in (1). Substituting the estimates in Table 1 for μ, σ and ξ, we estimate the endpoint of the distribution as 133.15 (just over 2 hours, 13 minutes), which makes sense based on the fact that the fastest running time in the data (up to 2023) is 133.73 (minutes). The endpoint of the distribution cannot be greater than the best previously observed value, but it should not be surprising that it is fairly close to that. However, this ignores the trend in running times, or any uncertainty in calculating this estimate.

The next model we fit is the linear trend model in (b), where in place of μ we use the time-varying function $\mu_t = \beta_0 + \beta_1(t - t_0)$ where t_0 is an appropriate (but arbitrary) centering time point. In this case we have taken $t_0 = 19$ (corresponding to calendar year 2016), in preparation for the following analysis with model (c) where t_0 is taken as a changepoint. In this case, the analysis yields the results shown in Table 2.

The NLLH is the key quantity needed to implement the likelihood ratio test mentioned in Section 2.2. Compared with model (a), twice the difference of NLLH values (i.e. the deviance) is

Parameter	Estimate	S.E.	z-ratio	p-value
β_0	139.335	0.517	269.602	0
$\log\sigma$	1.122	0.055	20.373	2.9×10^{-92}
ξ	–0.578	0.041	–13.953	2.0×10^{-44}
β_1	–0.211	0.036	–5.945	2.8×10^{-9}
NLLH	384.124			

Table 2: Table of parameter values for model (b), years 1–25, $r = 10$, $t_0 = 19$

$2 \times (393.713 - 384.124) = 19.178$ with one degree of freedom, and the chi-squared tail probability associated with that is 1.2×10^{-5}. This is therefore the p-value for the null hypothesis of no trend against the alternative hypothesis of a linear trend, and since the p-value is very much less than 0.05, we conclude that the linear trend is highly significant. This conclusion is reinforced even more strongly by the estimated β_1 parameter and the even smaller p-value associated with that. However, this simple calculation is only for one version of a trend model, and does not prove that our model is correct. It should be noted that the estimate of the GEV shape parameter ξ, –0.578, is even more negative than in the no-trend model (implying a shorter tail to the distribution) and the estimated end-point of the distribution for 2024, given algebraically by $\mu_{26} + \sigma/\xi$, is now 132.545 minutes, or 2:12:33, which is still more than two minutes slower than the time actually achieved by Chepngetich, but it should be emphasized that this is still an estimate and we have not, so far, taken account of the uncertainty in that estimate.

Now let us turn to model (c), where we write $\mu_t = \beta_0 + \beta_1(t - t_0) + \beta_2(t - t_0)_+$ with an additional parameter β_2 and the centering year t_0 again taken to be 19 to correspond to a changepoint in 2016. In this case the results are as in Table 3. The parameters β_1 and β_2 are harder to interpret in

Parameter	Estimate	S.E.	z-ratio	p-value
β_0	143.139	0.950	150.702	0
$\log\sigma$	1.031	0.064	16.172	8.0×10^{-59}
ξ	–0.617	0.047	–13.128	2.3×10^{-39}
β_1	0.109	0.082	–1.336	0.182
β_2	–1.731	0.447	–3.874	1.1×10^{-4}
NLLH	376.137			

Table 3: Table of parameter values for model (c), years 1–25, $r = 10$, $t_0 = 19$

this case, but the negative value of β_2 suggests strongly that the downward trend has accelerated since 2016. Compared with model (b), the deviance is $2 \times (384.124 - 376.137) = 15.974$ which is statistically significant as a χ^2_1 random variable (p-value 0.000064). All this points to the conclusion that the changepoint model is significantly better than the linear trend model.

In this case the estimated endpoint (i.e. best possible time) for 2024 is 127.244 minutes or 2:07:15, well under the time achieved by Chepngetich, but it should again be emphasized that this is only an estimate and we have taken no account so far of the uncertainty in that estimate.

The same analysis was repeated with different values of t_0. In fact, $t_0 = 19$ is not the optimal value from the point of view of minimizing NLLH; $t_0 = 13$ (corresponding to a change of slope after

2010) produces an NLLH value of 374.401, lower and therefore better than our previous value of 376.137. Treating t_0 as an unknown parameter, and constructing formal esitmates and tests based on that, creates further complications for the analysis and we shall not pursue that path; however, later in this paper we do discuss how the results change based on alternative assumptions about t_0. Since super shoes were only introduced in 2017, they are likely not the only reason that running times have improved more rapidly in recent years; other factors include the increased finanical incentives for setting records at the top marathon events, general improvements in equipment and training technology, and (in the case of women's racing) the trend away from women-only races which has improved the potential for women runners to pace or draft off top male runners.

3.1 Bayesian analyses under model (b)

We also consider Bayesian analyses of these datasets. An argument in favor of Bayesian analyses for these problems is that by representing the results in terms of a posterior distribution rather than an estimator and standard error, they give a more complete picture of the uncertainty in the resulting estimates, especially in cases where the posterior density is highly asymmetric, which is especially the case when estimating the endpoint of a distribution, which is effectively what we are doing when talking about the best possible marathon time. We shall also highlight the advantage of a Bayesian analysis for estimating *predictive distributions*, a point made in earlier discussions of athletic records [12], but reinforced in the present analysis.

As is traditional in Bayesian analyses, the posterior distribution is represented by a sample from the output of a Markov chain Monte Carlo (MCMC) simulation algorithm (details of the simulation are in Section 4). We first apply model (b) to the data from 1998–2023, with parameters β_0, $\log\sigma$, ξ and β_1. Figure 2 shows a trace plot of 5000 values from the posterior distribution for each of the four parameters after discarding an initial warmup part of the sample. The figures show all four parameters well mixed and without displaying any evident departure from a stationary distribution, which is what we want. Figure 3 shows posterior density plots for each of the four parameters, showing fairly symmetric distributions but also confirming the overall precision of the estimates. For example, in the case of ξ, the Bayesian analysis shows a posterior median of –0.562 with a 95% credible interval (from the 2.5 to the 97.5 percentile of the posterior distribution) of (–0.644,–0.478). In contrast, the MLE –0.578 and standard error 0.041 imply a 95% confidence interval (–0.658,–0.498). The differences are slight, but they do highlight the slightly right skewed nature of the posterior distribution.

From our perspective of trying to assess the plausibility of Chepngetich's record, the individual parameters of the extreme value distributions are of less interest than what they say about the endpoint of the distribution, i.e. the best result that could possibly occur under this model. As noted previously, the formula for the endpoint is $\mu_{26}+\sigma/\xi$ which may also be written as $\beta_0+7\beta_1+\sigma/\xi$ after taking into account that the trend in model (b) is centered around the projected changepoint. The posterior density plot for that parameter is given in Figure 4(i). This figure shows that the plausible values for the endpoint are roughly from 128 to 134 (minutes). This distribution has a hard upper boundary at 133.73, the fastest women's time in the Chicago Marathon prior to 2024, since the endpoint cannot be higher than the best previously existing value; the fact that the plotted density goes slightly beyond that value is an artifact of the smoothing operation applied to the posterior density. Of greater interest for the present discussion, however, is the other end of the posterior density plot, and this plot implies that the probability that the endpoint is less than

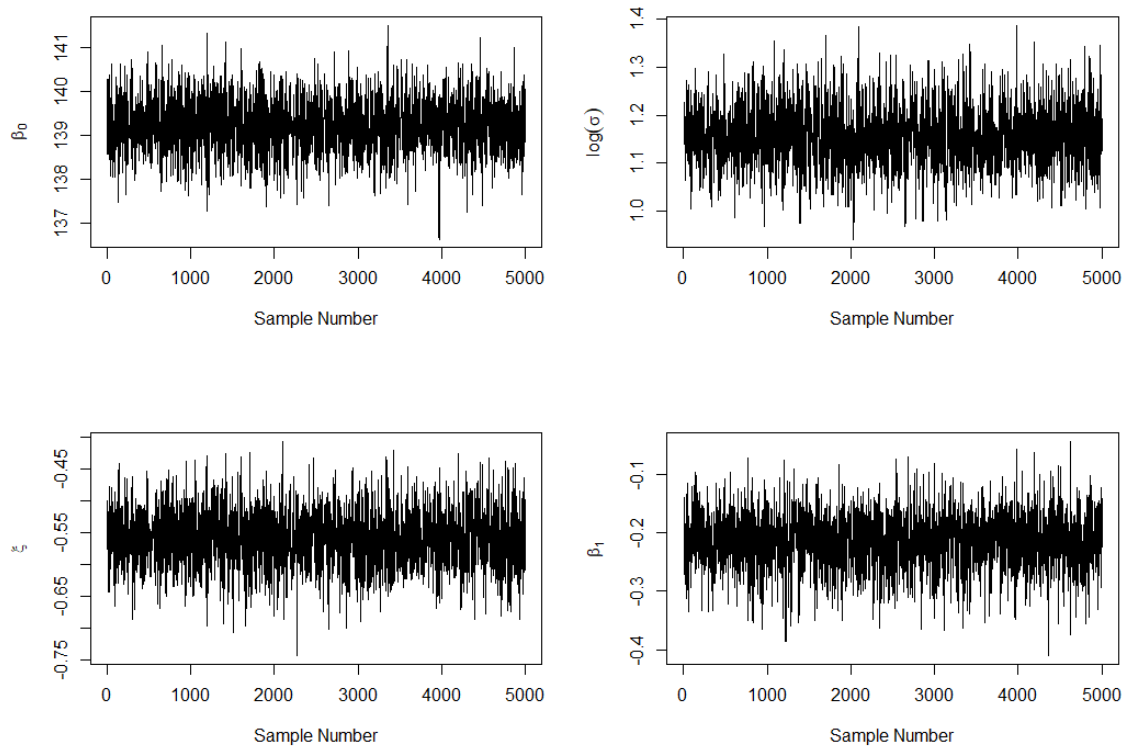

Figure 2: Trace plots in MCMC output for posterior distribution, model (b), data 1998–2023

130 (i.e. 2 hours, 10 minutes), based on all the results prior to 2024, is about 0.034 — a probability that would make Chepngetich's record seem unlikely but by no means impossible.

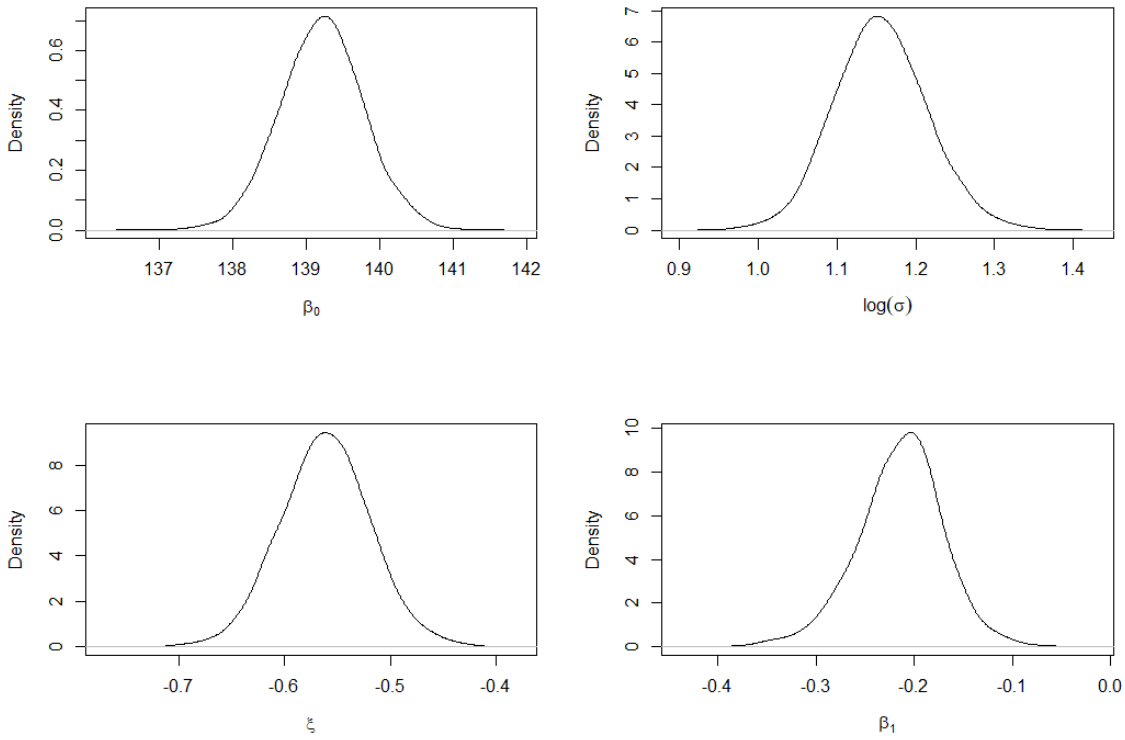

Figure 3: Posterior density plots, model (b), data 1998–2023

However, as argued in a previous somewhat similar analysis in [12], a more relevant calculation is the *predictive distribution* of the winning time, i.e. trying to predict what the winning time would be in 2024 based on data up to 2023, and then calculating a posterior distribution for that value. We assume that the distribution of the winning time is of the GEV form (1) where the parameter μ is replaced by $\beta_0 + 7\beta_1 + \sigma/\xi$ as just discussed. For each set of posterior values $(\beta_0, \sigma, \xi, \beta_1)$ from the MCMC sample, we generate 1000 values from the GEV distribution (1); these are collectively assumed to form a sample from the posterior predictive distribution of the winning time in 2024, conditional on all the data up to 2023. The posterior density of this quantity is shown in Figure 4(ii). From this plot, we estimate the posterior probability of a 2024 winning time better than 130 minutes to be about 0.011 — a smaller probability than we calculated for the endpoint of the distribution, but again, one that makes Chepngetich's record seem unlikely but not impossible.

As a direct comparison, [12] gave a similar analysis (but without trend) for the plausibility of the 3000 meters world record achieved in 1993 by the Chinese runner Wang Junxia. In fact [12] quoted a probability 0.00047 for the *conditional* probability of a performance equal or better than Wang's record, *given a new world record*. In the present case, this would translate to a posterior probability for $\Pr\{Y < 130 \mid Y < 131.88\}$ since 131.88 was the prevous world record. This probability, under our present analysis, comes to 0.347, which would imply that Chepngetich's record is not surprising at all.

However, for subsequent discussion, it seems that conditioning on breaking the previous world record is of less interest than the direct question of how likely it is that a woman runner could achieve a time less that 2 hours, 10 minutes, and our main point is that even the probability of 0.011 is large enough to make it seem a plausible result. In the following discussion, we shall consider a number of alternative variants on the analysis.

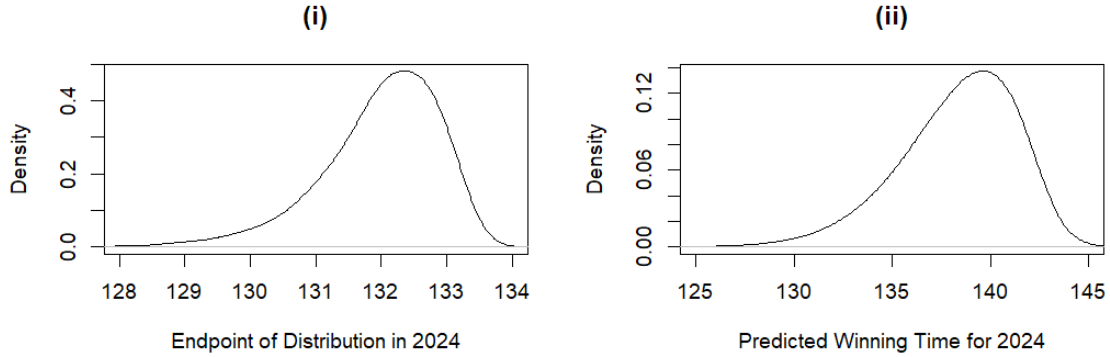

Figure 4: Posterior density plots, model (b) fitted to data from 1998–2023, for (i) estimated endpoint of the distribution in 2024, (ii) predicted winning time of the race in 2024.

3.2 Bayesian analyses under model (c)

The corresponding analysis under model (c) comtains even stronger evidence that Chepngetich's record is fully plausible. This model has five parameters β_0, σ, ξ, β_1 and β_2; we have run MCMC calculations under this model and produced figures similar to Figures 1 and 2; these will not be reproduced here. Instead, we jump straight to the analysis of the endpoint and predictive distribution, again taking 2024 as the target year conditioning on all the data up to 2023. The resulting posterior density plots are in Figure 5.

In the case of plot (i), 98% of the posterior distribution lies to the left of 130 minutes, implying that a performance under 2 hours, 10 minutes is virtually certain sooner or later. However, the left hand end of the distribution stretches down near the men's world record (currently 2 hours, 0 minutes, 35 seconds), which few people in the running community would consider plausible for women's performances. Thus, there is some model sensitivity to take into account here, that model (c) may represent a stronger trend post-2017 than is in fact plausible. We could judge this point better if we fitted a wider range of models. The second analysis resulting in plot (ii) leads to a posterior probability of 0.25, that the target time of 2 hours, 10 minutes would be broken in 2024. This result makes Chepngetich's performance seem fully plausible but again begs the question of model sensitivity. Therefore, we now turn to a model sensitivity comparison, where we repeat the preceding analyses under a variety of alternative specifications of the statistical model.

3.3 Model sensitivity analysis

In the preceding analyses, we used likelihood ratio tests to argue that model (c) fits the data much better than model (b), and also that model (c) makes Chepngetich's record seem much more

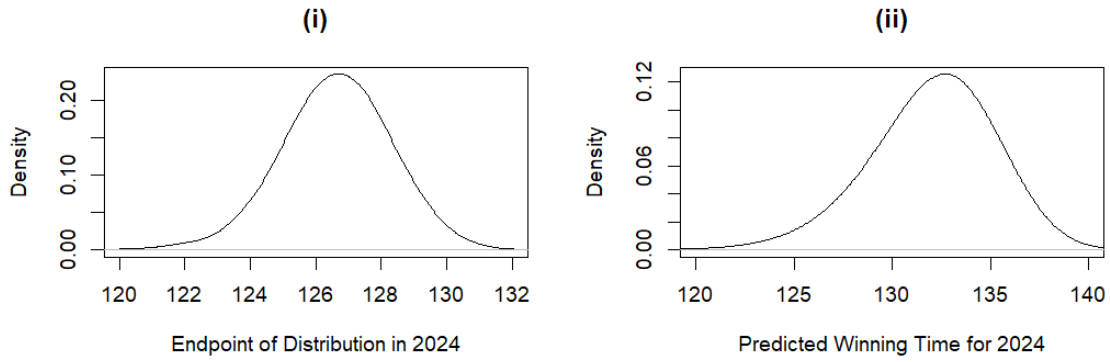

Figure 5: As in Figure 4, but for model (c).

plausible when assessed through Bayesian posterior analysis. However, model (c) also produces some results that seem implausible, such as the possibility of a women's marathon record several minutes faster than even Chepngetich's. This may be an artifact of the relatively small limited data used to estimate the trends. We also made a number of assumptions about the analysis itself, including the start year and the value of r (number of order statistics from each year included in the analysis). In addition, we noted that the changepoint model may in fact fit the data better with an earlier changepoint, such as 2010 (year 13 of the present analysis) instead of 2016. We can investigate these issues with a sensitivity analysis — essentially, repeating the whole analysis under different conditions.

Start Year	Endpoint				Winning Time			
	r=5	r=10	r=15	r=20	r=5	r=10	r=15	r=20
1998	0.17	0.034	0.031	0.091	0.004	0.011	0.013	0.014
2002	0.273	0.045	0.067	0.172	0.006	0.011	0.015	0.016
2005	0.211	0.07	0.121	0.213	0.015	0.022	0.024	0.024
2008	0.433	0.481	0.552	0.627	0.055	0.063	0.062	0.053
2005x	0.198	0.061	0.068	0.141	0.01	0.018	0.021	0.019
1998x	0.188	0.03	0.027	0.084	0.004	0.011	0.013	0.014
1998y	0.181	0.028	0.031	0.078	0.004	0.011	0.014	0.014

Table 4: Posterior probability that the endpoint or the winning time is under 130 minutes, based on model (b), for four values of r and different starting years. 2005x: analysis starting in 2005 but omitting 2007. 1998x: analysis with changepoint in 2013 instead of 2016. 1998y: analysis with changepoint in 2013 instead of 2016.

Tables 4 and 5 show some results from these comparisons. We used four values of r (5, 10, 15 and 20), and four different start years for the analysis (1998, 2002, 2005 and 2008), using both models (b) and (c). We also tried an analysis omitting the outlier year 2007, denoted 2005x in the tables. This was only done for start year 2005 because, in that case, 2007 was near the start of the series and therefore should have a stronger influence on the estimated trends. For start year 1998, we also tried an analysis in which 2013 or 2010 was the changepoint year; these analyses are

Start Year	Endpoint				Winning Time			
	r=5	r=10	r=15	r=20	r=5	r=10	r=15	r=20
1998	0.931	0.983	0.992	0.996	0.097	0.25	0.306	0.315
2002	0.967	0.988	0.996	0.996	0.168	0.324	0.392	0.34
2005	0.918	0.964	0.978	0.981	0.182	0.283	0.283	0.256
2008	0.596	0.73	0.771	0.847	0.079	0.116	0.132	0.125
2005x	0.92	0.973	0.976	0.98	0.151	0.262	0.297	0.258
1998x	0.686	0.722	0.796	0.874	0.037	0.069	0.092	0.106
1998y	0.477	0.469	0.584	0.656	0.029	0.057	0.065	0.066

Table 5: Same as Table 4, but based on model (c).

denoted 1998x and 1998y. For model (b), the changepoint year should not have any influence on the result so differences among years 1998, 1998x and 1998y should solely reflect sampling variability in the Monte Carlo calculations.

These tables show wide variations in the estimated probabilities (that either the endpoint of the distribution or the actual winning time should be less than 2 hours, 10 minutes, based on data up to 2023). Nevertheless, there are some consistent patterns. The estimated probabilities under model (c) are consistently higher than those under model (b), and even the smallest (around 0.004) are not so small that one could definitively say that Chepngetich's performance was not genuine. The comparison between the results for 2005 and 2005x (the latter omits 2007 from the calculations) shows that the influence of the "outlier year" 2007 is minimal (in all calculations except 2005x, data from 2007 has been kept in the analysis). However the last two rows (1998x and 1998y) do show that assuming an earlier year for the changepoint (2013 or 2010 respectively) does influence the result, in the direction that the estimated probability is smaller if we assume an earlier changepoint. However, even assuming 2010 as the changepoint year, the estimated probabilities for the winning time are between 0.029 and 0.066, which are not especially small. Overall, one can conclude that while Chepngetich undoubtedly has a very good run, her performance was not so exceptional that it should automatically raise doubts about its validity.

Start Year	Model (b)				Model (c)			
	r=5	r=10	r=15	r=20	r=5	r=10	r=15	r=20
1998	0.005	0.011	0.013	0.014	0.096	0.236	0.298	0.312
2002	0.006	0.011	0.015	0.016	0.154	0.3	0.37	0.332
2005	0.014	0.021	0.024	0.025	0.146	0.251	0.268	0.251
2008	0.041	0.055	0.058	0.052	0.069	0.112	0.131	0.125
205x	0.009	0.017	0.021	0.02	0.125	0.231	0.278	0.251
198x	0.004	0.011	0.013	0.014	0.037	0.068	0.093	0.106
198y	0.004	0.011	0.013	0.014	0.028	0.054	0.064	0.066

Table 6: Recalculating "Winning Time" probabilities of Tables 4 and 5, but under the alternative prior density

One other issue we should discuss is the influence of the prior distribution assumed for the GEV.

In the foregoing analyses, the unknown parameters are some subset of μ, β_0, β_1, β_2 and $\log\sigma$, and the prior density for each of these has been assumed uniform on $(-\infty, \infty)$, i.e. we have not imposed any limit on the range of these parameters for the MCMC sampler. This is consistent with common practice in Bayesian statistics, of a uniform prior for a location parameter or regression coefficients, and a $\frac{d\sigma}{\sigma}$ prior for the scale parameter (which is equivalent to a uniform prior on $\log\sigma$). The uniform prior is often criticized as over-simplistic or for having some undesirable properties, some alternatives being the Jeffreys prior [6] or some version of a reference prior [1]. The Jeffreys prior density is proportional to $|J|^{-1/2}$ where J is the Fisher information matrix, while a recent paper by Zhang and Shaby [15] has derived reference priors for the GEV distribution, but again, the starting point is the Fisher information matrix. The problem with applying these concepts in the present application is that the Fisher information is only defined when $\xi > -\frac{1}{2}$ [8, 13], whereas we have shown that in many cases the maximum likelihood estimator has $\hat{\xi} < -\frac{1}{2}$ and it turns out that much of the Bayesian posterior density is in this range as well. An alternative scheme, with less of a theoretical justification but some practical appeal, is to impose a non-uniform prior density on ξ, to restrict this parameter to a range consistent with previous analyses of similar data. Martins and Stedinger [7] made a proposal along these lines based on their experiences with hydrological data, and we have investigated some similar possibilities here. Specifically, it makes sense to restrict the range of ξ to $(-1, 0)$, and within this range, to impose a beta prior density of $|\xi|$. Table 6 shows the result of this process, applied to the winning time probabilities, for the prior density $\propto |\xi|^4(1-|\xi|)^4$ on $-1 < \xi < 0$, which has the effect of concentrating the prior mass near $\xi = -0.5$. As a side comment, we have tried alternative beta priors for $|\xi|$, some with considerable variation in the resulting estimated probabilities, but it would not seem realistic to make assumptions on the prior density of ξ that are completely inconsistent with the observational data. For the present discussion, the results in Table 6 are consistent with those in the right hand halves of Tables 4 and 5, and on that basis, we conclude that the choice of prior density does not have a great influence on the results.

4 Technical Details

Let $z^{(1)} \le z^{(2)} \le \ldots \le z^{(r)}$ denote the r best (i.e. smallest) running times in a given year. We make the following assumption (derived from extreme value theory) about the joint density:

$$f_r(z^{(1)}, \ldots, z^{(r)}) \;=\; \exp\left\{ -\left[1 + \xi\left(\frac{\mu - z^{(r)}}{\sigma}\right)\right]^{-1/\xi} \right\} \cdot \prod_{k=1}^{r} \left\{ \sigma^{-1}\left[1 + \xi\left(\frac{\mu - z^{(k)}}{\sigma}\right)\right]^{-1/\xi - 1} \right\}. \tag{2}$$

Except for a change of sign in each term of the form $\mu - z^{(k)}$, this is the same as formula (3.15) from [2]. The algebraic form of (2) assumes that each observation $z^{(k)}$, $1 \le k \le r$ is within the range of validity of the model; this means $1 + \xi\left(\frac{\mu - z^{(k)}}{\sigma}\right) > 0$; outside that range, the density is 0. To avoid constant repetition, we shall throughout assume these relationships are valid.

The formula is expressed in terms of parameters μ, σ, ξ, which are the location, scale and shape parameters of the Generalized Extreme Value (GEV) distribution, to which formula (2) reduces when $r = 1$, i.e. $\frac{\partial G}{\partial y}$ in equation (1). In simple analyses, we may treat μ, σ and ξ as constants, i.e. invariant from year to year. In more complicated models, as we propose here, these parameters

may depend on external covariates (including year of performance) in either a linear or nonlinear fashion.

Now let us extend the analysis to the case where there are T years of data, denoted by $t = 1, \ldots, T$, and the GEV parameters for year t are written $\mu_t,\ \sigma_t,\ \xi_t$. In this case the joint density for year t is given by (2) applied to the data $z_t^{(1)} \le z_t^{(2)} \le \ldots \le z_t^{(r)}$ for year t. In this case, the likelihood function L is given by multiplying the T joint densities of the form (2), and the negative log likelihood NLLH$= -\log L$ becomes

$$\text{NLLH} \;=\; \sum_{t=1}^{T} \left[\left\{ 1 + \xi_t \left(\frac{\mu_t - z_t^{(r)}}{\sigma_t} \right) \right\}^{-1/\xi_t} + r \log \sigma_t + \left(1 + \frac{1}{\xi_t} \right) \sum_{k=1}^{r} \log \left\{ 1 + \xi_t \left(\frac{\mu_t - z_t^{(k)}}{\sigma_t} \right) \right\} \right]. \tag{3}$$

For the models used in this paper, we assume σ_t and ξ_t are constants σ and ξ respectively, while μ_t is one of μ (model (a)) or $\beta_0 + \beta_1(t - t_0)$ (model (b)) or $\beta_0 + \beta_1(t - t_0) + \beta_2(t - t_0)_+$ (model (c)). Calculating the maximum likelihood estimators involves minimizing the NLLH with respect to the unknown parameters; for this we have used the functions `optim` and `nlm` in R [9]. Bayesian analysis used the adaptive Metropolis algorithm of [5].

5 Summary and Conclusions

This paper was motivated by Ruth Chepngetich's remarkable performance in the 2024 Chicago Marathon, establishing a new world record and becoming the first woman ever to break 2 hours, 10 minutes for the marathon, but also raising suspicions about possible cheating and in particular whether she might have been taking illegal drugs. We have taken data from the women's race of the Chicago Marathon from 1998–2023, initially using the 10 best performances each year but also considering analyses based on the best 5, 15 or 20 running times each year. We used extreme value theory to characterize the joint distribution of the best running times in each year, taking into account trends in the data, and thereby used both maximum likelihood and Bayesian methods to estimate the probability that either the "best possible performance" (endpoint of the distribution) or the actual winning time in 2024 would be better than 2 hours, 10 minutes, given all the data up to 2023. For the trend calculations, we have either assumed a linear trend (constant rate of improvement per year over the whole time period) or a changepoint model in which the slope of the trend is assumed to change after some intermediate time point known as the changepoint. The latter model is motivated especially by the introduction of "super shoes" in 2017 (initially the Nike Vaporfly model, but subsequently several modifications introduced by Nike and other running shoe companies). This suggested fixing the changepoint at 2016 (i.e. last year before the introduction of new shoes) though more detailed analysis suggested the true changepoint might be earlier than that, e.g. 2013 or 2010, suggesting other influences besides the new running shoes.

We varied a number of factors to allow us to assess sensitivities in the analysis, specifically, starting year of the analysis, number of running times considered for each year, year of the changepoint, and the prior distribution used for the Bayesian analysis. We also considered the effect of omitting the outlier year 2007, though it turned out that had very little effect. For the probability that the winning time in 2024 is better than 2 hours 10 minutes, we obtained estimates between 0.004 and 0.063 based on the linear trend model, and between 0.028 and 0.392 based on the changepoint

model. Although there is considerable variation among the different models considered, overall, these probabilities are not small enough to cast serious doubt on the validity of Chepngetich's performance. We conclude that, while an analysis of this nature can never prove or disprove whether an athlete was taking drugs, there is no evidence to suspect drug use based on the performance alone.

References

[1] J.-M. Bernardo. Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, 41(2):113–147, 1979.

[2] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag, London, https://doi.org/10.1007/978-1-4471-3675-0, 2001.

[3] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. *Bayesian Data Analysis (third edition)*. Chapman and Hall/CRC Press, New York, 2020.

[4] Joseph Guinness, Debasmita Bhattacharya, Jenny Chen, Max Chen, and Angela Loh. An observational study of the effect of Nike Vaporfly shoes on Marathon performance. *https://arxiv.org/pdf/2002.06105*, October 20, 2024.

[5] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7 (2):223–242, 2001.

[6] Harold Jeffreys. *Theory of Probability, First Edition*. Oxford University Press (second edition available from https://archive.org/details/in.ernet.dli.2015.2608/page/n5/mode/2up), 1939.

[7] Eduardo S. Martins and Jery R. Stedinger. Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3):737–744, 2000.

[8] P. Prescott and A.T. Walden. Maximum likelihood estimation of the parameters of the generalized extreme value distribution. *Biometrika*, 67:723–724, 1980.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[10] Michael E. Robinson and Jonathan A. Tawn. Statistics for exceptional athletics records. *Applied Statistics*, 44 (4):499–511, 1995.

[11] Richard L. Smith. Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86:27–43, 1986.

[12] Richard L. Smith. Letter to the editors: Statistics for exceptional athletics records. *Applied Statistics*, 46 (1):123–128, 1997.

[13] R.L. Smith. Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, 72:67–92, 1985.

[14] Jonathan A. Tawn. An extreme-value theory model for dependent observations. *Journal of Hydrology*, 101:227–250, 1988.

[15] Likun Zhang and Benjamin A. Shaby. Reference priors for the generalized extreme value distribution. *Statistica Sinica*, to appear, 2024.