Research Article

Brandon Onyejekwe* and Eric Gerber

Quantifying uncertainty in marathon finish time predictions

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: During a marathon, a runner's expected finish time is commonly estimated by extrapolating the average pace covered so far, assuming it is held constant for the rest of the race. Two problems arise when predicting finish times this way: the estimates do not consider in-race context that can determine if a runner is likely to finish faster or slower than expected, and the prediction is a single point estimate with no information about uncertainty. To address these issues, we implement a hierarchical Bayesian linear regression model that incorporates information from all splits in a race and allows quantification of uncertainty around the predicted finish times. Data from three marathons (Boston, New York, and Chicago) across 4 years (2021-2024) are utilized to establish the improved performance of this Bayesian approach over the traditional baseline method. Finally, we develop an app for runners to visualize their estimated finish distribution in real time.

Keywords: marathon; running; bayesian linear regression; uncertainty quantification

1 Introduction

A marathon is a long-distance road race where runners each complete 42.195 km (or 26.2 miles). Marathons, especially larger ones, can have tens of thousands of runners racing simultaneously, typically with a large number of spectators watching the race and cheering on the sidelines. As spectators usually remain at one spot along the course, they are typically only able to see a runner once. These spectators will often informally predict what time a given runner will finish the race, yet given their constrained view, there is very limited information to accurately predict when a runner will complete the multiple-hour race. Many marathons provide additional information using a chip in each runner's bib to track when runners complete certain portions of the race, often at every 5 km increment. These in-race splits are often reported live on a race's website or app, and are used to extrapolate the runner's finish time based on their pace. In this form of prediction, the pace is assumed to be held constant for the rest of the race, which brings rise to two significant problems.

First, the estimates do not consider the in-race context that can determine if a runner is likely to finish faster or slower. For example, marathon runners are commonly known to run slower during the second half of a race due to accumulated fatigue, and thus, the traditional prediction method will often underestimate the finish time. An individual runner's race pace can vary drastically due to other factors such as pacing strategy, preparation, and the impact of other runners in the field. Demographics, such as age and sex of the runner may also have some impact on the rest of their race. All of these effects are not properly captured via the traditional baseline prediction method of simple extrapolation.

The second major issue that arises from the commonly used prediction method is that the prediction is a single point estimate that has no additional information about the uncertainty surrounding a runner's finish time. Intuitively, we should feel more confident about a prediction made when a runner has completed 30 km (or approximately 75% of the race) than a prediction made when the runner has only completed 10 km (approximately 25%). Predictions generated under a more robust modeling framework can better reflect the uncertainty behind a point estimate, being narrower and more precise around an estimate as the runner gets closer to the finish of the race.

In this work, we utilize hierarchical Bayesian linear regression as an approach to address these two issues. This method incorporates additional information from the race to achieve more accurate predictions and allows uncertainty quantification around these predicted finish times.

Statistical and machine learning approaches have been applied to estimation of finish times for long-distance run-

***Corresponding author: Brandon Onyejekwe, Eric Gerber,** Northeastern University, Khoury College of Computer Sciences, Boston, MA, USA, e-mail: onyejekwe.b@northeastern.edu, e.gerber@northeastern.edu

ners in the hopes of addressing one or both of these issues. For example, [1] investigate linear regression and simple feed forward neural networks in predicting finish times for the Gothenburg Half Marathon, accounting for several factors. [2] found that K-nearest neighbors performed the best among several basic approaches for predicting the finish times of Boston marathon runners. Other papers have utilized case-based reasoning [3], artificial neural networks [4], and even ensembles of multiple ML methods [5]. These approaches unilaterally improved prediction over the simple extrapolation baseline in various contexts, but none addressed the second goal of effectively quantifying the uncertainty in the prediction.

This problem has also been tackled using Bayesian statistical methods with an eye towards quantifying uncertainty. In a blog post, [6] utilized nonparametric kernel smoothing methods on South Africa marathon data to demonstrate one approach to quantifying uncertainty. [7] also takes a nonparametric Bayesian approach, modeling the problem using dependent Dirichlet processes to analyze the effects of age and gender, examine marathon running patterns, and make finish time predictions. We seek to improve upon these predictions while more robustly examining uncertainty quantification. Their implementation of a nonparametric approach results in a search over an infinite model space, which may be more complex than our problem requires. We also focus more on having interpretable model results. In [8], a Bayesian model is implemented to predict finish times, but it specifically analyzes races of elite runners, whereas our work generalizes to a much wider range of runners.

Finally, rather than using primarily splits for prediction, there are various other factors that could be effective predictors of marathon finish times. In the survey [12], the authors compile and analyze a list of works across a 42 year span that involve equations for marathon finish prediction. They identify the most commonly used variables for prediction. In our work, we explore the effects of age [9, 10] and gender [11], as we have these data available. However, training information (prior race results, workout paces), lab tested measurements (VO2 max, heart rate and cadence data [13]), body measurements (BMI and body fat [14],) and race condition and strategy (packing [15], reference dependence [16, 17]) and many more are all additional potential factors.

2 Data

For this study we pulled data from the three World Marathon Majors in the United States: the Boston Marathon, the New York City Marathon, and the Chicago Marathon. Each event hosts tens of thousands of runners every April, November, and October, respectively [21]. For the Boston Marathon, most of the runners in the field of 30000 people qualify to compete by hitting notoriously difficult standards [1], while the remaining spots are for charity runners, who do not need to hit the qualification standards. The New York City and Chicago Marathons, each with over 50000 yearly runners, have lottery systems to select runners in addition to charity spots and time qualifiers. We scraped data for each marathon from the respective websites of each organization [18–20].

Our three datasets (Boston, New York, Chicago) each contain the name, age, gender, finish time, and in-race splits (5 km, 10 km, 15 km, 20 km, HALF, 25 km, 30 km, 35 km, and 40 km, all in seconds) for each finisher with complete race data from the respective marathon for each race held since the COVID-19 pandemic (2021-2024). Table 1 shows the number of runners for each marathon in each year, and Fig. 1 shows the distribution of finish times for each marathon. Fig. 2 shows the distribution of Boston Marathon finishes broken down by gender as well as by age.

By reformatting the data, we can get a more suitable set of features to perform our prediction task. For each split of the race, we can compute the average pace (total distance covered so far, divided by total time, in m/s) of an individual runner up until that split. This feature, which we call the *total_pace*, forms the basis of the traditional method, which assumes that pace will be held constant for the rest of the race. In Fig. 3, we directly compare true finish times with the extrapolated *total_pace* (which represents the traditional method's finish time estimates) at three different splits (10 km, 20 km, and 30 km). The red line represents the condition where the traditional method accurately predicts the finish time. For each of the 3 splits, most of the points lie above the traditional estimate line, which visually shows that extrapolating the total pace will typically underestimate the true finish time.

1 For example, the 2026 Boston Marathon cut-off times for 18-34 year old men and women are 2:55 and 3:25, respectively. Additionally, reaching the standard does not guarantee entry if too many people do so, which means runners often need to run faster than these times to qualify.

Race	Year	Total	Male	Female	Under 30	31-40	41-50	Above 50
Boston	2021	15121	7825	7296	1922	3578	4708	4913
	2022	24489	14079	10410	4014	6098	7418	6959
	2023	26028	14837	11169	4368	6406	7910	7344
	2024	25262	14467	10747	4207	6012	7503	7540
Chicago	2021	26753	14572	12181	5790	8731	7260	4972
	2022	39674	21049	18625	6702	12009	11896	9067
	2023	48785	25977	22808	8700	14490	14454	11141
	2024	52265	28197	24068	11435	16365	13872	10593
New York	2021	24628	13485	11128	5484	7814	6522	4808
	2022	46929	26224	20662	9422	13347	13351	10809
	2023	49406	27449	21868	11321	14451	13079	10555
	2024	54817	30424	24279	14585	16026	13309	10897

Tab. 1: Count of finishers for each marathon for each year, broken down by gender as well as by age groups

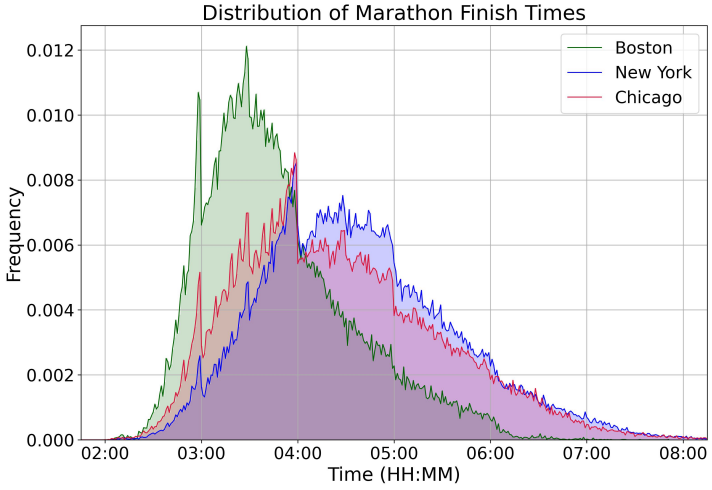

Fig. 1: Distribution of finish times for all finishers for each of the 3 marathons (Boston, New York, and Chicago) from 2021-2024

Another modification of the data was the addition of a *curr_pace* feature, which represents the pace of the most recently completed 5 km for the runner. At the 5 km mark, *total_pace* and *curr_pace* are the same, and for all other splits, the *curr_pace* is the runner's pace during the time between the immediate previous split and the current one. This feature can account for sudden or recent changes in a runner's pace, which may improve the accuracy of the final time prediction.

3 Methods

3.1 Model

The traditional method of extrapolating the current pace is used as a baseline (denoted **BL**). Here, the finish pace is assumed to be best approximated with *total_pace*. For the models we explore, we want to represent the finish pace as a linear combination of features. Thus, for a collection of N runners, we considered the following relationship.

$$y \sim \mathcal{N}(X\beta, \sigma) \tag{1}$$

where $y \in \mathbb{R}^N$ is a vector of each runner's finish pace (in m/s, which is then transformed into the finish time prediction, in seconds), and $X \in \mathbb{R}^{N\times D}$ is a feature matrix with D features (including bias column). The vector $\beta \in \mathbb{R}^D$ and value $\sigma \in \mathbb{R}^+$ are both parameters that need to be estimated. We explored the following candidate feature lists for X.

- **M1:** $[total\_pace]$
- **M2:** $[total\_pace, curr\_pace]$
- **M3:** $[total\_pace, curr\_pace, age, gender]$

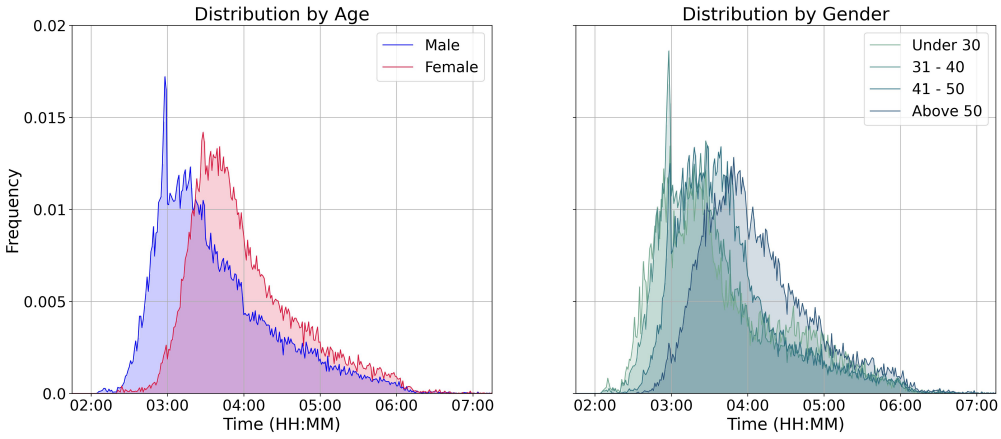

Fig. 2: Distribution of finish times, broken down by gender (left) and age (right) for the Boston Marathon.

Increasing the number of features used to predict the finish times should generally reduce the mean absolute error (MAE) and improve the accuracy of the predictions, as the model has more information to work with. We considered making predictions by directly incorporating all of the previous splits in the race to predict a future one, as this would contain the most prior information. However, we determined that this method has strong collinearity issues, as the previous splits of the runners are strongly correlated with each other. Furthermore, adding more features to a Bayesian model (especially a hierarchical one) leads to longer model runtimes and rapidly increases the number of parameters to estimate, which makes the marginal gains in prediction accuracy negligible. Thus, we opted to analyze in depth only the methods above and compare their performances to the baseline.

Bayesian regression is a useful method for providing information not only for point estimates, but also uncertainty [22]. By specifying prior distributions on the parameters of the linear regression model, and combining these priors with a likelihood function, we form a posterior distribution of possible finish times for a given individual. We use the posterior to generate both a point estimate (using the median of the distribution) and a credible interval: a central region of the distribution that we can use to quantify uncertainty.

In contrast to previous approaches, the hierarchical Bayesian linear regression [24] we implement is a compromise between two approaches: (1) pooling all the samples (corresponding to different stages) together to make parameter estimates and (2) running separate models for each stage. We assume that there is a relationship between the distance into the race, but don't want to make assumptions about what this relationship is, so our model allows the parameters to share information. Thus, we can modify Eq. 1 to the following:

$$y_s \sim \mathcal{N}(X\beta_s, \sigma_s) \tag{2}$$

where $s \in \{1, 2, ..., 8\}$ is the split of the race ($1 = 5$ km, $2 = 10$ km, ..., $8 = 40$ km). Here, $\beta_s \in \mathbb{R}^D$ and $\sigma_s \in \mathbb{R}^+$ become elements in the matrix β and vector σ, respectively.

For the Bayesian hierarchical framework, we set the following weakly-informative priors of a Gaussian on the fixed effects and a half-Cauchy on the uncertainty terms:

$$\beta_s \sim \mathcal{N}_D(0, I_D) \quad \text{and} \quad \sigma_s \sim \mathcal{HC}(0, 1) \tag{3}$$

3.2 Computation

We run the models using the R package **rstan**, a library for creating and running Bayesian models [23]. Sampling uses NUTS sampler, a Hamiltonian Monte Carlo method [25]. We ran four chains of 2000 samples (1000 warmup) each. The laptop used to run the model has a 2 GHz Quad-Core Intel Core i5 processor with 16GB RAM and 4 cores. Given these settings, there were no convergence issues.

We subsampled our training set to randomly select 2000 samples across the years 2021-2023. This was done to speed up the model training runtime while still having a representational and reasonably sized dataset to make inferences in our Bayesian framework. We tested how well this sample distribution matched the total population using the Kolmogorov-Smirnov test (H_0: samples come from same distribution, H_1: samples come from different distributions, $\alpha = 0.05$) and failed to reject the null hypothesis that they come from the same distribution (p=0.7216).

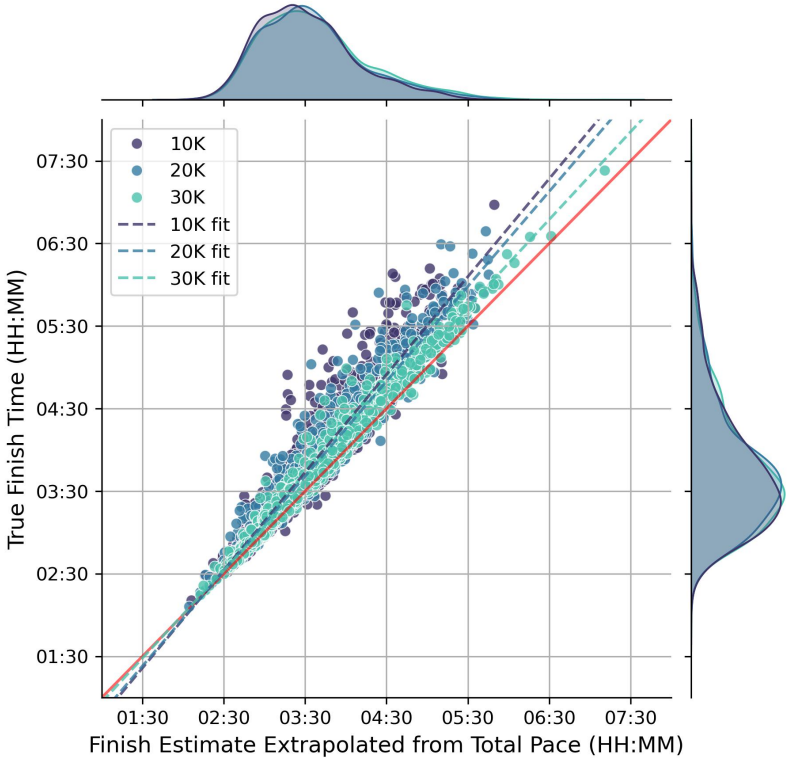

Fig. 3: For three stages of the race (10 km, 20 km and 30 km, in purple, blue, and cyan, respectively), the finish estimates extrapolated from the total pace so far (x-axis) are compared to the actual finish times (y-axis).

The models M1, M2, and M3 took roughly 50 seconds (16 parameters), 500 seconds (24 parameters), and 600 seconds (32 parameters) respectively. We also took 20000 samples for each test set and similarly tested to see how well it fit the population distribution (p=0.9158).

4 Results

To proceed with analysis, we partitioned the data into 3 training sets (one for each marathon) of runners from years 2021-2023, and 3 corresponding test sets of runners from the year 2024. For brevity, the results in the following section are specifically from analysis using the Boston data set. Tables showing the equivalent analyses from Chicago and New York are provided in the Appendix. Discussion of the major points of comparison between these three marathons can be found in Section 4.4. The decision to focus on M2 is explained in detail in Section 4.3.

4.1 Prediction Errors

As shown in Table 2, all three of our models, when applied to the Boston data, have similar MAE values at most splits of the race. The hierarchical models improve upon the baseline method at all levels, and significantly outperform it in the beginning and middle stages of races, as shown by the improvement percentages. It is especially important to have better finish estimates earlier in the race, as there is the greatest amount of uncertainty at these stages. The gap in MAE values decreases between the baseline model and our models in the latter stages of races (there is a gap of less than one minute between our models and the baseline model at 40 km compared to a gap of 6-8 minutes at 25 km). This makes sense because the baseline model also benefits from decreased uncertainty in the latter stages of the race. This demonstrates that the overall pace alone is a strong estimator of the true overall finish pace when the race is almost done.

As an example, consider the 15 km split. We see that the model predictions, on average, are roughly 8-10 minutes closer to the actual finish time than the baseline predictions. When looking at our use case, this jump in performance is considerable, as users will benefit from

	MAE				% Improve from BL		
Distance	BL	M1	M2	M3	M1	M2	M3
5K	24.431	15.391	15.392	15.408	0.37	0.37	0.369
10K	22.891	14.079	13.046	12.94	0.385	0.43	0.435
15K	20.834	12.073	10.879	10.82	0.42	0.478	0.481
20K	18.622	10.731	9.363	9.215	0.424	0.497	0.505
25K	14.777	8.378	6.939	6.809	0.433	0.53	0.539
30K	10.392	6.508	4.446	4.412	0.374	0.572	0.575
35K	5.23	3.758	2.574	2.652	0.281	0.508	0.493
40K	1.253	1.042	0.765	0.745	0.168	0.389	0.406
Overall MAE	14.828	9.003	7.929	7.878			
Overall R^2	0.787	0.897	0.903	0.904			

Tab. 2: Boston Marathon MAE at different splits of the race for the traditional method and the Bayesian linear regression models (M1, M2, and M3). Percent improvement from the traditional method for each model is also included.

having a much stronger prediction at a point in the race where there is high uncertainty.

Across all stages, M2 performs significantly better than M1, although the gap is small compared to the gap between those models and the traditional one. Intuitively, this makes sense because having *curr_pace* as an additional predictor gives enough information to get a slightly better estimate. Notably, M2 appears to have consistently better percentage improvements in the latter stages, while the percentage improvement of M1 dips. Thus, there is a greater effect in having *curr_pace* as an extra feature at those stages. While M3 generally shows a consistent improvement over M2, this is not true at all splits of the race and the overall improvement is negligible. We further analyze the effects of the extra features in M3 (age, gender) in Section 4.3.

We further breakdown the model performance via different groups of runners within the test set. In Fig. 4 the data is divided into 4 equally partitioned finish groups (Q4 being the fastest quarter of finishers, while Q1 being the slowest quarter of finishers). We then show the MAE values as bars for each group for both BL and M2. This breakdown highlights how the prediction errors vary depending on how quickly you run. Generally, the slower runners have higher prediction errors, which makes sense because there is more variability in possible finishes with slower paces.

We chose quartiles as a way to partition the groups because we felt it provided an intuitive way of investigating the effects of the predictors. Other methods of assessing how the test results vary based on the predictor values are possible.

We next categorize runners by their age and gender in Fig. 5. Runners are grouped by both their gender and their "age groups" [2]. The runners were divided into groups according to their age: G1 for under 30, G2 for 31-40, G3 for 41-50, and G4 for above 50. Men generally have slightly higher overall prediction errors than women for both BL and M2, especially at the beginning splits (there is a roughly 3 minute gap at 5 km that diminishes at the further splits). We can also visually see that the oldest age group (G4, in dark blue) has significantly higher BL and M2 MAE values at the beginning of races for both men and women.

4.2 Uncertainty Quantification

The key benefit of our approach lies in the ability to accurately quantify the uncertainty in the runners finish time using credible intervals of the posterior predictive distribution. When passing in a runner's feature predictors at a given distance into one of the models, we can create a p% credible interval $[t_1, t_2]$ such that the true finish time falls between t_1 and t_2 p% of the time.

To validate the credible intervals generated from the models, we check how well they fit with our assumptions. Specifically, we examine the credible interval sizes. On average, we expect the credible interval sizes to decrease as one gets further into the race, which fits with our intuition that one should be more certain of the estimate as they get closer to finishing. We also expect that a p% credible interval will be larger than a q% credible interval if $p > q$. Table 3 shows that these assumptions generally hold for 50%, 80%, and 95% credible intervals.

2 Age group used in this context is distinct from how age groups are used in these marathons, which generally are in 5 year increments.

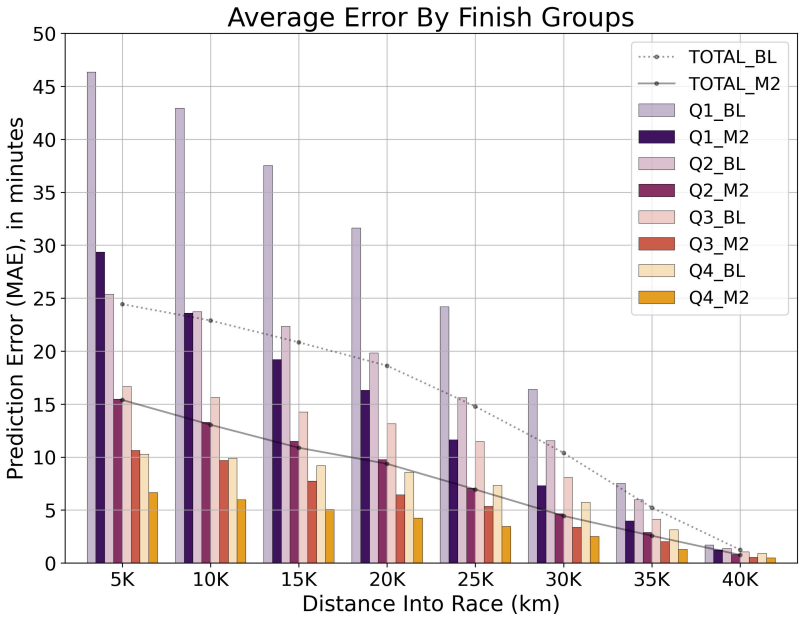

Fig. 4: MAE at different splits of the race for the traditional method (lighter shades) and M2 (darker shades), broken down for different finishing groups, differentiated by color. The dotted and solid lines are the overall MAE for traditional method and M2, respectively.

Each average interval size decreases and converges towards 0 as the race progresses and gets closer to finishing, and 50% intervals are narrower than 80% intervals, which are narrower than 95% intervals. While M2 and M3 both have roughly equal interval sizes, they are both consistently smaller than those for M1, showing that including the *curr_pace* helps to explain some variability in finish times across the race splits.

We also want to see if, for a given p% interval, approximately p% of runners truly finish within that interval. This gives us an approximation to our true goal that an individual's finish time has an p% chance of being within that predicted interval. Table 3 shows the proportions of intervals that contain the true value across different stages of the race for each model. The proportions are roughly around the expected proportions of 50%, 80%, and 95%. Note that, despite narrower intervals, M2 and M3 tend to contain a larger proportion of the true finish times, once again pointing to the utility of including additional information.

4.3 Model Selection

In selecting a model, it is important to balance performance and simplicity. In this section, we evaluate if the improvement we see from the M1, M2, and M3 models over baseline are significant, and if there are significant differences between the models themselves. We tested this using the Kolmogorov–Smirnov (KS) test (H_0: same distribution, H_1: different distributions, $\alpha = 0.05$). We fail to reject the null hypothesis that M2 and M3 come from the same distribution (p=0.2577). However, each of the other five pairwise comparisons involving BL, M1, M2, and M3 resulted in rejected null hypotheses (p<0.001 for all) [3]. Visually, the histograms of errors overlayed for all four predictions aligned with the results of the test: M2 and M3 appeared nearly the same, but each were significantly different from both M1 and BL. Additionally, the M3 versions of Fig 4 and Fig 5 looked nearly identical to the M2 versions, showing that the result plot broken down by age and gender (M3's two additional features) is not significantly changed by incorporating them into the model. Combining all of this, we conclude that M3 is not significantly different enough from M2 to justify its use, despite the increased complexity and marginally better MAE values. This suggests that any effect of age and gender is already adequately captured by the runner's total and current pace in the race.

Tab. 4 shows the posterior fitted parameters for M2 applied to Boston Marathon data. We notice that for far-

3 These results held for the Cramer-von Mises test (CVM) and the Anderson-Darling test (AD) as well. For the M2 and M3 comparison, $p = 0.1996$ for CVM and $p = 0.25$ for AD, and for all other comparisons, $p < 0.001$ for both

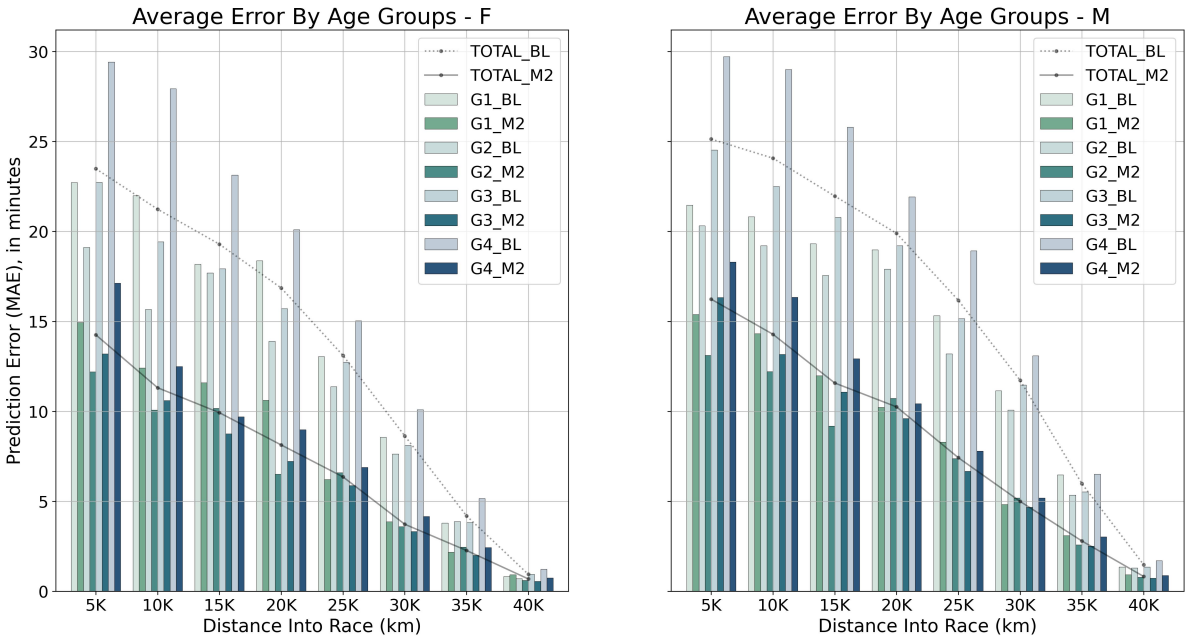

Fig. 5: MAE at different splits of the race for the baseline method (lighter shades) and M2 (darker shades), broken down for different age "groups", differentiated by color. The dotted and solid lines are the overall MAE for traditional method and M2, respectively. The left plot is for female runners, while the right plot is for male runners.

	50%			80%			95%		
Distance	M1	M2	M3	M1	M2	M3	M1	M2	M3
Credible Interval Sizes									
5K	21.328	21.331	21.378	40.839	40.859	40.923	63.362	63.387	63.509
10K	16.925	15.801	15.853	32.328	30.16	30.271	49.909	46.524	46.678
15K	16.761	15.238	15.13	32.022	29.066	28.868	49.427	44.831	44.493
20K	15.084	13.508	13.369	28.793	25.766	25.5	44.374	39.674	39.249
25K	13.116	10.939	10.682	25.009	20.852	20.35	38.489	32.05	31.258
30K	8.243	5.758	5.697	15.692	10.959	10.838	24.086	16.805	16.607
35K	5.553	3.956	3.873	10.578	7.53	7.366	16.211	11.546	11.283
40K	1.906	1.162	1.159	3.624	2.208	2.203	5.548	3.381	3.375
Proportion of True Finish Times Within Interval									
5K	0.504	0.502	0.506	0.763	0.767	0.767	0.896	0.894	0.895
10K	0.463	0.45	0.457	0.707	0.712	0.712	0.849	0.855	0.859
15K	0.499	0.502	0.508	0.756	0.77	0.772	0.883	0.891	0.891
20K	0.505	0.53	0.536	0.754	0.783	0.784	0.87	0.901	0.904
25K	0.547	0.573	0.574	0.791	0.818	0.806	0.898	0.912	0.914
30K	0.474	0.516	0.523	0.728	0.754	0.76	0.846	0.872	0.869
35K	0.536	0.589	0.568	0.796	0.816	0.809	0.888	0.907	0.904
40K	0.606	0.567	0.564	0.865	0.799	0.808	0.934	0.908	0.915

Tab. 3: Boston Marathon average credible interval sizes and the proportion of true finish times falling within credible intervals at each split of the race for the three model fits.

ther into the race, the total_pace becomes more important for the overall prediction than the curr_pace. We also notice the σ means decrease the farther into the race, as expected.

Examining the 5K parameter estimates for total_pace and curr_pace, we see that the means and standard deviations of total_pace are similar to those of curr_pace. As these two values are the same at this split when passed

	α - **Intercept**		β_1 - **total_pace**		β_2 - **curr_pace**		σ - **sigma**	
Distance	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0	-0.2545	0.0857	0.5109	0.7172	0.5048	0.7173	0.2178	0.0095
1	-0.2189	0.0614	-0.3803	0.1947	1.3891	0.1897	0.1549	0.0069
2	-0.1268	0.0604	-0.2213	0.146	1.2118	0.141	0.1445	0.0067
3	-0.128	0.0488	0.1813	0.0923	0.8251	0.0883	0.1233	0.0056
4	0.0304	0.0428	0.2169	0.0676	0.7508	0.0631	0.0999	0.0046
5	0.0697	0.023	0.5625	0.0285	0.4159	0.0259	0.0536	0.0026
6	0.0323	0.0139	0.8002	0.0135	0.1917	0.0121	0.0361	0.0016
7	0.012	0.0036	0.9359	0.0031	0.0606	0.0029	0.0108	0.0005

Tab. 4: Boston Marathon posterior parameter value estimates for M2

into the data, we expect this. We note the roughly equivalent MAE performance between this model at 5K with the performance of M1 (which only uses total_pace).

4.4 Marathon Comparison

The corresponding tables for New York and Chicago showing MAE values and credible interval information are included in the appendix A. Tables 2, 5, and 8 show that there is significant percent improvement between the models and the baseline for all three marathons. The credible intervals and parameter estimates for New York and Chicago generally appear to fit our expectations in the same ways as the Boston ones.

Looking back at Fig. 1, we can see that the Boston Marathon, as a whole, has a faster collection of runners than the other two marathons. This essentially results from the selection of runners, as Boston does not have a lottery that allows anyone to be able to run (only time qualifiers or charity runners can enter and run). This distribution difference affects the prediction accuracy of the disparate models. The Boston credible interval sizes also appear to be smaller overall compared to the corresponding New York and Chicago intervals. Additionally, races have different qualities (different courses, amount and placement of hills, typical race-day weather, levels of cheering support, etc.), which are all important factors.

5 Application

We develop an application to display how the M2 can be used to make predictions for a marathon race in real time. The *My Plot* tab of the app can be used to simulate or track a race; a user can select their marathon and sequentially enter in splits (in increments of 5 km) and the app will dynamically compute and update the displayed finish time statistics. One output is a plot (Fig. 6) which shows the predicted finish time probability distributions at different stages of the race. The centers of each curve represent the most probable finish times at that point of the race, and seeing multiple distributions together visually shows how the prediction changes over time. A narrower, taller distribution represents more precise predictions and narrower credible intervals. The other output is a table showing the median finish time prediction, as well as credible intervals (50%, 80%, and 95%) for each stage of the race. This tab can be a helpful tool for runners, coaches, and spectators to understand the distribution of possible finish times as the race is occurring.

6 Conclusion and Future Work

Bayesian linear regression can be used to address the issues present in the traditional method of estimating marathon finish times. It benefits from significantly improved point estimates by taking into account the context of in-race splits, while simultaneously providing additional context around the estimate with credible intervals to provide a sense of uncertainty.

Feature selection was an important consideration in addressing the problem, as the goal was to have a model that significantly improved the baseline method, but was still simple and interpretable. Adding too many features (such as using all prior splits, for example) would make the model harder to interpret, and significantly increase the time it took to fit. We also considered accounting for multiple observations by a single runner, but keeping track of this would similarly increase the complexity and number of parameters intractably. However, there are many features that we dont have access to, but are strong indicators of marathon performance that would likely

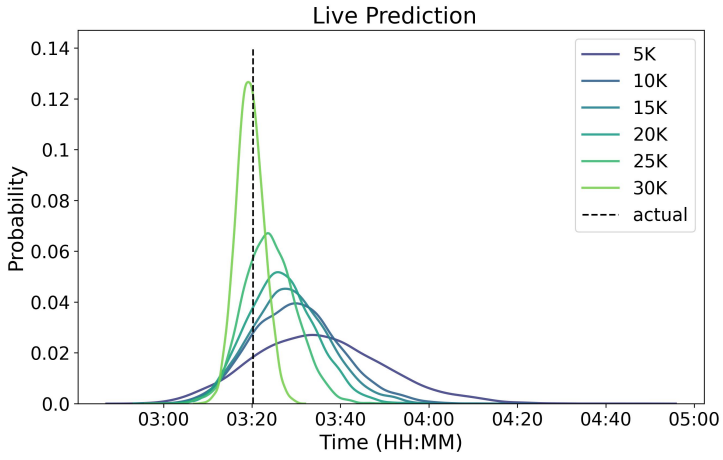

Fig. 6: Plot generated by the application after a runner inputs times. The probability distributions for finish times at each split of the race are shown.

result in better model estimates if they were incorporated. In future work, we could better quantify the effects of feature selection on the overall MAE of models.

The application only implements our model of choice (M2) for simplicity. We decided that implementing multiple models could be overwhelming and unnecessary for the user, distracting from our goal of providing fast and accurate information. However, by incorporating all of our models, there could be a visual comparison between them and could possibly reveal nuanced differences in the models if they give vastly different results for the same user. There is a subset of individuals in our test set where, for example, M1 had a better prediction than M2, so future work would involve examining that further and trying to find reasons why this affects some runners.

Despite establishing that our samples were representative of the population, sampling our training set has limitations in that it only allows the model to learn from a subset of the individual runners, not all of them. Similarly, sampling our test set means that we only evaluate the model performance on a subset of the data. We would like to further examine the effects of this sampling on our results in future work. We decided to exclude runners who did not finish the race or do not have some or all intermediate splits recorded, as we wanted to focus solely on runners who have all the information we want to use for prediction. Expanding from a complete case analysis to analyzing these runners (say, using a censored model) may help us uncover nuances in our analysis, such as reasons a runner dropped out of the race or a better understanding of the effects of gender and age.

The model described in this paper and the resulting application were developed to predict marathon finish times specifically for these three marathon races. Each of our three marathons had different parameter estimates, and changing the dataset to the results of a different marathon would alter the predictions to make the model more applicable to that specific race. We would like to do further analysis on how specific marathons affect the parameter estimates in future work. We also noticed differences when training the models on different years. We know, for example, that weather can drastically change between years, which can affect how the runners pace and finish. We also considered modeling the specific marathon as a hierarchical variable (similar to how we modeled splits), as this would allow us to build a single model where the parameters share information from the marathons. While this would significantly increase the number of parameters (especially if data from more marathons are included too), we would like to examine the performance of such a model in future work.

Finally, while our work focuses on predicting finish times specifically, the prediction task can even be adapted towards different goals. For example, the model and data can be modified to predict when a runner will cross a specific split in the race (say, the 30 km mark) instead of the finish. This can be helpful for a spectator stationed at 30 km that wants to know when a specific runner will pass by that point.

A Appendix

A.1 New York and Chicago Marathon Tables

The New York marathon results largely mimic those of the Boston marathon. We note the similar improvement in performance in terms of Overall MAE and R^2 in Tab. 5 for M2 and M3 above M1 and the baseline model. Note that in Tab. 6, we see that the proportion of true finish times within the various confidence levels is higher for the New York marathon than the Boston marathon at several points of the race for each of the three models investigated. This might have something to do with the New York marathon times being slightly less right skewed than the Boston marathon distribution of finish times. <Also compare Tab. 7 to Boston; maybe also reference Chicago results in the next section>.

	MAE				% Improve from BL		
Distance	BL	M1	M2	M3	M1	M2	M3
5K	21.426	19.385	19.397	19.19	0.095	0.095	0.104
10K	20.437	16.328	15.667	15.425	0.201	0.233	0.245
15K	19.057	16.443	13.248	13.228	0.137	0.305	0.306
20K	16.88	12.371	9.745	9.574	0.267	0.423	0.433
25K	12.0	9.722	7.118	7.026	0.19	0.407	0.414
30K	9.178	6.893	4.727	4.74	0.249	0.485	0.484
35K	4.945	4.142	2.578	2.571	0.162	0.479	0.48
40K	1.156	1.142	0.753	0.743	0.012	0.349	0.357
Overall MAE	13.149	10.819	9.167	9.076			
Overall R^2	0.871	0.927	0.937	0.937			

Tab. 5: New York Marathon MAE at different splits of the race for the traditional method and the Bayesian linear regression models (M1, M2, and M3). Percent improvement from the traditional method for each model is also included.

	50%			80%			95%		
Distance	M1	M2	M3	M1	M2	M3	M1	M2	M3
Credible Interval Sizes									
5K	37.266	37.229	37.185	71.831	71.817	71.704	112.984	112.968	112.866
10K	27.496	26.354	26.08	52.677	50.479	49.942	81.771	78.317	77.437
15K	27.684	23.164	23.064	53.035	44.272	44.081	82.397	68.513	68.229
20K	24.823	20.367	20.076	47.485	38.914	38.34	73.535	60.079	59.143
25K	15.794	13.569	13.364	30.108	25.886	25.462	46.332	39.799	39.143
30K	11.745	8.593	8.389	22.371	16.353	15.967	34.351	25.077	24.459
35K	7.89	5.783	5.808	15.024	11.007	11.045	23.038	16.885	16.931
40K	2.171	1.682	1.653	4.128	3.197	3.143	6.324	4.896	4.814
Proportion of True Finish Times Within Interval									
5K	0.468	0.466	0.474	0.806	0.806	0.81	0.971	0.967	0.971
10K	0.412	0.419	0.433	0.72	0.735	0.741	0.941	0.941	0.94
15K	0.382	0.45	0.45	0.743	0.797	0.796	0.968	0.962	0.962
20K	0.496	0.572	0.585	0.805	0.856	0.86	0.971	0.963	0.961
25K	0.419	0.518	0.526	0.734	0.826	0.813	0.937	0.96	0.956
30K	0.441	0.537	0.524	0.766	0.833	0.805	0.943	0.945	0.94
35K	0.498	0.667	0.672	0.814	0.899	0.899	0.961	0.964	0.963
40K	0.534	0.676	0.678	0.836	0.907	0.903	0.954	0.971	0.968

Tab. 6: New York Marathon average credible interval sizes and the proportion of true finish times falling within credible intervals at each split of the race for the three model fits.

	α - **Intercept**		β_1 - **total_pace**		β_2 - **curr_pace**		σ - **sigma**	
Distance	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0	-0.0212	0.0873	0.4674	0.7087	0.4399	0.7088	0.2407	0.0113
1	-0.0668	0.061	0.0883	0.1822	0.837	0.1795	0.1701	0.0079
2	-0.0443	0.0527	-0.1886	0.133	1.1318	0.1284	0.1546	0.0069
3	0.1877	0.0493	0.137	0.0811	0.7512	0.0764	0.1365	0.0063
4	0.0686	0.0319	0.5515	0.0492	0.3998	0.0462	0.0917	0.0042
5	0.0656	0.0184	0.6079	0.0246	0.3484	0.0229	0.0581	0.0025
6	0.0404	0.0134	0.7999	0.0129	0.1815	0.0126	0.0394	0.0018
7	0.0177	0.0035	0.9465	0.0038	0.0485	0.0037	0.0114	0.0005

Tab. 7: New York Marathon posterior parameter value estimates for M2

	MAE				% Improve from BL		
Distance	BL	M1	M2	M3	M1	M2	M3
5K	20.679	16.47	16.47	16.375	0.204	0.204	0.208
10K	18.328	14.183	13.634	13.481	0.226	0.256	0.264
15K	16.809	13.068	12.405	12.121	0.223	0.262	0.279
20K	15.742	10.999	9.632	9.327	0.301	0.388	0.408
25K	13.35	9.291	7.511	7.408	0.304	0.437	0.445
30K	9.231	6.639	4.887	4.856	0.281	0.471	0.474
35K	5.451	3.876	2.861	2.821	0.289	0.475	0.482
40K	1.149	1.106	0.861	0.838	0.037	0.25	0.27
Overall MAE	12.604	9.468	8.552	8.423			
Overall R^2	0.888	0.936	0.941	0.941			

Tab. 8: Chicago Marathon MAE at different splits of the race for the traditional method and the Bayesian linear regression models (M1, M2, and M3). Percent improvement from the traditional method for each model is also included.

	50%			80%			95%		
Distance	M1	M2	M3	M1	M2	M3	M1	M2	M3
Credible Interval Sizes									
5K	29.522	29.472	29.284	56.733	56.668	56.269	88.627	88.482	87.955
10K	26.067	25.511	25.648	49.989	48.892	49.184	77.724	76.022	76.486
15K	22.764	22.512	22.529	43.553	43.032	43.062	67.439	66.661	66.702
20K	19.201	15.818	15.549	36.676	30.175	29.676	56.615	46.479	45.698
25K	19.26	15.498	15.383	36.769	29.574	29.318	56.76	45.533	45.13
30K	12.543	9.315	9.147	23.899	17.736	17.402	36.731	27.211	26.665
35K	6.918	6.056	5.92	13.174	11.524	11.268	20.204	17.681	17.275
40K	1.845	1.424	1.413	3.509	2.706	2.687	5.376	4.146	4.117
Proportion of True Finish Times Within Interval									
5K	0.541	0.541	0.545	0.851	0.849	0.844	0.957	0.956	0.956
10K	0.543	0.56	0.566	0.86	0.863	0.87	0.96	0.96	0.964
15K	0.485	0.503	0.518	0.825	0.853	0.855	0.952	0.959	0.962
20K	0.498	0.503	0.513	0.827	0.818	0.814	0.941	0.923	0.921
25K	0.566	0.609	0.618	0.894	0.89	0.889	0.962	0.959	0.959
30K	0.544	0.616	0.609	0.855	0.868	0.868	0.954	0.944	0.946
35K	0.521	0.611	0.615	0.836	0.889	0.88	0.939	0.957	0.958
40K	0.532	0.585	0.595	0.806	0.829	0.832	0.929	0.937	0.934

Tab. 9: Chicago Marathon average credible interval sizes and the proportion of true finish times falling within credible intervals at each split of the race for the three model fits.

	α - **Intercept**		β_1 - **total_pace**		β_2 - **curr_pace**		σ - **sigma**	
Distance	Mean	Std	Mean	Std	Mean	Std	Mean	Std
0	-0.2776	0.0707	0.5195	0.7084	0.5142	0.7082	0.2155	0.0095
1	-0.1709	0.065	0.3148	0.2043	0.6924	0.1972	0.1852	0.0089
2	-0.2457	0.0459	0.549	0.1505	0.4719	0.1487	0.1611	0.0069
3	-0.0405	0.0361	0.2315	0.0725	0.7409	0.0696	0.1157	0.0054
4	0.0668	0.0339	0.2186	0.0617	0.7343	0.0576	0.1098	0.0047
5	0.0067	0.0198	0.6591	0.0242	0.3239	0.0222	0.0674	0.003
6	0.0131	0.0132	0.8712	0.0147	0.1138	0.0134	0.0441	0.002
7	0.0099	0.0033	0.9595	0.0032	0.0381	0.003	0.0103	0.0005

Tab. 10: Chicago Marathon posterior parameter value estimates for M2

References

[1] Johansson, M., Atterfors, J. and Lamm, J., 2023. Pacing Patterns of Half-Marathon Runners: An analysis of ten years of results from Gothenburg Half Marathon. *International Journal of Computer Science in Sport, 22*(1), pp.124-138.

[2] Hammerling, D., Cefalu, M., Cisewski, J., Dominici, F., Parmigiani, G., Paulson, C. and Smith, R.L., 2014. Completing the results of the 2013 Boston marathon. *PLoS One, 9*(4), p.e93800.

[3] Smyth, B., Lawlor, A., Berndsen, J. and Feely, C., 2022. Recommendations for marathon runners: on the application of recommender systems and machine learning to support recreational marathon runners. *User Modeling and User-Adapted Interaction, 32*(5), pp.787-838.

[4] Lerebourg, L., Saboul, D., Clemencon, M. and Coquart, J.B., 2023. Prediction of marathon performance using artificial intelligence. *International Journal of Sports Medicine, 44*(05), pp.352-360.

[5] Panwala, B. and Buch, S., 2024, October. Exploring Advanced Ensemble Learning Strategies in Machine Learning and Data Mining for Predictive Modeling of Marathon Running Time. In *International Conference on Advancements in Smart Computing and Information Security* (pp. 199-214). Cham: Springer Nature Switzerland.

[6] Collier, A.(2017) *Bayesian Marathon Predictions, Andrew B. Collier / @datawookie*, 28 February. Available at: https://datawookie.dev/blog/2017/02/bayesian-marathon-predictions (Accessed: 28 August 2025)

[7] F. Pradier, M., JR Ruiz, F. and Perez-Cruz, F., 2016. Prior design for dependent Dirichlet processes: An application to marathon modeling. *PloS one, 11*(1), p.e0147402.

[8] Patras, K., Predicting elite marathon performance: Medalists vs. non medalists for the 2023 BMW Berlin marathon.

[9] Hubble, C. and Zhao, J., 2016. Gender differences in marathon pacing and performance prediction. *Journal of Sports Analytics, 2*(1), pp.19-36.

[10] Hanley, B., 2016. Pacing, packing and sex-based differences in Olympic and IAAF World Championship marathons. *Journal of sports sciences, 34*(17), pp.1675-1681.

[11] Lehto, N., 2016. Effects of age on marathon finishing time among male amateur runners in Stockholm Marathon 1979–2014. *Journal of Sport and Health Science, 5*(3), pp.349-354.

[12] Keogh, A., Smyth, B., Caulfield, B., Lawlor, A., Berndsen, J. and Doherty, C., 2019. Prediction equations for marathon performance: a systematic review. *International journal of sports physiology and performance, 14*(9), pp.1159-1169.

[13] Berndsen, J., Smyth, B. and Lawlor, A., 2019, September. Pace my race: recommendations for marathon running. In *Proceedings of the 13th ACM conference on recommender systems* (pp. 246-250).

[14] Schmid, W., Knechtle, B., Knechtle, P., Barandun, U., Rüst, C.A., Rosemann, T. and Lepers, R., 2012. Predictor variables for marathon race time in recreational female runners. *Asian journal of sports medicine, 3*(2), p.90.

[15] Muñoz-Pérez, I., Castañeda-Babarro, A., Santisteban, A. and Varela-Sanz, A., 2024. Predictive performance models in marathon based on half-marathon, age group and pacing behavior. *Sport Sciences for Health, 20*(3), pp.797-810.

[16] Allen, E.J., Dechow, P.M., Pope, D.G. and Wu, G., 2017. Reference-dependent preferences: Evidence from marathon runners. *Management Science, 63*(6), pp.1657-1672.

[17] Markle, A., Wu, G., White, R. and Sackett, A., 2018. Goals as reference points in marathon running: A novel test of reference dependence. *Journal of Risk and Uncertainty, 56*(1), pp.19-50.

[18] *Boston Marathon Results* (2024) *Search Results | Boston Athletic Association*. Available at: https://www.baa.org/races/boston-marathon/results/search-results (Accessed: 28 August 2025).

[19] *Race results* (2024a) *New York Road Runners*. Available at: https://www.nyrr.org/tcsnycmarathon/results/race-results (Accessed: 28 August 2025).

[20] *Race results* (2024b) *Bank of America Chicago Marathon*. Available at: https://www.chicagomarathon.com/runners/race-results/ (Accessed: 28 August 2025).

[21] Dredge, M. (2025) *How to get into the World Marathon Majors, The Running Channel*. Available at: https://therunningchannel.com/how-to-get-into-the-world-marathon-majors/ (Accessed: 28 August 2025).

[22] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B., 1995. *Bayesian data analysis*. Chapman and Hall/CRC.

[23] Stan Development Team *Stan: Software for Bayesian Data Analysis* https://mc-stan.org/ (Accessed: 28 August 2025)

[24] *Regression models* (no date) *Stan Docs*. Available at: https://mc-stan.org/docs/stan-users-guide/regression.html#hierarchical-regression (Accessed: 28 August 2025).

[25] Hoffman, M.D. and Gelman, A., 2014. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J. Mach. Learn. Res., 15(1), pp.1593-1623.