**Research Article**

Brandon Onyejekwe* and Eric Gerber

# Quantifying uncertainty in marathon finish time predictions

https://doi.org/10.1515/sample-YYYY-XXXX
Received Sep 9, 2025; 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 simply a 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 in which runners complete 42.195 km (or 26.2 miles). Marathons, especially larger ones, can have tens of thousands of runners racing simultaneously, often with a large number of spectators watching the race and cheering on the sidelines. As spectators typically remain at one spot along the course, they are usually only able to see each runner once. These spectators will often informally estimate when a given runner will finish, yet given their constrained view, there is limited information to make accurate predictions on 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. This form of prediction assumes that the runner's average pace will be unchanged for the rest of the race, which gives rise to two significant problems.

The first major issue is that the estimates do not account for the in-race context that can determine whether 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, which means that the traditional prediction method will often underestimate the finish time. The race pace of an individual runner can vary due to other factors, such as their pacing strategy, preparation before the race, and the physical and mental impacts of other runners and spectators. A runner's demographics, such as their age and sex, 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 is that the prediction is simply a point estimate, with 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 estimate finish times for long-distance runners in the hopes of addressing one or both of these issues. For

***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

example, [1] investigates using linear regression and simple feed-forward neural networks to predict finish times for the Gothenburg Half Marathon, accounting for several factors. [2] found that K-nearest neighbors performed the best among several basic approaches to predict finish times for Boston marathon runners. Other papers have used 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 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] used non-parametric kernel smoothing methods on South Africa marathon data to demonstrate one approach to quantifying uncertainty. The researchers in [7] also take 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 have all been identified as 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 [18]. 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 [19–21].

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). Fig. 1 shows the distribution of finish times for each marathon, while Table 1 shows the number of runners by year, as well as the counts by gender and age. Fig. 2 shows the distribution of Boston Marathon finishes broken down by gender as well as by age. Generally, we see that males tend to have faster finish times than females, and younger runners tend to finish quicker than older ones.

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 this pace will remain 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 three 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.

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

**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 that 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.

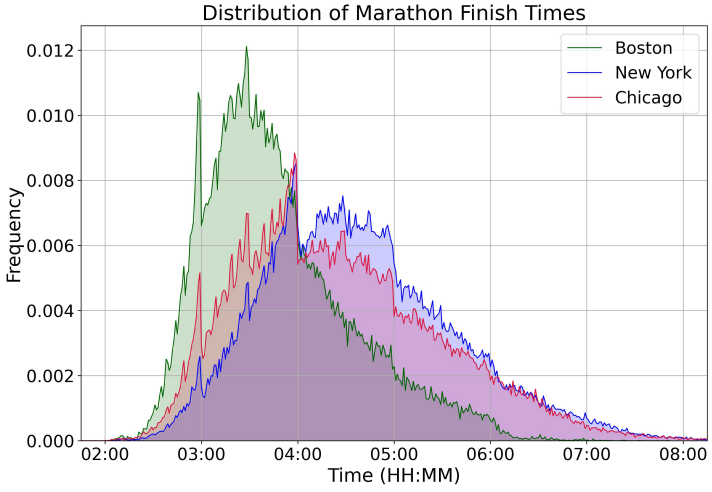


**Fig. 1:** Distribution of finish times for all finishers for each of the 3 marathons (Boston, New York, and Chicago) from 2021-2024

all other splits, *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**).This method assumes that the finish pace is 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 consider 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 the value $\sigma \in \mathbb{R}^+$ are both parameters that need to be estimated. We explore the following candidate feature lists for $X$.

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

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,

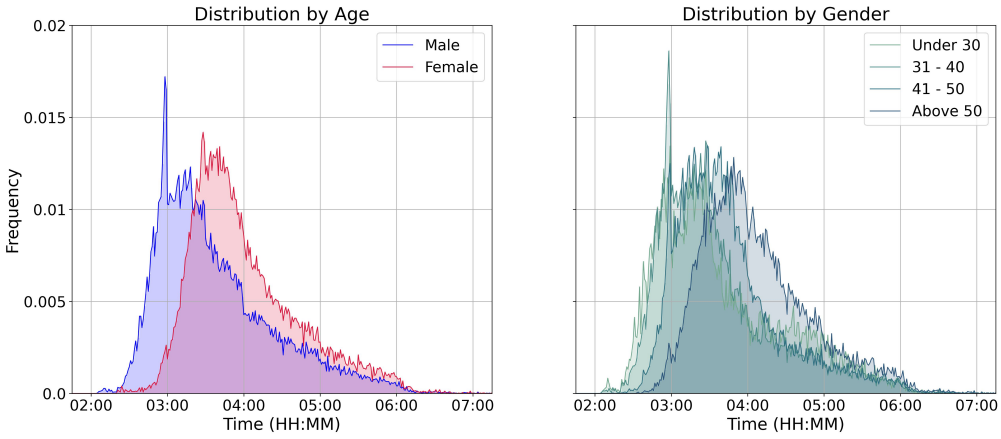


**Fig. 2:** Distribution of finish times, broken down by gender (left) and age (right) for the Boston Marathon.

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 for 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 then use this 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 other approaches, the hierarchical Bayesian linear regression [24] that we implement is a compromise between two approaches: (1) pooling all the samples (corresponding to different splits) together to make parameter estimates and (2) running separate models for each split. We assume that there is a relationship between the distance into the race and finish time, 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 $\beta$ and the vector $\sigma$, 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 used to create and run Bayesian models [23]. Sampling uses the NUTS sampler, a Hamiltonian Monte Carlo method [25]. We run 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 are no convergence issues.

We subsample our training set by randomly selecting 2000 samples from the years 2021-2023. This allows us to speed up the model training while maintaining a representational and reasonably sized dataset to make inferences in our Bayesian framework. We test how well this sample distribution matches the total population using the Kolmogorov-Smirnov test ($H_0$: samples come from the same distribution, $H_1$: samples come from different distributions, $\alpha = 0.05$) and fail to reject the null hypothesis that they come from the same distribution (p=0.7216). The models M1, M2, and M3 took roughly 60 seconds (24 parameters), 500 seconds (32 parameters), and 600

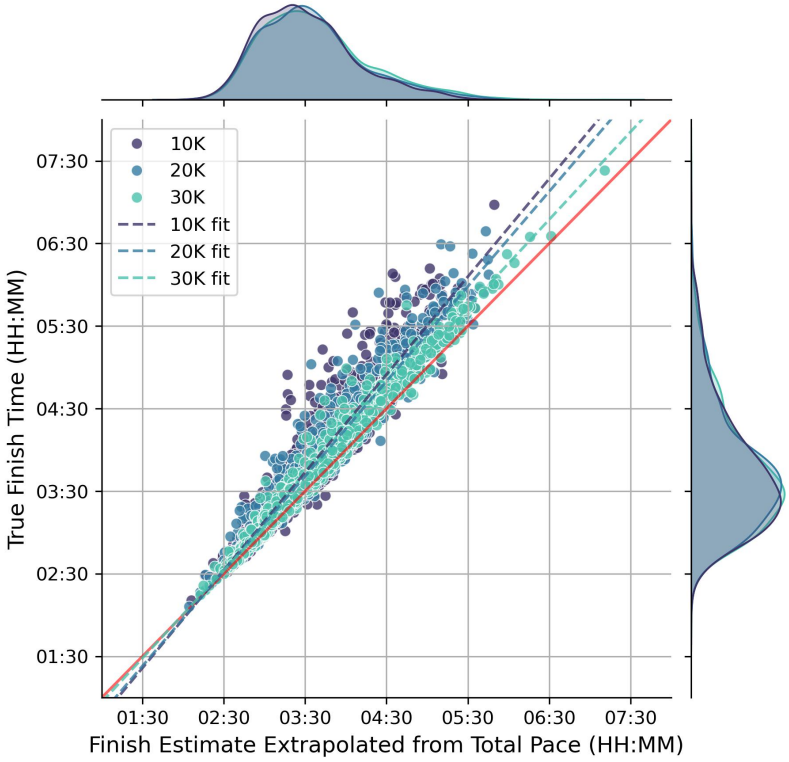


**Fig. 3:** For three stages of the race (10 km, 20 km and 30 km, in purple, blue, and cyan, respectively), the finish estimates extrapolated from the total pace so far (x-axis) are compared to the actual finish times (y-axis).

seconds (48 parameters), respectively. We also take 20000 samples for each test set and similarly test it to confirm that it fits the population distribution (p=0.9158).

# 4 Results

To proceed with analysis, we partitioned the data into three training sets (one for each marathon) of runners from years 2021-2023 and three corresponding test sets of runners from the year 2024. For brevity, the results in the following section are specifically from our analysis using the Boston data set. Tables showing the equivalent analyses from Chicago and New York are provided in the appendix. Discussion of the main 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 each race split. 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 most uncertainty at these stages. The gap in MAE values decreases between the baseline model and our models for the latter splits 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). We observe that the overall pace alone is a strong estimator of the true overall finish pace when the race is almost complete, which means that the baseline model also benefits from decreased uncertainty in the latter stages of the race.

As an example, consider the MAE values at 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 having a much stronger prediction at a point in the race where there is high uncertainty.

For all splits after 5 km, M2 performs significantly better than M1, although the gap is small compared to the gap between these models and the traditional one.

| | MAE | | | | % Improve from BL | | |
|---|---|---|---|---|---|---|---|
| Distance | BL | M1 | M2 | M3 | M1 | M2 | M3 |
| 5K | 24.431 | 15.382 | 15.382 | 15.406 | 0.37 | 0.37 | 0.369 |
| 10K | 22.891 | 14.075 | 13.047 | 12.942 | 0.385 | 0.43 | 0.435 |
| 15K | 20.834 | 12.073 | 10.882 | 10.819 | 0.42 | 0.478 | 0.481 |
| 20K | 18.622 | 10.727 | 9.366 | 9.219 | 0.424 | 0.497 | 0.505 |
| 25K | 14.777 | 8.379 | 6.935 | 6.816 | 0.433 | 0.531 | 0.539 |
| 30K | 10.392 | 6.513 | 4.447 | 4.411 | 0.373 | 0.572 | 0.576 |
| 35K | 5.23 | 3.757 | 2.573 | 2.651 | 0.282 | 0.508 | 0.493 |
| 40K | 1.253 | 1.042 | 0.765 | 0.745 | 0.168 | 0.389 | 0.406 |
| Overall MAE | 14.828 | 9.002 | 7.928 | 7.879 | | | |
| 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.

Intuitively, this makes sense because having *curr_pace* as an additional predictor gives enough information to get a slightly better estimate. In particular, M2 appears to have consistently better percentage improvements in the latter splits, while the percentage improvement of M1 dips. Thus, there is a greater effect in having *curr_pace* as an extra feature at these splits. Although M3 generally shows a consistent improvement over M2, this is not true for all splits of the race, and the overall improvement is negligible. We further analyze the effects of the extra features in M3 (age and gender) in Section 4.3.

We further break down the model performance on different groups of runners within the test set. In Fig. 4, runners are 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 based on the runner's speed. Generally, the slower runners have higher prediction errors, which makes sense because there is more variability in possible finishes with slower paces. We choose quartiles to partition the groups because we feel that they provide an intuitive way to investigate 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] shown in Table 1 (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 quantify the uncertainty in a runner's 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. 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 of M1, which shows that including *curr_pace* as a feature helps explain some variability in finish times and allows for more precise uncertainty estimates.

**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.

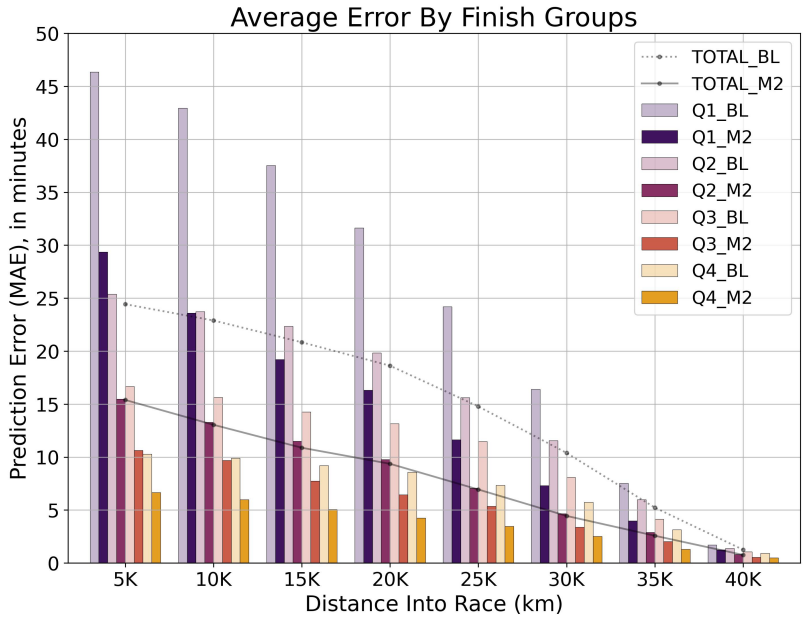


**Fig. 4:** MAE at different splits of the race for the traditional method (lighter shades) and M2 (darker shades), broken down for different finishing groups, differentiated by color. The dotted and solid lines are the overall MAE for traditional method and M2, respectively.

| | 50% | | | 80% | | | 95% | | |
|---|---|---|---|---|---|---|---|---|---|
| Distance | M1 | M2 | M3 | M1 | M2 | M3 | M1 | M2 | M3 |
| Credible Interval Sizes | | | | | | | | | |
| 5K | 21.322 | 21.324 | 21.392 | 40.837 | 40.829 | 40.953 | 63.343 | 63.342 | 63.568 |
| 10K | 16.928 | 15.79 | 15.857 | 32.34 | 30.126 | 30.264 | 49.908 | 46.482 | 46.69 |
| 15K | 16.773 | 15.235 | 15.125 | 32.043 | 29.069 | 28.864 | 49.472 | 44.831 | 44.491 |
| 20K | 15.063 | 13.516 | 13.372 | 28.755 | 25.78 | 25.5 | 44.31 | 39.698 | 39.242 |
| 25K | 13.125 | 10.929 | 10.699 | 25.023 | 20.843 | 20.384 | 38.503 | 32.04 | 31.331 |
| 30K | 8.245 | 5.764 | 5.705 | 15.7 | 10.97 | 10.853 | 24.109 | 16.817 | 16.63 |
| 35K | 5.558 | 3.96 | 3.876 | 10.583 | 7.534 | 7.374 | 16.223 | 11.551 | 11.292 |
| 40K | 1.908 | 1.162 | 1.158 | 3.628 | 2.208 | 2.202 | 5.555 | 3.38 | 3.374 |
| Proportion of True Finish Times Within Interval | | | | | | | | | |
| 5K | 0.502 | 0.502 | 0.507 | 0.763 | 0.765 | 0.768 | 0.893 | 0.894 | 0.894 |
| 10K | 0.462 | 0.452 | 0.458 | 0.708 | 0.71 | 0.711 | 0.847 | 0.855 | 0.859 |
| 15K | 0.498 | 0.5 | 0.508 | 0.756 | 0.77 | 0.773 | 0.884 | 0.892 | 0.891 |
| 20K | 0.506 | 0.529 | 0.535 | 0.755 | 0.782 | 0.784 | 0.87 | 0.901 | 0.903 |
| 25K | 0.549 | 0.574 | 0.573 | 0.793 | 0.817 | 0.805 | 0.896 | 0.912 | 0.914 |
| 30K | 0.474 | 0.516 | 0.524 | 0.727 | 0.755 | 0.76 | 0.845 | 0.871 | 0.87 |
| 35K | 0.537 | 0.592 | 0.569 | 0.794 | 0.816 | 0.81 | 0.888 | 0.908 | 0.904 |
| 40K | 0.606 | 0.564 | 0.567 | 0.865 | 0.798 | 0.81 | 0.935 | 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.

We also want to verify that, 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 a 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 the proportions for M1, M2, and M3 are roughly equal for each interval.

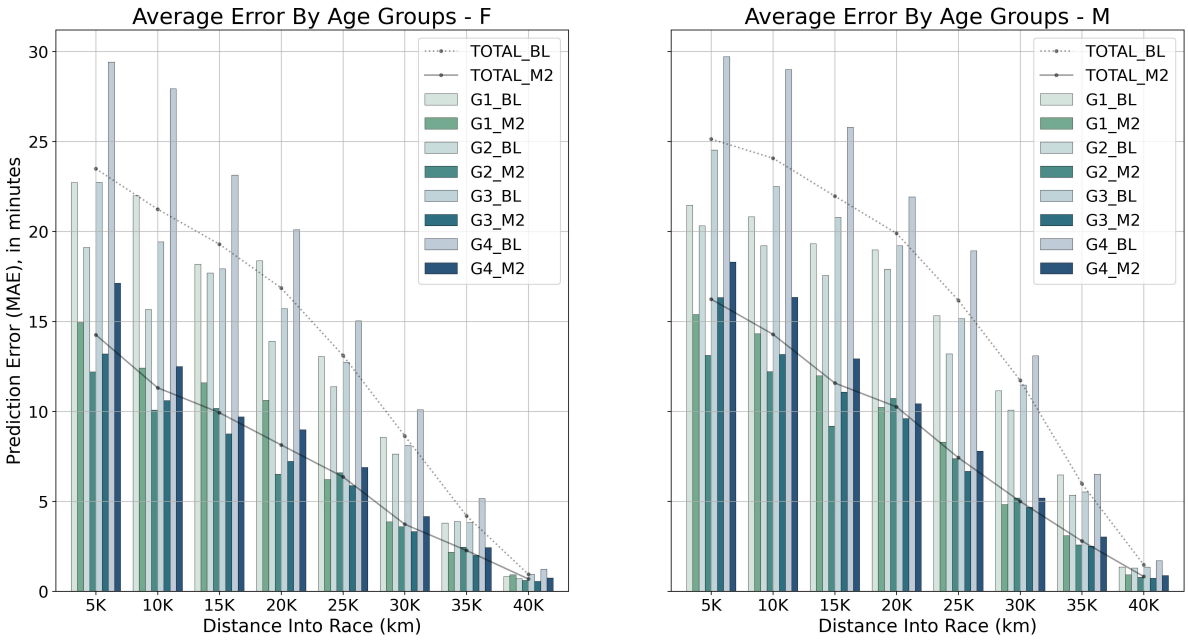


**Fig. 5:** MAE at different splits of the race for the baseline method (lighter shades) and M2 (darker shades), broken down for different age "groups", differentiated by color. The dotted and solid lines are the overall MAE for traditional method and M2, respectively. The left plot is for female runners,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

## 4.3 Model Selection

When selecting a model, it is important to balance performance and simplicity. In this section, we evaluate whether the improvements we see from M1, M2, and M3 over the baseline are significant, as well as whether there are significant differences between the three models. We test 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=0.2328). 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 overlaid for all four predictions aligned with the results of the test: M2 and M3 appeared nearly the same, but each was significantly different from both M1 and BL. Additionally, the M3 versions of Figs. 4 and 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 information, 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 farther distances into the race, the mean of the parameter *total_pace* increases, while the mean of the parameter *curr_pace* decreases. This fits with our intuition that the feature *total_pace* becomes more important for the overall prediction than the feature *curr_pace* as a race progresses. We also notice that the mean of the parameter $\sigma$ decreases the farther into the race. This is expected and reflects the increasing level of confidence in the model predictions, as well as smaller intervals. The standard deviations of all parameters (other than $\beta_1$ and $\beta_2$ for 5 km) are small, reflecting confidence that the model is fit well. Upon further examinination of the 5 km parameter estimates for *total_pace* and *curr_pace* ($\beta_1$ and $\beta_2$), we see that the means and standard deviations of both parameters are very similar. As these two values are the same at this split when passed into the data, we expect this result. We note the roughly equivalent MAE performance between this model at 5 km with the performance of M1 (which only uses *total_pace*).

**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, $p = 0.1916$ for CVM and $p = 0.25$ for AD, and for all other comparisons, $p < 0.001$ for both

| | $\alpha$ - **Intercept** | | $\beta_1$ - **total_pace** | | $\beta_2$ - **curr_pace** | | $\sigma$ - **sigma** | |
|---|---|---|---|---|---|---|---|---|
| Distance | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 5K | -0.2556 | 0.0813 | 0.5029 | 0.7151 | 0.513 | 0.7156 | 0.2176 | 0.0094 |
| 10K | -0.2197 | 0.0622 | -0.371 | 0.1972 | 1.3801 | 0.1918 | 0.1548 | 0.0068 |
| 15K | -0.1265 | 0.0592 | -0.2268 | 0.1461 | 1.2172 | 0.1404 | 0.1445 | 0.0066 |
| 20K | -0.1279 | 0.0506 | 0.1797 | 0.0937 | 0.8266 | 0.0894 | 0.1233 | 0.0057 |
| 25K | 0.0301 | 0.0427 | 0.2182 | 0.0684 | 0.7497 | 0.0638 | 0.0998 | 0.0047 |
| 30K | 0.0692 | 0.0231 | 0.5628 | 0.0283 | 0.4157 | 0.0258 | 0.0536 | 0.0026 |
| 35K | 0.0327 | 0.0136 | 0.7998 | 0.013 | 0.1919 | 0.0117 | 0.0361 | 0.0016 |
| 40K | 0.0121 | 0.0036 | 0.9358 | 0.003 | 0.0606 | 0.0028 | 0.0108 | 0.0005 |

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

### 4.4 Marathon Comparison

The corresponding tables for New York and Chicago showing MAE values and credible interval information are included in the appendix. Like Table 2, the appendix Tables A.1, and A.4 (for New York and Chicago, respectively) show that there is a significant percent improvement between the models and the baseline for the three marathons. We note similar behavior between the models: M2 and M3 having comparable performance, yet bettering the M1 percent improvement, especially at the latter stages. Tables A.2 and A.5 show the credible interval information, which generally appear to fit our expectations. Finally, the parameter estimates are shown in Tables A.3 and A.6.

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 is open for everyone to have a chance at running (only time qualifiers or charity runners can enter and run). This distribution difference affects the prediction accuracy of the disparate models. The credible interval sizes for Boston also appear to be smaller overall compared to New York and Chicago. Additionally, races have different qualities (such as the 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. Fig. 6 shows one of the outputs of the app: a plot displaying 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 in 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 from the app 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 while the race is taking place.

## 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 is an important consideration in addressing the problem, as the goal is to have a model that performs better than the baseline method but is still simple and interpretable. Adding too many features (such as using all prior splits, for example) makes the model harder to interpret and significantly increases the time it takes 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 do not have access to, but are strong indicators of marathon performance that would likely result in better

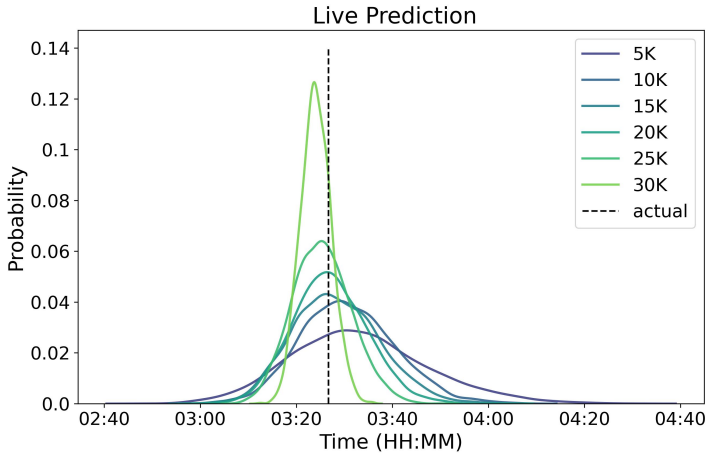


**Fig. 6:** Plot generated by the application after a runner inputs times. The probability distributions for finish times at each split of the race are shown.

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 in the app, 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. For example, although overall M1 outperforms M2, there are individual samples in our test set where M1 had a better prediction than M2. 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

As mentioned in Section 4.4, the results for the New York and Chicago Marathons largely mirror those of the Boston Marathon. Tabs. A.1 and A.4 compare the MAE values for New Yokr and Chicago, respectively. Similarly, Tabs. A.2 and A.5 show the credible interval tables for each marathon, while Tabs. A.3 and A.6 show the parameter estimates for the M2 model for each marathon.

| | MAE | | | | % Improve from BL | | |
|---|---|---|---|---|---|---|---|
| Distance | BL | M1 | M2 | M3 | M1 | M2 | M3 |
| 5K | 21.426 | 19.402 | 19.392 | 19.174 | 0.094 | 0.095 | 0.105 |
| 10K | 20.437 | 16.345 | 15.655 | 15.423 | 0.2 | 0.234 | 0.245 |
| 15K | 19.057 | 16.468 | 13.251 | 13.229 | 0.136 | 0.305 | 0.306 |
| 20K | 16.88 | 12.381 | 9.749 | 9.574 | 0.267 | 0.422 | 0.433 |
| 25K | 12.0 | 9.721 | 7.111 | 7.026 | 0.19 | 0.407 | 0.414 |
| 30K | 9.178 | 6.895 | 4.727 | 4.739 | 0.249 | 0.485 | 0.484 |
| 35K | 4.945 | 4.143 | 2.578 | 2.57 | 0.162 | 0.479 | 0.48 |
| 40K | 1.156 | 1.142 | 0.753 | 0.743 | 0.012 | 0.348 | 0.357 |
| Overall MAE | 13.149 | 10.828 | 9.166 | 9.073 | | | |
| Overall $R^2$ | 0.871 | 0.927 | 0.937 | 0.937 | | | |

**Tab. A.1:** 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 | 37.241 | 37.23 | 37.187 | 71.78 | 71.842 | 71.731 | 112.878 | 112.988 | 113.034 |
| 10K | 27.498 | 26.339 | 26.074 | 52.68 | 50.423 | 49.926 | 81.782 | 78.253 | 77.428 |
| 15K | 27.726 | 23.173 | 23.048 | 53.134 | 44.285 | 44.061 | 82.507 | 68.523 | 68.19 |
| 20K | 24.811 | 20.37 | 20.068 | 47.482 | 38.91 | 38.322 | 73.539 | 60.05 | 59.129 |
| 25K | 15.789 | 13.57 | 13.366 | 30.09 | 25.886 | 25.46 | 46.301 | 39.799 | 39.146 |
| 30K | 11.768 | 8.596 | 8.396 | 22.41 | 16.357 | 15.974 | 34.428 | 25.074 | 24.474 |
| 35K | 7.904 | 5.796 | 5.804 | 15.047 | 11.031 | 11.037 | 23.073 | 16.924 | 16.917 |
| 40K | 2.17 | 1.682 | 1.653 | 4.126 | 3.198 | 3.143 | 6.318 | 4.898 | 4.817 |
| Proportion of True Finish Times Within Interval | | | | | | | | | |
| 5K | 0.467 | 0.465 | 0.473 | 0.805 | 0.805 | 0.811 | 0.97 | 0.967 | 0.971 |
| 10K | 0.413 | 0.419 | 0.432 | 0.717 | 0.735 | 0.741 | 0.942 | 0.939 | 0.939 |
| 15K | 0.383 | 0.45 | 0.448 | 0.746 | 0.8 | 0.795 | 0.969 | 0.961 | 0.962 |
| 20K | 0.496 | 0.572 | 0.584 | 0.806 | 0.856 | 0.857 | 0.971 | 0.962 | 0.961 |
| 25K | 0.419 | 0.519 | 0.527 | 0.733 | 0.827 | 0.812 | 0.935 | 0.958 | 0.957 |
| 30K | 0.442 | 0.535 | 0.526 | 0.764 | 0.832 | 0.806 | 0.942 | 0.943 | 0.939 |
| 35K | 0.5 | 0.669 | 0.668 | 0.816 | 0.899 | 0.9 | 0.959 | 0.964 | 0.964 |
| 40K | 0.534 | 0.68 | 0.677 | 0.834 | 0.907 | 0.904 | 0.954 | 0.971 | 0.968 |

**Tab. A.2:** 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.

| | $\alpha$ - **Intercept** | | $\beta_1$ - **total_pace** | | $\beta_2$ - **curr_pace** | | $\sigma$ - **sigma** | |
|---|---|---|---|---|---|---|---|---|
| Distance | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 5K | -0.0217 | 0.0896 | 0.4364 | 0.7025 | 0.471 | 0.7023 | 0.2407 | 0.0113 |
| 10K | -0.0651 | 0.0609 | 0.085 | 0.1797 | 0.8398 | 0.1774 | 0.17 | 0.0077 |
| 15K | -0.0439 | 0.0516 | -0.1887 | 0.1287 | 1.1317 | 0.1242 | 0.1546 | 0.0069 |
| 20K | 0.1885 | 0.05 | 0.1343 | 0.0814 | 0.7537 | 0.0769 | 0.1365 | 0.006 |
| 25K | 0.0688 | 0.0315 | 0.5513 | 0.0492 | 0.4 | 0.0459 | 0.0917 | 0.0043 |
| 30K | 0.0661 | 0.0185 | 0.6076 | 0.0253 | 0.3485 | 0.0235 | 0.0581 | 0.0024 |
| 35K | 0.0403 | 0.0139 | 0.8 | 0.0135 | 0.1814 | 0.0129 | 0.0394 | 0.0018 |
| 40K | 0.0178 | 0.0036 | 0.9464 | 0.0036 | 0.0486 | 0.0036 | 0.0114 | 0.0005 |

**Tab. A.3:** New York Marathon posterior parameter value estimates for M2.

| | MAE | | | | % Improve from BL | | |
|---|---|---|---|---|---|---|---|
| Distance | BL | M1 | M2 | M3 | M1 | M2 | M3 |
| 5K | 20.679 | 16.469 | 16.466 | 16.369 | 0.204 | 0.204 | 0.208 |
| 10K | 18.328 | 14.178 | 13.631 | 13.477 | 0.226 | 0.256 | 0.265 |
| 15K | 16.809 | 13.078 | 12.387 | 12.131 | 0.222 | 0.263 | 0.278 |
| 20K | 15.742 | 10.998 | 9.634 | 9.326 | 0.301 | 0.388 | 0.408 |
| 25K | 13.35 | 9.293 | 7.511 | 7.408 | 0.304 | 0.437 | 0.445 |
| 30K | 9.231 | 6.64 | 4.889 | 4.854 | 0.281 | 0.47 | 0.474 |
| 35K | 5.451 | 3.876 | 2.86 | 2.818 | 0.289 | 0.475 | 0.483 |
| 40K | 1.149 | 1.107 | 0.861 | 0.838 | 0.037 | 0.25 | 0.27 |
| Overall MAE | 12.604 | 9.468 | 8.549 | 8.422 | | | |
| Overall $R^2$ | 0.888 | 0.936 | 0.941 | 0.941 | | | |

**Tab. A.4:** 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 | 29.476 | 29.46 | 29.299 | 56.644 | 56.645 | 56.275 | 88.459 | 88.483 | 87.999 |
| 10K | 26.08 | 25.48 | 25.646 | 50.013 | 48.84 | 49.171 | 77.787 | 75.962 | 76.461 |
| 15K | 22.779 | 22.512 | 22.525 | 43.562 | 43.036 | 43.074 | 67.471 | 66.685 | 66.677 |
| 20K | 19.209 | 15.831 | 15.55 | 36.681 | 30.2 | 29.665 | 56.61 | 46.501 | 45.664 |
| 25K | 19.273 | 15.488 | 15.39 | 36.802 | 29.566 | 29.339 | 56.79 | 45.51 | 45.146 |
| 30K | 12.542 | 9.309 | 9.141 | 23.896 | 17.729 | 17.399 | 36.744 | 27.209 | 26.67 |
| 35K | 6.916 | 6.055 | 5.923 | 13.173 | 11.529 | 11.276 | 20.205 | 17.686 | 17.281 |
| 40K | 1.843 | 1.425 | 1.412 | 3.508 | 2.71 | 2.685 | 5.372 | 4.153 | 4.113 |
| Proportion of True Finish Times Within Interval | | | | | | | | | |
| 5K | 0.541 | 0.541 | 0.544 | 0.849 | 0.85 | 0.843 | 0.957 | 0.957 | 0.956 |
| 10K | 0.547 | 0.561 | 0.565 | 0.86 | 0.862 | 0.87 | 0.961 | 0.96 | 0.964 |
| 15K | 0.482 | 0.503 | 0.519 | 0.827 | 0.851 | 0.857 | 0.954 | 0.959 | 0.961 |
| 20K | 0.497 | 0.503 | 0.513 | 0.826 | 0.817 | 0.814 | 0.941 | 0.924 | 0.923 |
| 25K | 0.566 | 0.61 | 0.619 | 0.894 | 0.889 | 0.889 | 0.96 | 0.959 | 0.96 |
| 30K | 0.543 | 0.617 | 0.61 | 0.855 | 0.87 | 0.867 | 0.954 | 0.944 | 0.946 |
| 35K | 0.521 | 0.612 | 0.614 | 0.834 | 0.888 | 0.879 | 0.939 | 0.958 | 0.958 |
| 40K | 0.53 | 0.587 | 0.596 | 0.807 | 0.83 | 0.833 | 0.928 | 0.936 | 0.933 |

**Tab. A.5:** 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.

| | $\alpha$ - **Intercept** | | $\beta_1$ - **total_pace** | | $\beta_2$ - **curr_pace** | | $\sigma$ - **sigma** | |
|---|---|---|---|---|---|---|---|---|
| Distance | Mean | Std | Mean | Std | Mean | Std | Mean | Std |
| 5K | -0.276 | 0.0716 | 0.4938 | 0.7076 | 0.5394 | 0.708 | 0.2155 | 0.0099 |
| 10K | -0.169 | 0.0639 | 0.3101 | 0.2052 | 0.6964 | 0.1988 | 0.1851 | 0.0093 |
| 15K | -0.2455 | 0.0473 | 0.5462 | 0.1497 | 0.4748 | 0.1481 | 0.1612 | 0.0071 |
| 20K | -0.0394 | 0.0355 | 0.2291 | 0.0724 | 0.7428 | 0.0692 | 0.1158 | 0.0053 |
| 25K | 0.066 | 0.0337 | 0.2201 | 0.0614 | 0.733 | 0.0575 | 0.1098 | 0.0048 |
| 30K | 0.0065 | 0.0191 | 0.6586 | 0.0232 | 0.3245 | 0.0213 | 0.0673 | 0.0031 |
| 35K | 0.0135 | 0.0133 | 0.8708 | 0.0147 | 0.114 | 0.0134 | 0.0441 | 0.002 |
| 40K | 0.0099 | 0.0033 | 0.9595 | 0.0033 | 0.0381 | 0.0031 | 0.0103 | 0.0005 |

**Tab. A.6:** 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] 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).

[19] *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).

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

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