Research article

Zhenyu Gao*, Yixing Li and Zhengxin Wang

Restoring the real world records in Men's swimming without high-tech swimsuits

https://doi.org/10.1515/jqas-2019-0087
Received August 21, 2019; accepted July 16, 2020; published online August 7, 2020

Abstract: The recently concluded 2019 World Swimming Championships was another major swimming competition that witnessed some great progresses achieved by human athletes in many events. However, some world records created 10 years ago back in the era of high-tech swimsuits remained untouched. With the advancements in technical skills and training methods in the past decade, the inability to break those world records is a strong indication that records with the swimsuit bonus cannot reflect the real progressions achieved by human athletes in history. Many swimming professionals and enthusiasts are eager to know a measure of the real world records had the high-tech swimsuits never been allowed. This paper attempts to restore the real world records in Men's swimming without high-tech swimsuits by integrating various advanced methods in probabilistic modeling and optimization. Through the modeling and separation of swimsuit bias, natural improvement, and athletes' intrinsic performance, the result of this paper provides the optimal estimates and the 95% confidence intervals for the real world records. The proposed methodology can also be applied to a variety of similar studies with multi-factor considerations.

Keywords: bias analysis; probability and statistics; swimming; time series; world records.

1 Introduction

Competitive swimming has a long history of over a hundred years and has become one of the most prominent events at the Summer Olympic Games. In the world's highest level competitive swimming events, swimmers aim to achieve a higher place on the podium and challenge the world records. The two world's best stages for competitive swimmers are the International Swimming Federation (FINA) World Championships and the swimming events at the Summer Olympics. The swimming World Championships was first held in 1973, and has now grown to include 42 events in the biennially meeting. Swimming at the Summer Olympics can date back to the inaugural 1896 Summer Olympic Games, and will have a total of 36 events in the incoming 2020 Tokyo Olympic Games (FINA 2020b).

Over the past 50 years, visible improvement in swimming time is observed in all competitive events due to the advancements in technical skills, training methods, equipment technologies, etc. The progression of swimming world records can best reflect the improvement made in this sport. For example, world record of the Men's 100 m freestyle has changed from 51.94 s in 1970 to the current world record of 46.91 s in 2009 (FINA 2020c). Since the year 2009, however, world records in many swimming events have remained unchanged. This abnormal phenomenon is attributed to a technology revolution a decade ago. From February 2008 to December 2009, with the use of the controversial high-tech swimsuits, the swimming community experienced an extraordinary improvement in performance in just two years (Brammer, Stager, and Tanner 2012). Compared to a traditional swimsuit, the swimsuits made of non-textile materials can further reduce drag against water and elevate the swimmer's body position through water. According to Tang (2008) and Craik (2011), the high-tech swimsuits are made of thin sheets of polyurethane or other non-textile materials in order to minimize drag, maximize support to the muscles, improve core stability, and increase buoyancy. Speedo's research showed that their high-tech design 'The LZR Racer' can reduce drag or water resistance by 38% compared to a

***Corresponding author: Zhenyu Gao**, Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA, e-mail: zhenyu.gao@gatech.edu
Yixing Li, Daniel Guggenheim School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA, e-mail: liyixingustc@gmail.com
Zhengxin Wang, H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA, e-mail: zhengxw@g.clemson.edu

traditional swimsuit made of Lycra, which translates into approximately a 4% increase in speed for swimmers (Tang 2008). With the introduction of the high-tech swimsuits, more than 130 world records were broken in 2008 and 2009. Seeing these significantly improved swimming times, some people argued that wearing the high-tech swimsuits is 'technology doping'. In view of the dispute, on January 1, 2010, FINA enforced new rules and banned the use of all swimsuits made of non-textile materials. Today, there are specific measures to regulate different characteristics of the swimsuit, such as the thickness, buoyancy, permeability, and body coverage (FINA 2017).

Although the high-tech swimsuit era ended at the beginning of 2010, world records created during that period were preserved. Apparently, assisted by the swimsuit bonus, world records set in the high-tech swimsuit era cannot mirror the real progressions made by human athletes over time. Consequently, 10 years on, some of them still remain untouched. Looking at those 'unreal' world records, swimming professionals and enthusiasts may come up with a question: what would the real world records be had the high-tech swimsuits never been allowed in history? This paper presents a methodology that rigorously quantifies the swimsuit bias in different events and restores the real world records without high-tech swimsuits. The scope of this work mainly focuses on medium-to-short distance events in Men's Swimming, yet the method can also be extended to other events and even other sports. The remainder of this paper is organized as follows. Section 2 contains a literature survey of past studies on similar topics. Section 3 provides a review of the current world records. Section 4 describes the data used to conduct this study and conducts an initial gap analysis. Section 5 applies various statistical and optimization methods on the collected data to answer the main question. Finally, Section 6 provides a summary and concludes the paper.

2 Literature review

Several previous studies in the literature identified the swimsuit bias and conducted preliminary analyses regarding its impact. Dyer (2015) conducted a systematic review of the controversial sports technologies, and contained details of the debate on the use of full body swimsuits and the 'fastskin' swimsuits. Stager, Brammer, and Tanner (2012) used the longitudinal progression of athletic records to calculate the predictions of 2008 Olympic Games swimming events. Through comparisons, they confirmed the existence of swimsuit bias because 17 out of 26 events were significantly faster than predicted. In a similar paper, Brammer et al. (2012) fitted improvement curve on historical Olympic performances to predict the mean swim time of the top eight swimmers in swimming events at the 2012 Olympic Games. They concluded that the 2008 Olympic Games results were biased by the banned tech suits and the predicted 2012 results will realign with the prediction curves. Foster, James, and Haake (2012) described a most relevant quantitative study by far. They built a five-parameter regression model on the top 25 times in each year since 1948 to assess the effects of swimsuit technology. They found out that the introduction of polyurethane panels in 2008 increased performance by 1.5–3.5%; the use of full body polyurethane suits in 2009 increased this performance further and by up to 5.5%.

Some other quantitative studies conducted on swimming but on different concerns showed the trend of using statistical methods to analyze swimming training and competitions. Costa et al. (2010) tracked world-ranked male swimmers performance during five consecutive seasons from 2003 to 2008 in Olympic freestyle events. They used descriptive statistics, analysis of variance (ANOVA), and correlation analysis to find out that the performance enhancement was approximately 3–4% for the time-frame. Cornett, Brammer, and Stager (2015) investigated a controversy of the 2013 FINA World Swimming Championships. Through statistical analysis, they identified that the swimming performances were biased depending on the direction and lane in which the swimmers swam.

By far, nevertheless, none of the previous works attempted to quantify the high-tech swimsuits' influences on swimming world records. The most relevant previous studies only provide very rough ranges on how much athletes' performance can be improved by wearing high-tech swimsuits, which are not adequate to 'correct' the current world records. In addition, most of the quantitative studies in the literature used descriptive statistics and a single parametric regression to analyze swimming data, which include too many assumptions to simplify the problem and underestimate the actual complexity in this type of problems. Conversely, restoring the world records is a challenging task that must be done by using more advanced statistical methods on a more customized dataset, which is precisely the overarching objective of this paper.

3 Current world records

Under the proposed methodology and procedure, this paper attempts to restore the world records of four events in Men's Swimming: 50 m freestyle, 100 m freestyle, 200 m

Table 1: Current world records of the four selected events.

Event	Time (seconds)	Holder	Year
50 m freestyle	20.91	Cesar Cielo	2009
100 m freestyle	46.91	Cesar Cielo	2009
200 m freestyle	1:42.00	Paul Biedermann	2009
4 × 100 m freestyle relay	3:08.24	United States	2008

Table 2: Best records of the four selected events since 2010.

Event	Time (seconds)	Holder	Year
50 m freestyle	21.04	Caeleb Dressel	2019
100 m freestyle	46.96	Caeleb Dressel	2019
200 m freestyle	1:43.14	Yannick Agnel	2012
4 × 100 m freestyle relay	3:09.06	United States	2019

freestyle, and 4 × 100 m freestyle relay. They are selected because their current world records are the most representative ones that were influenced by the "swimsuit bonus" (FINA 2020c). Meanwhile, these four events are always in the spotlight of Men's Swimming. Details of the four current world records are shown in Table 1.

All four world records in Table 1 were set in either 2008 or 2009 when the high-tech swimsuits were allowed to use in competitions. World records of both 50 and 100 m freestyle were set by Brazilian swimmer Cesar Cielo in 2009. World record of 200 m freestyle was set by German swimmer Paul Biedermann during the 2009 World Championships held in Rome, Italy. The United States relay team set the 4 × 100 m freestyle relay world record during the 2008 Olympic Games held in Beijing, China. Since the high-tech swimsuits were prohibited, these world records have remained unchanged for over a decade. However, owing to the advancements in other aspects of the sport in the post high-tech swimsuit era, many athletes have achieved fast times that are not too far away from these world records. Table 2 shows the best records of the four events since 2010 (FINA 2020c). Without using the high-tech swimsuits, these best records are good indicators of the real progression of human athletes in swimming. Some records in Table 2 are very close to their corresponding world records in Table 1. For example, the best record for 100 m freestyle since 2010, set by American Caeleb Dressel in the 2019 World Championships, is only 0.05 s slower than the world record. If the high-tech swimsuits were never used in history, it is possible that the times in Table 2 would become the current world records.

4 Dataset and initial analysis

Given that the scope of this study is on world records, we only use data that were produced by the world's best swimmers in the best competitions. Under this standard, we choose data from the finals and semifinals in Olympic Games (held once every four years) and World Championships (held every two years in odd-numbered years). Since the high-tech swimsuits were only used in 2008 and 2009, the final and semifinal results from the 2008 Olympic Games and the 2009 World Championships can best represent the top swimming times with high-tech

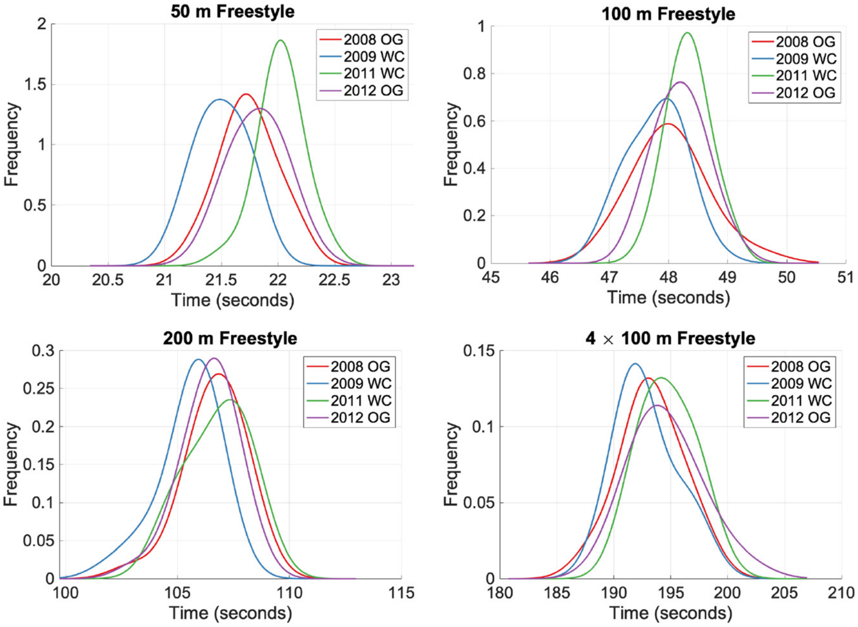

Figure 1: Empirical distributions of the raw data.

swimsuits. Final and semifinal results from the two best competitions right after the high-tech swimsuits era – the 2011 World Championships and the 2012 Olympic Games are used to represent the top swimming times without high-tech swimsuits. The dataset is not large, but it cannot be expanded due to the restricted scope of the study. Additional data from the National Championships of some countries from 2008 to 2012 may be carefully selected to increase the dataset size, yet more complex judgments have to be used in this process. In view of this, the proposed methodology also has uncertainty consideration to address the small data challenge.

Figure 1 shows the empirical distributions of the dataset used in this study. Each plot in Figure 1 displays how data from the four major competitions are distributed for a specific event, where Kernel density estimation (KDE) with normal kernel function is applied on the data. A few observations can be made by looking at these distributions. First, top times from the high-tech swimsuits era (red and blue curves) are in general faster than top times from the post high-tech swimsuits era (green and purple curves). Second, times improved noticeably from one year to the next within each of the two eras - top times from 2012 to 2009 are faster than top times from 2011 to 2008 respectively. This observation highlights the need to include the natural improvement achieved by top athletes year after year into the model. The only exception to the two observations above is the 200 m freestyle event, in which the top times from the 2012 Olympic Games (post high-tech [HT] swimsuit era) is slightly faster than top times from the 2008 Olympic Games (HT swimsuit era). This might indicate that the top athletes' natural improvement over those four years is at least comparable to the benefits of high-tech swimsuits in this event.

In an initial gap analysis, we try to quantify the enhancive chance of reaching to the "World's top time" of each event by wearing high-tech swimsuits. In the swimming community, there is a commonly acknowledged criterion for the top time of each event. By swimming faster than the top times, swimmers are considered world class and have decent chances of reaching to the finals of World's major competitions. The definitions of top time for the four selected events are given in the second row of Table 3. The rest of Table 3 compares the probabilities of obtaining top times between 2008 and 2012. It can be observed that by wearing high-tech swimsuits, swimmers had significantly higher chances of achieving the top times. In the meantime, the effect of high-tech swimsuits is more influential in short distance events. In 50 and 100 m freestyle, top swimmers had around 30% higher chances of achieving top times when wearing the high-tech swimsuits. Yet in 200 m freestyle, this difference decreases to around 12%. Please note that although the 4 × 100 m freestyle relay has the longest total distance, it is still considered as a short distance event because each swimmer only swims 100 m. This initial comparison result further confirms the swimsuit bias, and shows that it has different impacts in short and medium distance events.

5 Methodology

To achieve the overarching objective and provide valuable reference to 'correct' the current world records, many factors must be taken into account in the modeling process. In this work, the following three aspects are included when analyzing the dataset:

(1) *Swimsuit Bias*: among the four major competitions, results from the 2008 Olympic Games and the 2009 World Championships include the "swimsuit bonus", while results from the 2011 World Championships and the 2012 Olympic Games do not. Quantifying the swimsuit bias is the key to restoring the current world record for each event.

(2) *Natural Improvement*: even though the dataset only spans across four years, which is a relatively short time period, it is still imperative to consider the natural improvement in swim performance over four years. If this time dependent performance factor is not included, results from four different years could not be compared on the same ground level that best reflects the effects brought by the swimsuits.

(3) *Swimmers' True (Intrinsic) Performance in 2008*: after the swimsuit bias is removed and the natural

Table 3: Initial gap analysis.

		50 m freestyle	100 m freestyle	200 m freestyle	4 × 100 m FR
Top time (seconds)		<22	<48	<106	<193
Top time chance (%)	With HT swimsuit	91.24	54.12	41.75	62.37
	Without HT swimsuit	60.31	23.19	29.38	27.31
Difference (%)		30.93	30.93	12.37	35.06

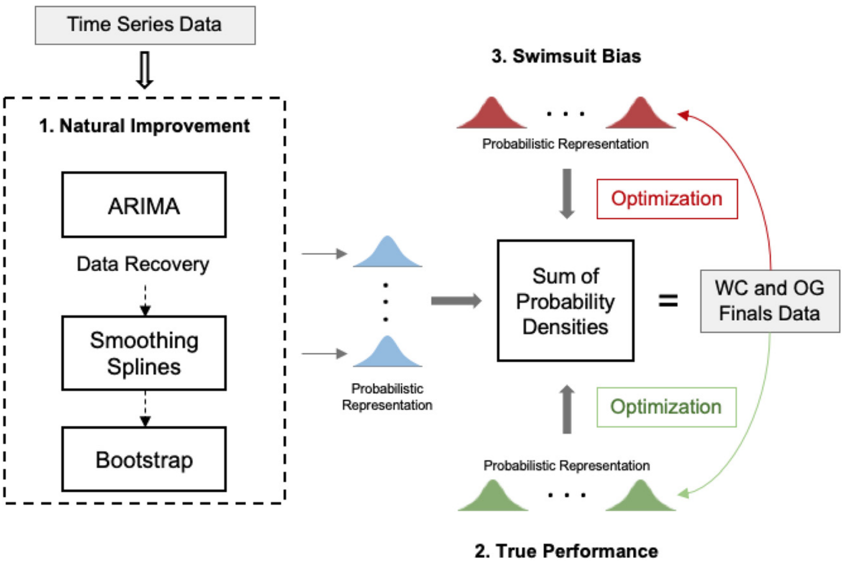

Figure 2: Flowchart of the proposed methodology.

improvement is deducted, the whole dataset should be left with the swimmers' intrinsic performance in 2008, the earliest year in the dataset. This last aspect serves as the reference basis of the overall analysis.

With the above three factors in mind, this study utilizes various probabilistic and statistical methods to model them. A flowchart of the proposed methodology is displayed in Figure 2. A critical point in this study is that we treat each of three factors using non-deterministic representations. We use probability distributions to represent the swimsuit bias, natural improvement, and swimmers' true performance in each event to acknowledge the variations among athletes and swimsuits. In the Sections 5.1 and 5.2, for each event we first use time series techniques (Autoregressive Integrated Moving Average [ARIMA] models), smoothing splines, and bootstrap to obtain the probabilistic representation of natural improvement. Subsequently, we treat the distributions of the raw dataset as the sums of probability densities from all of natural improvement, swimsuit bias, and true performance, and use an optimization setup to find optimal parameters for swimsuit bias and true performance. Overall, this is an approach that probabilistically separates the three factors from the collected dataset. In the end, we provide optimal estimates for the restored read world records as well as their 95% confidence intervals.

5.1 ARIMA

The modeling of natural improvement achieved by human swimmers from 2008 to 2012 starts with the Autoregressive Integrated Moving Average (ARIMA) model. This part of the analysis is conducted with two objectives in mind: (1) further demonstrating the deviation from natural improvement during the high-tech swimsuit years, and (2) estimating and recovering the true natural improvement data to support next part of the analysis. To study this time-dependent pattern, another dataset that contains the best time records in the finals of the Olympic Games and World Championships in the past 40 years is collected (FINA 2020c).

The ARIMA model is a forecasting technique that is able to predict to future value of a series based entirely on the previous values of the series (Box et al. 2015). In a nonseasonal ARIMA (p, d, q) model, p is the number of autoregressive terms, d is the number of nonseasonal differences needed for stationarity, and q is the number of lagged forecast errors in the prediction equation. Because of limited space, a detailed mathematical introduction of ARIMA model is not included in this article. In this study, we use ARIMA to restore the real time-dependent performance progression of each event without high-tech swimsuits. For each event, we determine an ARIMA (p, d, q) model and use Kalman filter to handle missing values (here, records in years 2008 and 2009 are treated as missing data). More specifically, we take state space form of ARIMA model from the output returned by ARIMA and pass it to Kalman filter. Based on results before 2008 and after 2009, we are able to obtain the estimated best performances in these two years without the influence of high-tech swimsuits.

The visual ARIMA results are shown in Figure 3. Each plot in Figure 3 has three different elements. The black dots are the best records made without the use of high-tech

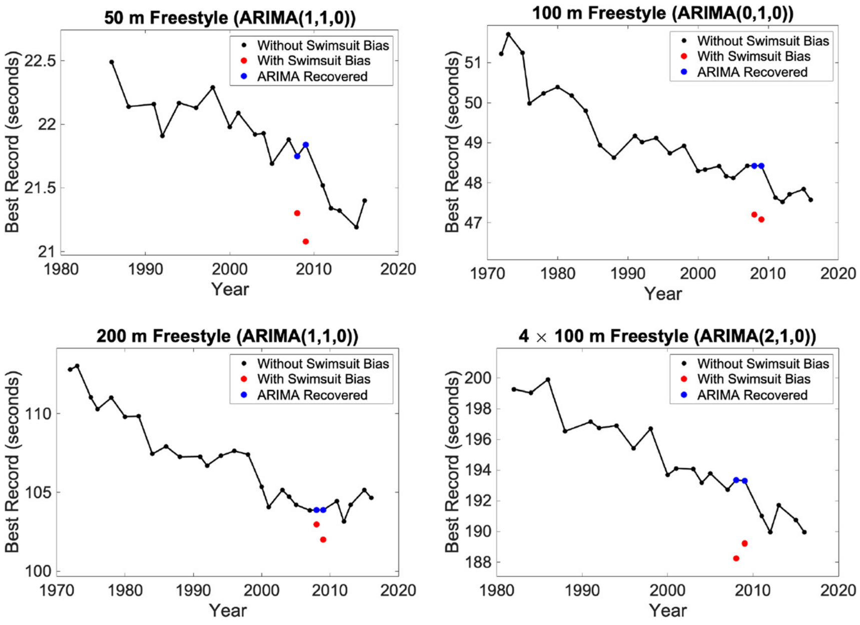

Figure 3: ARIMA results for the performance progress.

swimsuits; the red dots are the records in 2008 and 2009 when the high-tech swimsuits were allowed; the blue dots are the estimated results given by ARIMA and Kalman filter. It is seen that in each event, the two red dots are well below the black line which connects the black and blue dots, indicating the huge impact of the swimsuit bias on the records in 2008 and 2009. In the following subsection, the restored best records (black plus blue dots) are used to model the improved performance over the years.

5.2 Splines

At the beginning of this section, we mentioned that natural improvement must be considered when comparing swim times from different years. In this work, we use the cubic smoothing splines on the restored best records data to model the natural yearly improvements. The spline method is selected because of its proper flexibility and better stability at the boundaries. For each event, the objective here is to find a function $g(x)$ that minimizes

$$\sum_{i=1}^{N}(y_i - g(x_i))^2 + \lambda\int g''(t)^2dt \qquad (1)$$

where N is the dataset size and λ is a nonnegative tunning parameter that controls the bias-variance trade-off of the smoothing spline (James et al. 2015). To identify an optimal λ value, it is recommended to use the leave-one-out cross-validation (LOOCV) error, given by

$$\mathrm{RSS}_{cv}(\lambda) = \sum_{i=1}^{N}\left(y_i - \hat{g}_\lambda^{(-1)}(x_i)\right)^2 \qquad (2)$$

where $\hat{g}_\lambda^{(-1)}(x_i)$ indicates the fitted function at x_i, using all but the ith training observation. In this study, we aim to obtain very smooth splines to model the improved performance. Therefore, a very large λ is used to avoid rough interpolation results.

Using the 2008 results as the base, for each event we need to model three different improvement progresses: δ_1 (from 2008 to 2009), δ_2 (from 2008 to 2011), and δ_3 (from 2008 to 2012). Also, this modeling process must have uncertainty consideration due to two facts: (1) the uncertainty of the model itself, and (2) different swimmers have different progresses in a given period. The resampling method bootstrap is employed to model the nondeterministic δ_1, δ_2, and δ_3. The procedure is as follows:

(1) From the original sample X of sample size N, we randomly draw N observations with replacement to produce a new bootstrap sample $X_{b,1}$.
(2) Use $X_{b,1}$ to fit the cubic smoothing spline and compute the three statistics $\hat{\delta}_{1,1}$, $\hat{\delta}_{2,1}$, and $\hat{\delta}_{3,1}$.
(3) Repeat 1 and 2 for 5000 times, and obtain $\mho : \{\hat{\delta}_{1,i}, \hat{\delta}_{2,i}, \hat{\delta}_{3,i}, 1 \le i \le 5000\}$.
(4) Fit Normal distributions on the estimates in $\mho$ to obtain the nondeterministic δ_1, δ_2, and δ_3.

Visual result of the described 'bootstrap + cubic smoothing splines' process is shown in Figure 4. In each plot, the black solid line is the fit to the restored best records data (black dots). The gray lines represent the family of 5000 bootstrap results, which are used to account for uncertainty. After fitted with Normal distributions, the mean and standard deviation values for all the nondeterministic δ_1's, δ_2's, and

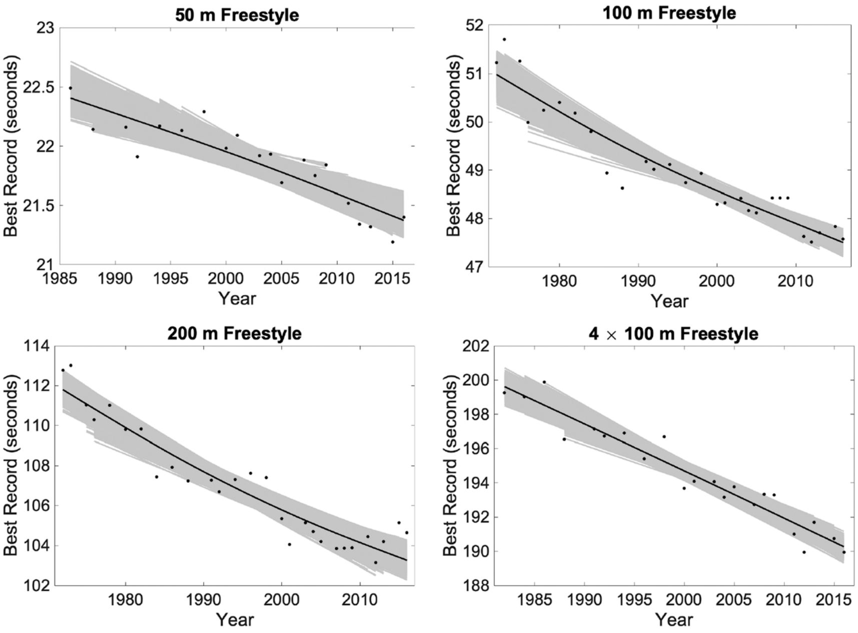

Figure 4: Bootstrap and spline results for natural improvement.

δ_3's are given in Table 4. It can be observed that for each event, we have $\mu_{\delta_3} < \mu_{\delta_2} < \mu_{\delta_1}$, which corresponds to faster swim times as each year goes by. In the meantime, $\sigma_{\delta_3} > \sigma_{\delta_2} > \sigma_{\delta_1}$ indicates a larger extent of uncertainty when estimating the progression over a longer time span. The distributions in Table 4 are used by the next subsection, in which the swimsuit bias is quantified through an optimization setup.

5.3 The final modeling and optimization scheme

Three different factors jointly influenced Men's Swimming dataset from 2008 to 2012. The first factor is the true swimming times of elite swimmers in 2008, denoted by β_{2008}. We can treat β_{2008} as the "intrinsic" true performance of elite swimmers at the beginning of the time period. The second factor to include is the performance improvement caused by better techniques, training methods, etc., over the four years. This natural improvement in performance had already been quantified in the last subsection, and here we further denote δ_1 as $\delta_{2008\to2009}$, δ_2 as $\delta_{2008\to2011}$, and δ_3 as $\delta_{2008\to2012}$. Finally, and most important, we must include the swimsuit bias for data in 2008 and 2009. We use τ_s to denote the swimsuit bias.

In the following analysis, we treat all three factors as nondeterministic quantities. For example, the swimmers' base performance in each event β_{2008} should be a distribution instead of a deterministic value, because even among elite swimmers, difference in their real swim time exists. In addition, different swimmers achieve different improvements each year, so the time factors δ's are also probability distributions. Lastly, the swimsuit bias τ_s is not uniform to all swimmers as its magnitude could be influenced by material, manufacturer, swimmers' body length, swimmers' technique, etc. Under the nondeterministic setting, Table 5 shows the assignment of factors in each year's result.

Table 4: Descriptive statistics of δ_1, δ_2, and δ_3 for all events.

Event	δ_1 (2008 → 2009)		δ_2 (2008 → 2011)		δ_3 (2008 → 2012)	
	μ_{δ_1}	σ_{δ_1}	μ_{δ_2}	σ_{δ_2}	μ_{δ_3}	σ_{δ_3}
50 m freestyle	0.037	0.005	−0.110	0.014	−0.147	0.018
100 m freestyle	−0.066	0.005	−0.199	0.016	−0.265	0.021
200 m freestyle	−0.158	0.021	−0.467	0.064	−0.620	0.087
4 × 100 m FR	−0.278	0.018	−0.835	0.053	−1.114	0.070

Table 5: Factor assignment from 2008 through 2012.

Year	Factors
2008	$\beta_{2008} + \tau_s$
2009	$\beta_{2008} + \delta_{2008\to2009} + \tau_s$
2011	$\beta_{2008} + \delta_{2008\to2011}$
2012	$\beta_{2008} + \delta_{2008\to2012}$

In Table 5, results from all four years include the factor β_{2008}, which is the intrinsic performance of swimmers at the beginning of the time period; results from 2008 to 2009 include the swimsuit bias factor τ_s; results from 2009, 2011, and 2012 include the time-dependent natural improvement. In this work we use Normal distribution to model all the listed factors. The Normal distribution is selected for two reasons: (1) it is the most commonly used probability distribution to reflect the natural variability among a group, and (2) it is the maximum entropy distribution when knowing the mean and the standard deviation of data. Under this assumption, all the factors are represented as follows:

$$\begin{aligned}\beta_{2008} &\sim \mathcal{N}(\mu_1, \sigma_1^2)\\ \tau_s &\sim \mathcal{N}(\mu_2, \sigma_2^2)\\ \delta_{2008\to2009} &\sim \mathcal{N}(\mu_3, \sigma_3^2)\\ \delta_{2008\to2011} &\sim \mathcal{N}(\mu_4, \sigma_4^2)\\ \delta_{2008\to2012} &\sim \mathcal{N}(\mu_5, \sigma_5^2)\end{aligned} \tag{3}$$

The factors β, τ, and δ are independent random variables. In the following procedure, we treat each distribution in the raw data as a sum of independent random variables according to Table 5. In probability theory, when X and Y are two independent and continuous random variables with density functions $f_X(x)$ and $f_Y(y)$, then the density function $f_Z(x)$ for $Z = X + Y$ is the convolution of f_X and f_Y. For Normal density, the convolution of two Normal densities with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is again a Normal density with mean $\mu_1 + \mu_2$ and variance $\sigma_1^2 + \sigma_2^2$. Hence, for each event, the distributions of the raw data can be expressed as follows:

$$\begin{aligned}&y_{2008}|\beta_{2008}, \tau_s \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\\ &y_{2009}|\beta_{2008}, \tau_s, \delta_{2008\to2009} \sim \mathcal{N}(\mu_1 + \mu_2 + \mu_3, \sigma_1^2 + \sigma_2^2 + \sigma_3^2)\\ &y_{2011}|\beta_{2008}, \delta_{2008\to2011} \sim \mathcal{N}(\mu_1 + \mu_4, \sigma_1^2 + \sigma_4^2)\\ &y_{2012}|\beta_{2008}, \delta_{2008\to2012} \sim \mathcal{N}(\mu_1 + \mu_5, \sigma_1^2 + \sigma_5^2)\end{aligned} \tag{4}$$

The four distributions in Equation (4) involve a total of 10 parameters, among which the six time dependent related ones on the δ's had already been calculated by using ARIMA, smoothing splines, bootstrap, and are summarized in Table 4. Now for each event, the target of this final step is to estimate the following four parameters for the true performance in 2008 and swimsuit bias: μ_1, μ_2, σ_1, and σ_2.

In the subsequent process, for each event we find the optimal values of μ_1, μ_2, σ_1, σ_2 through maximizing the likelihood of all four years' data of the event, which can be written as:

$$\mathcal{L}(\mu_1, \mu_2, \sigma_1, \sigma_2) = \prod_{t=2008}^{2012} \prod_{i=1}^{N} y_t(x_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \tag{5}$$

When maximizing the log-likelihood of Equation (5), the two target distributions $\beta_{2008} \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $\tau_s \sim \mathcal{N}(\mu_2, \sigma_2^2)$ are independent. Therefore, the parameters (μ_1, σ_1) and (μ_2, σ_2) can be optimized sequentially. The optimization process is conducted as follows:

(1) *Initialization:* From data, provide the ranges (upper and lower bounds) and initial guesses for $\mu_1, \mu_2, \sigma_1, \sigma_2$:

$$\mathcal{U}_1, \mathcal{U}_2, \mathcal{S}_1, \mathcal{S}_2, \mu_{1,i}, \mu_{2,i}, \sigma_{1,i}, \sigma_{2,i}$$

(2) Since the two parameters in (μ_1, σ_1) are expected to be orders of magnitude larger than their counterparts in (μ_2, σ_2), first optimize (μ_1, σ_1) using the current values of (μ_2, σ_2):

$$(\mu_1, \sigma_1) \leftarrow \underset{\mu_1 \in \mathcal{U}_1, \sigma_1 \in \mathcal{S}_1}{\operatorname{argmin}} \mathcal{L}(\mu_1, \mu_2, \sigma_1, \sigma_2)$$

(3) Optimize (μ_2, σ_2) using the updated (μ_1, σ_1):

$$(\mu_2, \sigma_2) \leftarrow \underset{\mu_2 \in \mathcal{U}_2, \sigma_2 \in \mathcal{S}_2}{\operatorname{argmin}} \mathcal{L}(\mu_1, \mu_2, \sigma_1, \sigma_2)$$

(4) *Repeat* steps 2 and 3 *until* a convergence criterion is met.

In this subject matter, we conduct the optimization process under one constraint on σ_2. Apart from the common acknowledgment that the high-tech swimsuit hardly undermines swimming performance, the addition of this constraint has another consideration. During the optimization process, it is found out that among the four parameters to optimize in each event: $\mu_1, \mu_2, \sigma_1, \sigma_2$, the objective function is least sensitive to the value of σ_2 because of its smallest order of magnitude and relatively 'weak role'. In this case the value of σ_2, standard deviation for swimsuit bias, is likely to be overestimated because the estimation process of σ_2 from small dataset is more vulnerable to the role of chance. Therefore, the constraint $\mu_2 + 6 \cdot \sigma_2 \leq 0$ is applied to make more robust estimation on the dispersion of swimsuit bias. A summary of the final

Table 6: Optimized parameters for the two unknown distributions (seconds).

Event	Optimal μ_1	Optimal σ_1	Optimal μ_2	Optimal σ_2
50 m freestyle	22.04	0.23	−0.40	0.06
100 m freestyle	48.50	0.37	−0.56	0.09
200 m freestyle	1:47.14	1.22	−0.95	0.16
4 × 100 m FR	3:15.63	2.47	−2.41	0.40

Table 7: Restored world records of the four selected events.

Event	Current WR	Swimsuit Bias	Real WR (Optimal)	Real WR (95% CI)
50 m freestyle	20.91	$\mathcal{N}(-0.4, 0.06^2)$	21.31	[21.19, 21.43]
100 m freestyle	46.91	$\mathcal{N}(-0.56, 0.09^2)$	47.47	[47.29, 47.65]
200 m freestyle	1:42.00	$\mathcal{N}(-0.95, 0.16^2)$	1:42.95	[1:42.63, 1:43.27]
4 × 100 m FR	3:08.24	$\mathcal{N}(-2.41, 0.40^2)$	3:10.65	[3:09.85, 3:11.45]

optimization results for the two unknown distributions are given in Table 6.

With the complete optimized results for μ_1, μ_2, σ_1, σ_2, now we can try to restore the real world record for each event. Among the current world records displayed in Table 1, three were created in 2009 and one was created in 2008. We now subtract each swimsuit bias from its respective current world record and arrive at the optimal estimate and a 95% confidence interval for the real world record of that event. The restored world records for the four events are shown below in Table 7.

Using the restored world records, now we are able to compare them against the post high-tech swimsuits era best records shown in Table 2. It can be seen that the best records since 2010 for 50 m freestyle (21.04), 100 m freestyle (46.96), and 4 × 100 m freestyle relay (3:09.06) are well below the lower bounds of the 95% confidence interval of their respective restored world records (21.19, 47.29, and 3:09.85, respectively). So it is relatively safe to say that Caeleb Dressel and the United States relay team would own the world records of these three events had the high-tech swimsuits never been allowed in history. On the other hand, the best record for 200 m freestyle (1:43.14), although is not below the lower bound of the restored 95% confidence interval, falls into the range of the 95% confidence interval. This indicates that it is a very competitive record and is comparable to the status of a world record.

6 Conclusions

In this paper, we present a methodology that can be used to restore the real world records in Men's Swimming had the high-tech swimsuits never been used in history. Compared to the other existing methods in swimming analytics literature, this work considers and includes influential factors from various dimensions, and uses nondeterministic treatments in the modeling process. In this project we first select an appropriate range for data collection, conduct preliminary analysis on the collected dataset, and confirm the existence of swimsuit bias. The quantitative analysis then starts with applying ARIMA time series model, cubic smoothing splines, and resampling method to model the natural time dependent improvement achieved in the sport. Then, probabilistic modeling and optimization are utilized to learn the parameters of the nondeterministic distributions of swimsuit bias and athletes' true performance for different events. The final result includes optimal estimates and their 95% confidence intervals for the restored real world records. Topic-wise, we believe that the result of this work can serve as a reference answer to one of the most challenging questions in the swimming community. Method-wise, the proposed methodology covers typical elements that are very common in other sports, such as the time dependent progress, effect of advanced technologies, variability among athletes and equipments, etc. Therefore, this methodology can also provide reference for other research activities in and beyond sports.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.
Research funding: None declared.
Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

Box, G. E., G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. 2015. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, New Jersey.

Brammer, C. L., J. M. Stager, and D. A. Tanner. 2012. "Beyond the "High-tech" Suits: Predicting 2012 Olympic Swim Performances." *Measurement in Physical Education and Exercise Science* 16: 183–93.

Cornett, A., C. Brammer, and J. Stager. 2015. "Current Controversy: Analysis of the 2013 Fina World Swimming Championships." *Medicine & Science in Sports & Exercise* 47: 649–54.

Costa, M. J., D. A. Marinho, V. M. Reis, A. J. Silva, M. C. Marques, J. A. Bragada, T. M. Barbosa. 2010. "Tracking the Performance of World-Ranked Swimmers." *Journal of Sports Science and Medicine* 9: 411. 24149635.

Craik, J. 2011. "The Fastskin Revolution: From Human Fish to Swimming Androids." *Culture Unbound: Journal of Current Cultural Research* 3: 71–82.

Dyer, B. 2015. "The Controversy of Sports Technology: A Systematic Review." *SpringerPlus* 4: 524.

FINA. 2017. "Fina Requirements for Swimwear Approval (FRSA)." Regulations Valid for Swimwear to be Approved with Effect from January 1, 2017.

FINA. 2020b. *Swimming Events*. Also available at http://www.fina.org/calendar.

FINA. 2020c. *Swimming Records*. Also available at http://www.fina.org/fina-rankings/filter/records.

Foster, L., D. James, and S. Haake. 2012. "Influence of Full Body Swimsuits on Competitive Performance." *Procedia Engineering* 34: 712–7.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2015. "An Introduction to Statistical Learning." In *Springer Texts in Statistics Book Series*.

Stager, J. M., C. L. Brammer, and D. A. Tanner. 2012. "Identification of a Bias in the Natural Progression of Swim." In *Biomechanics and Medicine in Swimming XI*.

Tang, S. K. Y. 2008. *The Rocket Swimsuit: Speedo's Lzr Racer*. Also available at http://sitn.hms.harvard.edu/flash/2008/issue47-2/.