



This is a peer-reviewed, final published version of the following in press document, Attribution-NonCommercial 4.0 International CC BY-NC 4.0 <https://creativecommons.org/licenses/by-nc/4.0/> and is licensed under Creative Commons: Attribution-Noncommercial 4.0 license:

**Jaén-Carrillo, Diego ORCID logoORCID: <https://orcid.org/0000-0003-0588-0871>, Margarit-Boscà, Arcadi ORCID logoORCID: <https://orcid.org/0009-0006-0035-1427> and de Waal, Simon J ORCID logoORCID: <https://orcid.org/0000-0002-3841-3471> (2026) A laboratory-free digital twin for trail running: Population-level validation of GPS-calibrated performance prediction across athletes and race formats. International Journal of Sports Science & Coaching. doi:10.1177/17479541261492262 (In Press)**

Official URL: <https://doi.org/10.1177/17479541261492262>

DOI: <http://dx.doi.org/10.1177/17479541261492262>

EPrint URI: <https://eprints.glos.ac.uk/id/eprint/16588>

### **Disclaimer**

The University of Gloucestershire has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

The University of Gloucestershire makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

The University of Gloucestershire makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

The University of Gloucestershire accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.



# A laboratory-free digital twin for trail running: Population-level validation of GPS-calibrated performance prediction across athletes and race formats

Diego Jaén-Carrillo<sup>1</sup> , Arcadi Margarit-Boscà<sup>2</sup> , and Simon Jake de Waal<sup>3</sup>

## Abstract

The wide range of physiological demands in trail running makes performance modelling challenging, and existing models require resource intensive laboratory testing and lack individual predictive resolution. This study aimed to develop a two-layer physics-informed machine learning (ML) pipeline that predicts pre-race finish times without laboratory assessment – a physics-based simulation layer, whose parameters were calibrated from athlete's historical Global Positioning System (GPS) race records, combined with an XGBoost stacking layer that learned systematic residuals of the physics model. The pipeline was evaluated under leave-one-athlete-out (LOAO) cross-validation across 579 trail running races from 62 competitive athletes (44 males, 18 females; age:  $32.3 \pm 7.1$  years; ITRA Performance Index:  $739 \pm 109$ ) spanning five different race formats. The XGBoost stacking layer achieved a mean absolute percentage error (MAPE) of 11.2% (MAE = 22.0 min, RMSE = 43.9 min,  $R^2 = 0.938$ ), significantly outperforming the physics layer alone (MAPE = 11.7%;  $W = 73.5$ ,  $p = 0.010$ ). Feature importance analysis revealed the physics-layer prediction (37.6%) and log-transformed race distance (48.3%) together accounted for 85.9% of the XGBoost model's explanatory power, confirming the physics model as a structured informative prior. Accuracy varied by format (MAPE: 9.9–21.9%) and the stacking layer reduced the physics model's proportional bias from  $r = -0.564$  to  $r = -0.206$  and corrected elevation-dependent underestimation by up to 78.6% in races >3000 m elevation gain. This framework is the first population-validated, laboratory-free digital twin for trail running performance prediction, providing both athlete-specific accuracy and mechanistic interpretability from GPS data alone.

## Keywords

Aerobic efficiency, altitude, energy cost, global positioning system, pacing, physiological demands, wet bulb globe temperature

Reviewers: Felipe Garcia-Pinillos (University of Granada, Spain)

Jeff Rothschild (Auckland University of Technology, New Zealand)

## Introduction

Trail running, competitive foot races across natural terrain with substantial changes in gradient, surface, and altitude, has emerged as one of the fastest-growing endurance disciplines, with organised events now spanning formats from sub-10 km uphill races and VK (+1000 m elevation gain) events to ultra-distance mountain courses exceeding 100 km.<sup>1,2</sup> This competition diversity imposes a correspondingly wide range of physiological demands. The metabolic cost of locomotion varies non-linearly with slope – increasing as a function of incline up to +15% and directly proportional to slope thereafter, while decreasing until –15% and then rising linearly at steeper downhill slopes – due to the absence of the bouncing-ball mechanism and the resulting positive and negative only work in gradients exceeding  $\pm 15\%$ .<sup>3</sup> This unique characteristic of trail running

results in cumulative elevation gain and loss, rather than distance alone, as primary determinants of finish time. Additional modulators include altitude-related reductions in maximal aerobic power,<sup>4,5</sup> surface conditions,<sup>6</sup> and thermal environment.<sup>7</sup> Therefore, prediction models developed for

<sup>1</sup>Department of Sport Science, University of Innsbruck, Innsbruck, Austria

<sup>2</sup>Department of Sport Science, Universidad de Murcia, Murcia, Spain

<sup>3</sup>Community of Sport and Exercise, School of Education, Health and Science, University of Gloucestershire, Gloucestershire, UK

## Corresponding author:

Simon Jake de Waal, Community of Sport and Exercise, School of Education, Health and Science, University of Gloucestershire, Gloucestershire, UK.

Email: sdewaal@glos.ac.uk

road running systematically misestimate trail finish times because they assume a constant energetic cost per kilometre and cannot account for the directional accumulation of terrain-induced speed deviations across every segment of the course.<sup>8</sup>

In adjacent ultra-endurance disciplines, gradient-boosted machine learning (ML) models have demonstrated predictive utility for group-level race outcomes in 48- and 72-h fixed-distance events,<sup>8,9</sup> yet these approaches remain population-level and do not transfer directly to the variable-distance, terrain-specific context of trail running. While ML has been broadly applied to sport outcome prediction, with established methodological frameworks for classification tasks such as win/loss forecasting across team sport formats,<sup>10</sup> the continuous regression problem of predicting individual endurance performance from physiological and environmental data presents structurally different challenges that these frameworks were not designed to address. The challenge of modelling trail running performance has attracted increasing research attention, but existing approaches each carry important limitations. Although empirical group-level models capture population-level trends, explaining 81% of segment speed variance across 517 runners using slope, terrain type, they cannot generate athlete-specific predictions in the absence of individual physiological information.<sup>11</sup> Approaches that incorporate laboratory-measured determinants such as maximal aerobic capacity, ventilatory thresholds, or gradient-specific running economy<sup>6,12</sup> improve specificity but require dedicated equipment and testing protocols, restricting their application to small, tested samples. More broadly, even the most comprehensive multi-domain ML frameworks for athletic performance prediction, integrating physiological, biomechanical, and psychological data from laboratory assessments, have achieved  $R^2$  values up to 0.903 while remaining logistically intensive to deploy at scale.<sup>13</sup> Race-progress frameworks such as Trail Running Assessment of Performance<sup>14</sup> enable real-time checkpoint-based prediction but depend on within-race observations and do not generate pre-race estimates. Physics-based approaches grounded in optimal control theory<sup>15</sup> offer mechanistic insight into pacing optimisation but have not been validated in large, heterogeneous athlete populations. A recent study<sup>16</sup> demonstrated the feasibility of a physics-based digital twin calibrated entirely from historical GPS race records, without laboratory assessment, in a single-athlete proof-of-concept (leave-one-race-out CV:  $R^2 = 0.864$ , MAE = 18.2 min, MAPE = 11.1%). However, whether that framework generalises across a diverse athlete population spanning all competitive trail formats remains unknown.

Digital twins, computational representations of individual physical or physiological systems, parameterised from real measurements and used to simulate, predict, and test responses to changes in state or environment, have been identified as a high-potential technology in sport science.<sup>17,18</sup> Previous sport-related applications have focused principally on biomechanical optimisation, injury-risk

assessment, and rehabilitation, typically integrating high-frequency inertial measurement units (IMUs), electromyography, and musculoskeletal simulation pipelines.<sup>19</sup> These systems yield detailed insight into joint loading and tissue mechanics under controlled laboratory or clinical conditions, but they require specialised instrumentation that extends far beyond the GPS devices and self-reported demographics available in large-scale naturalistic race datasets. No digital twin for trail running has been validated at population scale using data available exclusively in competition settings, GPS course files and athlete demographics, across the full spectrum of competitive formats.

The term digital twin, as used here, denotes an individualised computational model calibrated from an athlete's own race history that enables prospective simulation of performance outcomes under varying course and environmental conditions — a definition consistent with its application in endurance sport contexts. The two-layer architecture couples a mechanistic physics model, which encodes domain knowledge about terrain-induced metabolic cost, with an XGBoost stacking layer that corrects systematic residuals arising from individual variation in pacing strategy and terrain-specific efficiency; neither layer alone achieves both physiological structure and individual specificity. The present study introduces a two-layer physics-informed ML pipeline that extends the single-athlete proof-of-concept<sup>16</sup> to a multi-athlete, multi-format population requiring no laboratory assessment, with three main objectives:

- (i) to quantify overall predictive accuracy of a two-layer pipeline relative to a physics layer alone;
- (ii) to characterise how accuracy and prediction bias vary across race formats and total elevation profiles; and
- (iii) to assess the contribution of a physics-layer prediction as a structured informative prior within a ML layer through systematic feature importance analysis.

## Methods

This study adopted a retrospective observational cross-sectional design. No experimental intervention was performed; all performance data were derived from consumer GPS sport watch files recorded passively during competitive trail races. The study was approved by the Ethics Committee of the University of Innsbruck (Certificate 36/2023) and conducted in accordance with the Declaration of Helsinki; written informed consent was obtained from all participating athletes.

### Participants and race dataset

The cohort comprised 62 competitive trail runners (44 males, 18 females; age:  $32.3 \pm 7.1$  years old; height:  $173.8 \pm 9.7$  cm; body mass  $63.3 \pm 9.1$  kg). Participants were recruited through snowball non-probability methods across Spain, Germany,

**Table 1.** Race dataset characteristics by format category.

| Race format | Distance range | N   | %    | Distance mean (km) | Distance range (km) | Gain mean (m D+) | Gain range (m D+) |
|-------------|----------------|-----|------|--------------------|---------------------|------------------|-------------------|
| VK          | uphill-only    | 14  | 2.4  | $5.8 \pm 1.1$      | 3.8–7.2             | $904 \pm 442$    | 487–2018          |
| MT          | <15 km         | 111 | 19.2 | $11.3 \pm 3.1$     | 3.5–15.0            | $990 \pm 601$    | 357–3123          |
| STS         | 15–34.9 km     | 298 | 51.5 | $23.2 \pm 5.2$     | 15.1–34.8           | $1402 \pm 595$   | 449–5290          |
| STL         | 35–59.9 km     | 107 | 18.5 | $45.2 \pm 6$       | 35.4–59.3           | $2528 \pm 768$   | 910–5349          |
| UT          | $\geq 60$ km   | 49  | 8.5  | $87.6 \pm 25.6$    | 60.5–172.4          | $3937 \pm 1504$  | 1382–8894         |
| Total       |                | 579 | 100  | $30.0 \pm 22.5$    | 3.5–172.4           | $1734 \pm 1124$  | 357–8894          |

VK: Vertical Kilometre/uphill-only (loss <100 m, gain  $\geq 400$  m, dist <20 km). MT: Mountain Trail. STS: Short Trail-S. STL: Short Trail-L. UT: Ultradistance. N: number of races. D+: cumulative positive elevation gain. SD: standard deviation.

Austria, and Italy. Competitive level was assessed using the ITRA Performance Index ( $739 \pm 109$ , range 511–944), indicating a predominantly sub-elite to national-level competitive profile. Each athlete provided four self-reported variables: body mass, height, age, and sex. All remaining model inputs were derived from race data: physiological performance parameters (GPS-estimated speed at the second ventilatory threshold [VT2], aerobic efficiency coefficient [ $\alpha$ ], and pacing decay coefficient [ $\mu$ ]) were estimated from GPS race records via leave-one-race-out (LORO) cross-validation (see Section 2.3.2), maximum heart rate (HRmax) was extracted as the peak value recorded across all available races,<sup>20</sup> and race-level variables (distance, elevation gain and loss, altitude profile, and technicality index) were computed from GPS course files. Thermal stress conditions (WBGT) were estimated from ERA5 climate reanalysis data<sup>21</sup> and ambient weather conditions recorded at the race location and date. A total of 579 trail running races from 62 athletes were included in the analysis after quality filtering (see Section 2.2), corresponding to  $9.3 \pm 3.5$  races per athlete. Race characteristics are summarised in Table 1. The dataset covers a broad range of competitive formats, from sub-10-km Vertical Kilometre events to ultra-distance races exceeding 100 km.

Geographically, races were held across 26 countries, with the dataset predominantly concentrated in two European mountain regions: the Iberian Peninsula, accounting for 51.0% of races (Spain 50.3%, Portugal 0.3%, Andorra 0.3%), and the Alpine arc, accounting for 41.0% (Austria 18.3%, Italy 12.3%, Switzerland 4.0%, France 2.2%, Germany 4.2%). A minority of races (8.0%) took place outside Europe, including events in South America (Argentina 2.4%, Colombia 1.6%, Chile 0.3%), Asia (China 1.1%), and other countries. This geographic distribution ensures that the dataset captures diverse course typologies, from the low-to-mid altitude, technical terrain of the Iberian ranges to the high-altitude glaciated routes of the Alps, providing a representative sample of mainstream European competitive trail running.

### Course profile extraction

Raw GPS activity files in Garmin FIT (including gzip-compressed .fit.gz), GPX, and TCX formats were parsed directly from consumer sport watches (Garmin, Suunto,

COROS), extending the single-format barometric extraction pipeline introduced previously<sup>16</sup> to a multi-device, multi-format context. Geographic coordinates were extracted from per-record messages, converting FIT semi-circle units to decimal degrees via the standard factor ( $180/2^{31}$ ). Inter-point horizontal distances were computed using the Haversine formula. Cumulative elapsed time was derived from embedded timestamps and used to compute per-kilometre GPS travel speed.

Raw altitude sequences were smoothed prior to elevation computation using a centred moving-average window of width max (3,  $N/200$ ) samples, where  $N$  is the total number of GPS records in the file. Unlike the fixed 30-s sliding-median filter applied to constant 1 Hz COROS recordings in the predecessor study,<sup>16</sup> this adaptive window scales with recording density and device model, suppressing high-frequency altimetric noise across heterogeneous hardware without attenuating genuine elevation change. This adaptive approach ensures cross-device comparability across the Garmin, Suunto, and COROS units present in the dataset, whose native recording cadences differ. Cumulative elevation gain ( $D+$ ) and loss were computed as the sum of positive and negative differences of the smoothed altitude series, respectively.

The full GPS track was partitioned into consecutive 1-km segments. For each segment, cumulative elevation gain (m  $D+$ ), cumulative elevation loss (m  $D-$ ), mean altitude (m a.s.l.), and mean GPS speed ( $\text{km}\cdot\text{h}^{-1}$ , derived from the timestamp span of GPS records within the segment), were extracted and stored. Summary course features retained per race were: total distance (km), total  $D+$  and  $D-$  (m), mean and maximum altitude (m a.s.l.), number of 1-km segments, and a technicality index approximating the ITRA terrain difficulty scale (1–5) computed as a clipped linear function of the standard deviation of per-point slope grades.

### Predictive model

The present prediction pipeline extends the previously published<sup>16</sup> single-athlete proof-of-concept to a multi-athlete, multi-race framework by integrating two sequential computational layers: (1) a physics-based simulation model

translating individual physiological parameters and course terrain data into finish-time estimates, and (2) an ML correction layer that learns systematic residuals not captured by the deterministic model. The physics model is executed first, and its output serves as a primary stacking feature for the gradient-boosted regressor, allowing the ML layer to correct the structured errors of the physics model without re-learning the underlying biomechanics from scratch.

### Physics layer

**Segment Speed Equation.** The physics model predicts race finish time by simulating speed over each 1-km segment of the GPS-derived course profile. For segment  $s$  with mean altitude  $a_s$  (m a.s.l.) and net slope  $i_s$  (m/m), the predicted segment speed  $v_s$  ( $\text{km}\cdot\text{h}^{-1}$ ) is:

$$v_s = v_{\text{base}} \times f_{\text{alt}}(a_s) \times f_{\text{dur}}(T) \times f_{\text{pace}}(d, D, \mu) \times f_{\text{heat}}(\text{WBGT}) \times (1 - \text{delta\_tech}) / f_{\text{GAP}}(i_s)$$

where  $v_s$  refers to segment speed ( $\text{km}\cdot\text{h}^{-1}$ );  $v_{\text{base}}$  to baseline speed;  $f_{\text{alt}}$  to altitude factor;  $f_{\text{dur}}$  to duration-fatigue factor;  $f_{\text{pace}}$  to pacing factor;  $f_{\text{heat}}$  to heat penalty;  $\text{delta\_tech}$  to technicality penalty;  $f_{\text{GAP}}$  to grade-adjusted pace factor;  $T$  to elapsed time (min);  $d$  to cumulative distance at segment midpoint (km);  $D$  to total race distance (km); and  $\mu$  to pacing coefficient.

The baseline speed  $v_{\text{base}}$  integrates the three athlete-specific calibration parameters with a distance-intensity correction:

$$v_{\text{base}} = VT2 \times \alpha \times f(D)$$

$VT2$  ( $\text{km}\cdot\text{h}^{-1}$ ) is the primary performance anchor;  $\alpha$  (dimensionless) is an aerobic efficiency scalar capturing individual deviations from the population-level VT2-to-race-speed relationship. The distance-intensity correction  $f(D)$  accounts for the higher fraction of VT2 at which athletes race shorter events – an empirical term absent from the original single-athlete model<sup>16</sup>:

$$f(D) = 1 + 0.662 \times \exp(-D/8)$$

The reference points are established as  $D \sim 5 \text{ km} \rightarrow f \sim 1.35$  (VK/short trail);  $D \sim 10 \text{ km} \rightarrow f \sim 1.19$ ;  $D \sim 25 \text{ km} \rightarrow f \sim 1.03$ ;  $D \geq 50 \text{ km} \rightarrow f \sim 1.00$ . Of note, correction vanishes at ultra distances.

Finish time is the sum of segment travel times  $T_s = L_s / v_s$ , where  $L_s = 1 \text{ km}$  is the segment length (shorter for the final partial segment). The duration factor  $f_{\text{dur}}$  and pacing factor  $f_{\text{pace}}$  are updated iteratively as elapsed time and distance accumulate over successive segments.

**Grade-Adjusted Pace Factor.** The grade-adjusted pace factor encodes the energetic cost of running or walking on sloped terrain relative to flat ground. Following the previous digital twin framework for trail running,<sup>16</sup> it is derived from the fifth-degree polynomial energy cost equations for running and walking established previously<sup>3</sup>:

$$C_{\text{run}}(i) = 155.4i \wedge 5 - 30.4i \wedge 4 - 43.3i \wedge 3 + 46.3i \wedge 2 + 19.5i + 3.6$$

$$C_{\text{walk}}(i) = 280.5i \wedge 5 - 58.7i \wedge 4 - 76.8i \wedge 3 + 51.9i \wedge 2 + 19.6i + 2.5$$

$i$ : net slope (m/m), clamped to  $[-0.45, +0.45]$ ;  $C$ : metabolic energy cost ( $\text{J}\cdot\text{kg}^{-1}\cdot\text{m}^{-1}$ ).

The optimal locomotion cost  $C_{\text{opt}}(i)$  selects walking when it is energetically cheaper than running. Empirically, this transition occurs at approximately 20% grade,<sup>3</sup> consistent with the spontaneous gait transitions observed in competitive trail runners on steep uphill:

$$C_{\text{opt}}(i) = \min(C_{\text{run}}(i), C_{\text{walk}}(i)) \text{ if } i > 0.20, \text{ else } C_{\text{run}}(i)$$

This walk-run transition at 20% grade directly addresses a key limitation of the previous trail running digital twin framework,<sup>16</sup> which applied the running cost function uniformly across all gradients and consequently overestimated finish times on steep technical courses with sustained sections above 20% grade – a discrepancy explicitly identified in the error analysis of that study.

The GAP factor normalises  $C_{\text{opt}}$  by the flat-ground running cost ( $3.6 \text{ J}\cdot\text{kg}^{-1}\cdot\text{m}^{-1}$ ), so that  $f_{\text{GAP}}(0) = 1$  on flat terrain and  $f_{\text{GAP}}(i) > 1$  on uphill:

$$f_{\text{GAP}}(i) = C_{\text{opt}}(i) / 3.6$$

**Environmental and Physiological Correction Factors.** Five multiplicative correction factors modulate  $v_{\text{base}}$ , all shared with the architecture previously published<sup>16</sup> unless noted otherwise. Four are applied at the segment level; the heat factor is constant for the entire race.

### Altitude factor

Hypoxia at altitude  $a$  (m a.s.l.) reduces maximal oxygen uptake. The functional form was introduced by previous works<sup>15,16</sup>:

$$f_{\text{alt}}(a) = \max(0.50, 1 - 11.7 \times 10 \wedge -9a \wedge 2 - 4.01 \times 10 \wedge -6a)$$

The factor is applied per segment at the mean segment altitude, capturing the altitude profile along the course. The floor of 0.50 prevents non-physiological predictions at extreme altitudes ( $>8000 \text{ m a.s.l.}$ ).

### Duration factor

Prolonged exercise progressively impairs aerobic capacity. A linear decay with elapsed time  $T$  (minutes) has been

previously used.<sup>15</sup> Note that a previous single-athlete model<sup>16</sup> employed a Banister TRIMP-based fatigue term; the present model replaces this with the linear formulation, which requires no within-race heart-rate data and thus scales to the multi-athlete population dataset:

$$f_{dur}(T) = \max(0.40, (940 - T) / 1000)$$

The constants 940 (min) and 1000 are adopted from a previous work.<sup>15</sup> The former is the fatigue time-constant (the elapsed time at which aerobic capacity would reach zero without the floor), and together they yield an initial factor of 0.94 at race start and a linear decay of  $0.001 \text{ min}^{-1}$ . The factor is updated at each segment so that later segments receive a progressively larger fatigue penalty. The lower bound of 0.40 ensures physiologically plausible speeds for races approaching and exceeding 15 h.

### Pacing factor

Within-race speed relative to mean race speed follows a linear decay with race progress<sup>16</sup>:

$$f_{pace}(d, D, \mu) = \max(0.40, 1 + \mu \times (d / D))$$

where  $d$  is the cumulative distance at the segment midpoint and  $D$  is total race distance. The pacing coefficient  $\mu$  encodes the athlete's characteristic race strategy:  $\mu < 0$  represents positive pacing (progressive deceleration),  $\mu = 0$  even pacing, and  $\mu > 0$  negative pacing.  $\mu$  is one of the three calibrated parameters (Section 2.3.2).

### Heat factor

Thermal stress reduces performance at a rate of approximately 0.5% per °C WBGT above the neutral threshold of 18°C<sup>22</sup>:

$$f_{heat}(WBGT) = \max(0.70, 1 - 0.005 \times \max(0, WBGT - 18))$$

WBGT is assigned at the race level from weather data at the start location and date. The floor of 0.70 limits extreme heat penalties.

### Technicality penalty

Trail terrain complexity is quantified by a technicality index  $T_{ITRA}$  approximating the ITRA difficulty scale (from 1 to 5),<sup>14</sup> computed from the grade standard deviation of the GPS track clipped to,<sup>1,5</sup> following the concept introduced previously.<sup>16</sup> Segments above the baseline level ( $T_{ITRA} = 2$ ) incur a speed penalty relative to the athlete's performance level:

$$\delta_{tech} = \max(0, (T_{ITRA} - 2) \times 0.046 / VT2)$$

Dividing by VT2 makes the penalty proportional to the

athlete's absolute speed, so that technical terrain has a larger relative impact on slower athletes.

### Physics calibration

**Calibration Parameters.** The physics model has three athlete-specific parameters estimated from GPS race history: (i) VT2 ( $\text{km} \cdot \text{h}^{-1}$ ), anchoring the overall performance level; (ii)  $\alpha$  (dimensionless), the aerobic efficiency scalar, capturing race-specific energy economy relative to the population mean; and (iii)  $\mu$  (dimensionless), the pacing decay coefficient, encoding the athlete's characteristic race strategy. In the single-athlete proof-of-concept,<sup>16</sup> VT2 was measured in a laboratory incremental treadmill test and only  $\alpha$  and  $\mu$  were calibrated from race data. The present multi-athlete framework extends this to a three-parameter calibration, treating VT2 as an additional free parameter, because laboratory physiological assessment is not available for most athletes in the population-level dataset. Together, these three parameters constitute the individual performance fingerprint that is kept fixed across all distance predictions for that athlete.

**Grid Search and LORO Cross-Validation.** For each athlete, the optimal parameter triplet (VT2\*,  $\alpha$ \*,  $\mu$ \*) is identified by exhaustive grid search over a discrete parameter space, extending the two-parameter grid used previously<sup>16</sup> to include VT2:

$$VT2 \text{ in } \{8.00, 8.25, \dots, 22.00\} \text{ km}$$

$$\cdot \text{h}^{-1} (49 \text{ values, step } 0.25)$$

$$\alpha \text{ in } \{0.550, 0.575, \dots, 1.075\} (23 \text{ values, step } 0.025)$$

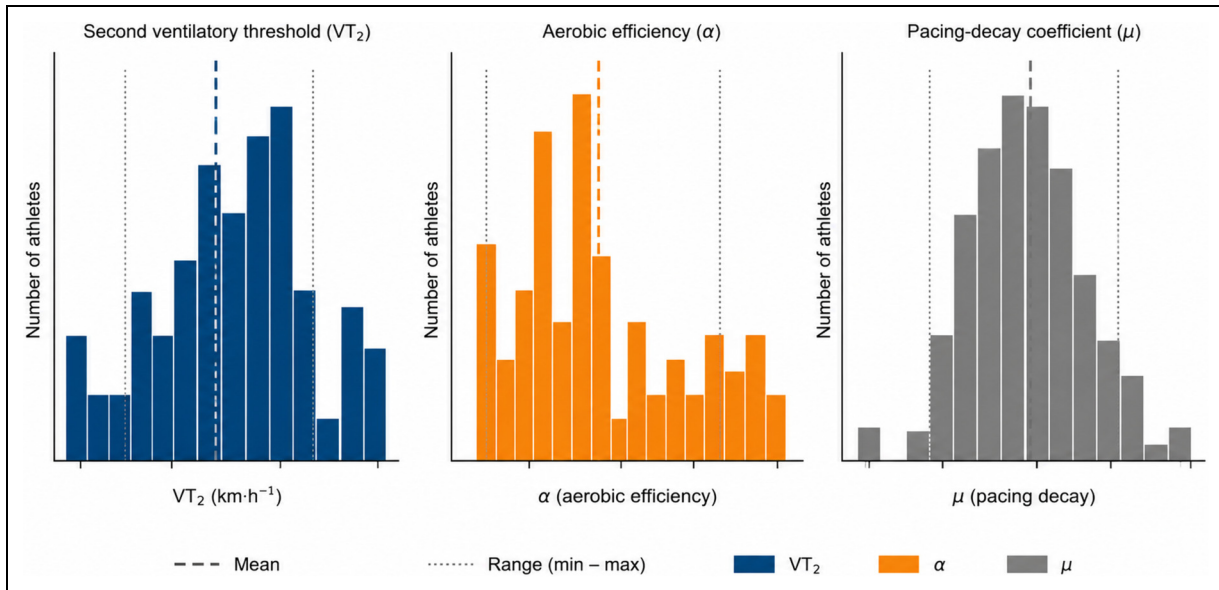
$$\mu \text{ in } \{-0.300, -0.275, \dots,$$

$$+ 0.150\} (15 \text{ values, step } 0.025)$$

The full grid spans  $49 \times 23 \times 15 = 16,905$  parameter combinations. The objective function is the coefficient of determination ( $R^2$ ) between predicted and actual finish times across all available calibration races for that athlete. The in-sample optimal triplet maximises  $R^2$  simultaneously over all races.

Calibration accuracy is estimated with LORO cross-validation: each race  $r_i$  is held out in turn, the model is recalibrated on the remaining races, and the held-out race is predicted. Mean absolute percentage error (MAPE) across held-out races quantifies individual-level accuracy. Athletes with fewer than three annotated races are calibrated in-sample only; LORO is applied when three or more races are available.

**GPS-estimated speed at VT2.** Unconstrained grid search can produce non-physiological (VT2,  $\alpha$ ) pairs in which an implausibly high VT2 is offset by a compensating low  $\alpha$ , or vice versa, a form of parameter confounding



**Figure 1.** Distribution of the three athlete-specific parameters calibrated from historical GPS race data across the full cohort ( $n = 62$ ): second ventilatory threshold speed (VT<sub>2</sub>), aerobic efficiency coefficient ( $\alpha$ ), and pacing-decay coefficient ( $\mu$ ). Histograms illustrate the frequency distribution of each parameter. Dashed vertical lines indicate the cohort mean, whereas dotted vertical lines denote the observed minimum and maximum values. The broad distributions and substantial inter-individual variability support the ability of the GPS-based calibration procedure to generate athlete-specific performance profiles from historical race data alone.

absent from the original model<sup>16</sup> because VT<sub>2</sub> was fixed from laboratory measurement. To stabilise calibration for athletes whose initial in-sample solution yielded  $\alpha \geq 0.95$ , a value more than two standard deviations above the cohort mean ( $0.7 \pm 0.1$ ) and indicative of the optimiser pegging  $\alpha$  against its plausible ceiling while compensating with an underestimated VT<sub>2</sub>, a GPS-anchored recalibration was applied to 12 of the 62 athletes.

For each 1-km GPS segment within uphill slopes of 2–25%, the terrain- and altitude-corrected flat-sea-level equivalent speed is computed:

$$v_{adj} = v_{GPS} \times f_{GAP}(i_s) / f_{alt}(a_s)$$

Uphill segments in this slope range were selected because aerobic demand is dominant – minimal elastic energy return from running mechanics – and the GAP correction is most reliable within the Minetti polynomial's validated domain.<sup>3</sup> The first and last 10% of each race are excluded to remove start-sprint and finish-kick artefacts. The 90th percentile of  $v_{adj}$  pooled across all calibration races of the athlete provides a robust upper estimate of the VT<sub>2</sub>× $\alpha$  product, which pins VT<sub>2</sub> in the subsequent calibration. With VT<sub>2</sub> fixed, only  $\alpha$  is searched (i.e., 23 combinations), eliminating parameter confounding entirely. While ensemble ML models trained on incremental exercise test data can predict ventilatory and anaerobic thresholds with  $R^2$  values of up to 0.803,<sup>23</sup> this GPS-anchored approach provides a laboratory-free alternative to direct CPET-based threshold determination.

The pacing coefficient  $\mu$  is estimated directly from GPS speed profiles via ordinary least squares. Denoting the relative speed at segment midpoint  $x_i = d_{mid,i}/D$  as  $v_{rel,i} = v_{GPS,i}/\langle v_{GPS} \rangle$ , the linear pacing model  $v_{rel} = 1 + \mu * x$  gives the closed-form estimator:

$$\mu_{hat} = \text{Sum}_i[x_i(v_{rel,i} - 1)] / \text{Sum}_i[x_i^2]$$

Per-race estimates are combined as a weighted mean across the athlete's calibration races, weighted by the number of valid GPS segments in each race.

**Distribution of Calibrated Parameters.** To evaluate the physiological plausibility of the calibration procedure, the distributions of the three athlete-specific parameters (VT<sub>2</sub>,  $\alpha$  and  $\mu$ ) were examined across the complete cohort (Figure 1). Mean VT<sub>2</sub> was  $15.1 \pm 2.6 \text{ km} \cdot \text{h}^{-1}$  (range: 9.5–19.8  $\text{km} \cdot \text{h}^{-1}$ ), mean  $\alpha$  was  $0.7 \pm 0.1$  (range: 0.6–0.9), and mean  $\mu$  was  $0.017 \pm 0.037$  (range: –0.10 to 0.10). The distributions were consistent with the broad performance spectrum represented in the dataset and supported the physiological plausibility of the GPS-based calibration framework.

#### Machine learning stacking layer

**Feature Engineering.** After physics calibration, a pooled XGBoost regressor is trained on all ML-eligible race observations from 62 athletes. The model uses 28 input features organised in five groups (Table 2).

**Table 2.** Input features of the XGBoost stacking layer organised by group.

| Feature group          | n | Features                                                                                                             |
|------------------------|---|----------------------------------------------------------------------------------------------------------------------|
| Course characteristics | 9 | phys_per_km, gain_m, loss_m, alt_mean_m, alt_max_m, technicality, gain_per_km, loss_per_km, vert_ratio (gain/loss)   |
| Athlete calibration    | 3 | VT2, alpha, $\mu$ – individual physiological fingerprint                                                             |
| Athlete demographics   | 5 | sex (binary), body mass (kg), height (m), HR_max (bpm), age (yr)                                                     |
| Race conditions        | 1 | WBGT (°C)                                                                                                            |
| Physics stacking       | 1 | phys_per_km – physics model predicted finish time (min)                                                              |
| Derived features       | 9 | log(dist), log(gain+1), log(loss+1), gain/loss ratio, sqrt(dist); binary flags is_vk, is_ultra, is_flat; phys_per_km |

VT2: GPS-estimated speed at the second ventilatory threshold;  $\alpha$ : aerobic efficiency coefficient;  $\mu$ : pacing decay coefficient; HR\_max: maximum heart rate; WBGT: Wet-Bulb Globe Temperature; gain\_per\_km: elevation gain per kilometre; loss\_per\_km: elevation loss per kilometre; vert\_ratio: gain-to-loss ratio; phys\_per\_km: physics-predicted time per kilometre; physics\_pred: finish time predicted by the physics model (Layer 1).

Log-transforms of distance and elevation features reduce the influence of extreme values in right-skewed distributions and help the gradient boosting algorithm find optimal split points. Race-format binary flags (is\_vk for VK and short races with >300 m *D+* and <50 m *D-*; is\_ultra for distances  $\geq 60$  km; is\_flat for gain density <10 m *D+*/km) provide explicit format signals that modulate the physics model's systematic behaviour across race types.

A physics error filter was applied before training: races where  $|\text{physics\_pred} - \text{actual}|/\text{actual} > 0.40$  were excluded from the ML dataset, because in these outliers the stacking feature carries a severely misleading signal that does not generalise. This filter removed 31 races from the 610-race cleaned dataset, yielding 579 races for ML training and evaluation. The 40% threshold was selected as the point at which the marginal gain in training data retention (12 additional races from relaxing to 50%) is outweighed by the risk of including stacking features with severely degraded signal quality; tightening to 30% would exclude a further 16 valid difficult races with no material gain in physics-feature accuracy. The 31 excluded races were concentrated in two format extremes: short, very steep VK and MT races (approximately 76% of exclusions) and ultra-trail races exceeding 60 km (approximately 15%). Excluded races were distributed across both over-prediction (approximately 40%) and under-prediction (approximately 60%), indicating no systematic directional bias that would introduce survivorship effects into the retained training set. No standardisation or normalisation was applied to continuous features prior to training; XGBoost's split-based learning is invariant to feature scale, and the log-transformed features already address right-skew without requiring explicit scaling.

**Table 3.** XGBoost hyperparameter configuration.

| Hyperparameter   | Value |
|------------------|-------|
| n_estimators     | 600   |
| max_depth        | 4     |
| learning_rate    | 0.04  |
| subsample        | 0.80  |
| colsample_bytree | 0.80  |
| min_child_weight | 3     |
| reg_alpha (L1)   | 0.50  |
| reg_lambda (L2)  | 2.00  |
| random_state     | 42    |

n\_estimators: number of boosting rounds; max\_depth: maximum tree depth; learning\_rate: shrinkage applied to each tree's contribution; subsample: fraction of training samples used per tree; colsample\_bytree: fraction of features sampled per tree; min\_child\_weight: minimum sum of instance weights required in a leaf node; reg\_alpha (L1): L1 regularisation term on weights; reg\_lambda (L2): L2 regularisation term on weights; random\_state: seed for reproducibility.

**XGBoost Architecture and Hyperparameters.** An XGBoost gradient-boosted tree ensemble<sup>24</sup> was trained with the hyperparameter configuration shown in Table 3. Tree depth was limited to 4 to prevent overfitting on a dataset of ~600 samples. The number of estimators was set to 600 to exploit the expanded 28-feature input. Stronger regularisation (L1 = 0.50, L2 = 2.00) compensates for the increased feature dimensionality.

**LOAO Cross-Validation.** In each of the 62 folds, all races belonging to one athlete constitute the test set; the XGBoost model is retrained from scratch on the remaining athletes' races and used to predict the held-out athlete's finish times without any parameter sharing. This protocol simulates the real deployment scenario in which predictions are required for a new athlete whose race history was not available during model development, a substantially stricter evaluation than the single-athlete LORO approach used previously.<sup>16</sup> Final LOAO metrics are reported as MAPE and MAE averaged across all held-out races.

**Model performance.** Under LOAO cross-validation on 579 races from 62 athletes, the XGBoost stacking model achieved a MAPE of 11.2% and a MAE of 22.0 min. The physics model alone yielded a LOAO MAPE of 11.7%, confirming a systematic improvement from the ML correction layer across the full race sample. These figures are consistent in magnitude with the leave-one-race-out physics-only MAPE of 11.1% reported for a single elite athlete,<sup>16</sup> suggesting that the model generalises its predictive accuracy from the individual to the population level. Per-athlete results are reported in Supplementary Table S1.

### Statistical analysis

Continuous variables are reported as mean  $\pm$  standard deviation (SD) with minimum–maximum range. Predictive

accuracy was quantified using two metrics: the MAPE (%), defined as  $\text{MAPE} = (1/n) \sum |\hat{y}_i - y_i| / y_i \times 100$ , where  $\hat{y}_i$  and  $y_i$  denote predicted and actual finish times of race  $i$ , respectively; and the MAE (minutes), providing an absolute, practically interpretable measure of error.

Physics calibration accuracy was assessed using LORO cross-validation, in which each race is withheld from calibration in turn and the model is recalibrated on the remaining races. The full two-layer pipeline was evaluated using LOAO cross-validation across 62 folds, in which all races from one athlete constitute the test set and the ML layer is retrained on the remaining athletes. This scheme ensures that neither the physics calibration parameters nor the ML model of the held-out athlete contribute to training, providing an unbiased estimate of generalisation to unseen athletes. To assess the contribution of the ML stacking layer, per-athlete LOAO MAPE was compared between the full pipeline (physics + ML) and the physics layer alone using the Wilcoxon signed-rank test (two-tailed), a non-parametric procedure appropriate for paired comparisons on non-normally distributed data. Effect size was estimated as  $r = Z/\sqrt{N}$ , with benchmarks of  $r = 0.10$ ,  $0.30$ , and  $0.50$  denoting small, medium, and large effects, respectively.<sup>25</sup> Agreement between predicted and actual finish times was further examined using Bland–Altman analysis.<sup>26</sup> For each model (i.e., physics layer and XGBoost stacking), the mean difference (bias) and 95% limits of agreement ( $\text{LoA} = \text{bias} \pm 1.96 \text{ SD of differences}$ ) were computed, where differences were defined as predicted minus actual finish time. Proportional bias, the tendency of prediction error to vary systematically with the magnitude of the outcome, was assessed by computing the Pearson correlation coefficient between the mean of predicted and actual times and their difference. The significance threshold was set at  $\alpha = 0.05$ . All analyses were conducted in Python 3.12 using the SciPy (v1.13), NumPy (v1.26), and XGBoost (v2.1) libraries.

## Results

### Overall predictive performance

The proposed two-layer physics-ML pipeline was evaluated across 579 trail running races from 62 athletes using LOAO cross-validation. Across all race formats, the XGBoost stacking layer achieved a MAPE of 11.23% (MAE = 22.0 min, RMSE = 43.9 min,  $R^2 = 0.938$ ), compared with 11.67% (MAE = 22.6 min, RMSE = 42.7 min,  $R^2 = 0.942$ ) for the physics-only layer. A Wilcoxon signed-rank test on per-race absolute percentage errors confirmed a statistically significant, though small, improvement attributable to the ML stacking layer ( $W = 73.5$ ,  $p = 0.010$ , normal approximation with continuity correction). The XGBoost layer produced lower errors than the physics layer in 38 of 62 athletes (61.3%). The marginally higher  $R^2$  of the physics layer (0.942 vs. 0.938) despite worse MAPE

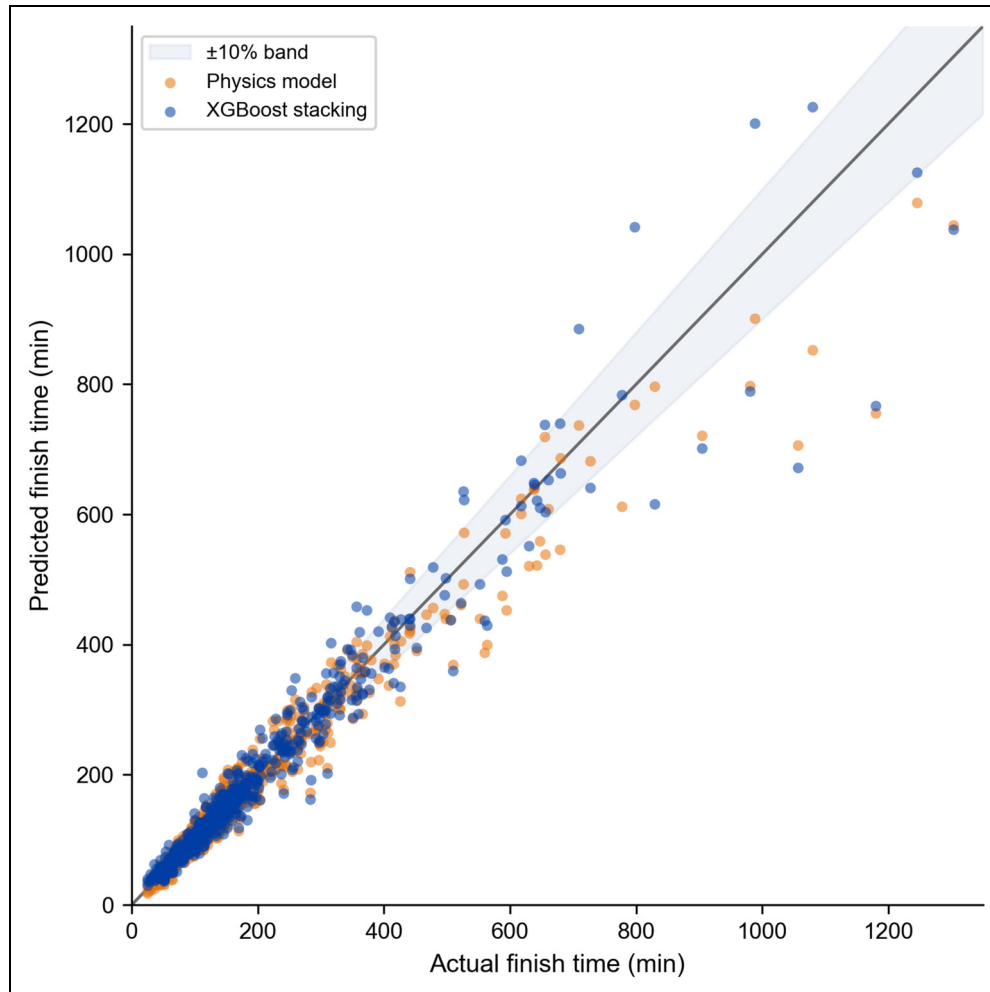
reflects the disproportionate influence of ultra-trail races on total variance: the physics model's predictions for these high-duration, high-leverage events remain strongly correlated with actual finish times even under systematic underestimation, whereas the XGBoost correction of this bias introduces modest additional dispersion in the same races, reducing  $R^2$  while improving calibration across all formats. Predicted versus actual finish times are shown for both layers in Figure 2. For both models, predictions were tightly clustered around the line of identity, with the large majority of races falling within the  $\pm 10\%$  error band.

### Feature importance of the XGBoost stacking layer

Feature importance analysis of the XGBoost stacking model revealed that two predictors accounted for 85.9% of the total explanatory power: log-transformed race distance (log\_dist, 48.3%) and the physics-layer prediction (physics\_pred, 37.6%). The remaining predictors each contributed less than 6.5% individually, with race distance in kilometres (dist\_km, 6.5%), square-root distance (dist\_sqrt, 1.3%), and elevation gain (gain\_m, 1.3%) as the next most informative variables. The cumulative importance reached 95.0% with five features. These results indicate that the physics prediction serves as a strong informative prior that the ML layer refines, rather than replaces. The dominance of log\_dist reflects the proportional scaling of finish time with race length across a 10-fold distance range; the dominance of physics\_pred reflects that the stacking layer receives domain knowledge as a structured input and learns only its systematic residuals, not the underlying biomechanics from scratch. This concentration of importance is therefore a sign that the stacking architecture is functioning as designed.

### Predictive accuracy by race format category

Accuracy varied systematically by race format category. The XGBoost stacking layer achieved its lowest MAPE in Short Trail Long races (STL: 9.9%, 95% CI [8.5–11.5%], MAE = 26.2 min) and Short Trail-S races (STS: 10.3%, 95% CI [9.2–11.5%], MAE = 13.8 min), followed by Ultra Trail (UT: 10.3%, 95% CI [8.3–12.4%], MAE = 64.7 min). Mountain Trail races showed a markedly higher error for the XGBoost layer (MT: 14.1%, 95% CI [11.5–16.4%], MAE = 8.3 min), and the highest error was observed in VK (21.9%, 95% CI [13.5–30.5%], MAE = 8.7 min), likely reflecting the atypical distance-to-elevation profile of this format. The wide confidence interval for VK and its observed range (0.5–53.5%) reflect the inherent instability of category-level estimates at  $n = 14$ ; format-specific conclusions for VK should therefore be interpreted with caution. The physics layer produced lower MAPE than XGBoost only in MT races (12.9% vs. 14.2%), while the XGBoost layer outperformed the physics layer in the remaining four categories, including VK (21.9% vs.



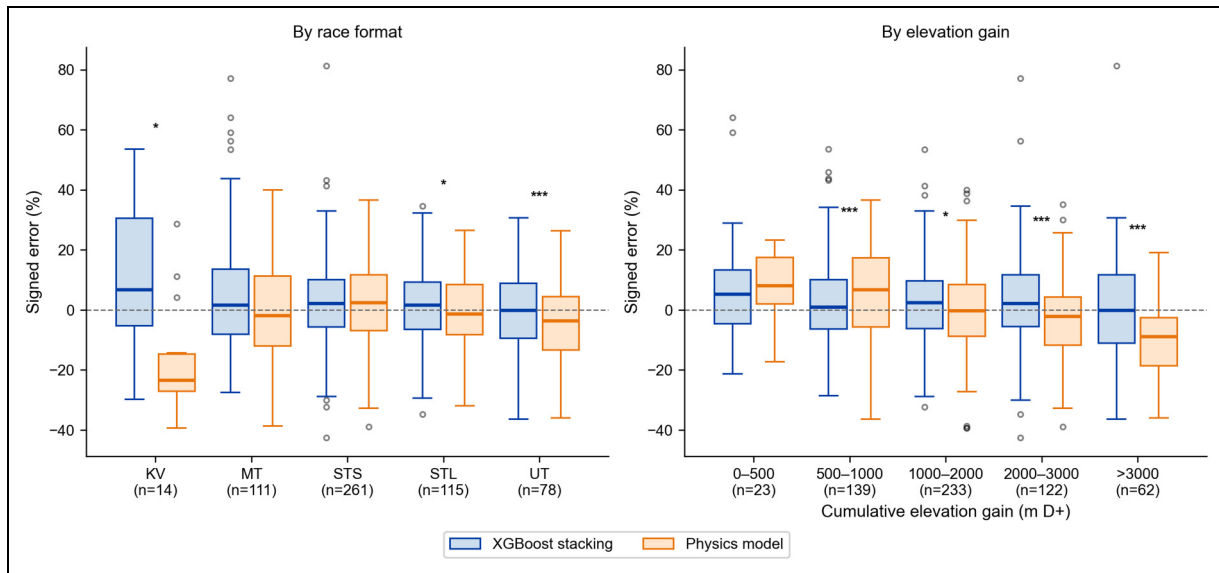
**Figure 2.** Predicted versus actual finish times for all 579 races under LOAO cross-validation. Each point represents a single race; orange: physics model (Layer 1); blue: XGBoost stacking (Layer 2). The diagonal line indicates perfect agreement and the shaded band the  $\pm 10\%$  error zone.

23.4%). Per-category Wilcoxon signed-rank tests on absolute percentage errors revealed a statistically significant difference between the XGBoost and physics layers only for STS races ( $W = 14.1$ ,  $p = 0.015$ ), where the XGBoost layer outperformed the physics layer by 1.09 percentage points (10.3% vs. 11.4%). For VK, MT, STL, and UT races, no significant difference was observed between layers. Signed prediction error (directional bias) by race format and elevation profile is reported in Section 3.4; readers should interpret MAPE and signed error jointly, as high MAPE with low signed bias indicates symmetric error, whereas high MAPE with large signed bias indicates systematic directional failure.

#### Prediction bias by race category and elevation profile

The following analysis characterises the directional component of prediction error for both layers; overall accuracy (MAPE) by format is reported in Section 3.3. Figure 3

displays the signed prediction error (predicted – actual, %) by race format category (left panel) and elevation gain range (right panel), reflecting the systematic directional bias of each layer rather than overall accuracy. By race format, the physics layer showed marked systematic underestimation in VK races (median signed error:  $-23.5\%$ ), with XGBoost substantially reducing this bias (median:  $+6.7\%$ ;  $W = 10$ ,  $p = 0.005$ ). In Ultra Trail races, the physics layer also exhibited strong underestimation of finish time (mean:  $-44.3$  min), whereas the XGBoost layer reduced this to  $-14.3$  min ( $W = 845$ ,  $p < 0.001$ ). Similarly, the physics layer underestimated STL races (mean:  $-5.4$  min) compared to the XGBoost layer ( $+0.8$  min;  $W = 2,538$ ,  $p = 0.026$ ). No significant directional bias difference was observed for MT ( $p = 0.324$ ) or STS races ( $p = 0.587$ ). By elevation gain, the physics layer showed overestimation for races with moderate gain (500–1000 m  $D+$ : mean  $+5.9$  min;  $W$ ,  $p < 0.001$ ) and progressively greater underestimation at higher gain ranges (2000–3000 m  $D+$ : mean



**Figure 3.** Signed prediction error (predicted – actual, %) for the XGBoost stacking layer and physics layer by race format category (left panel) and elevation gain range in metres  $D+$  (right panel). Elevation gain bins are defined as left-closed, right-open intervals: [0, 500), [500, 1000), [1,000, 2000), [2,000, 3000), and  $\geq 3000$  m  $D+$ . Significance markers indicate results of uncorrected Wilcoxon signed-rank tests (\*  $p < 0.05$ ; \*\*  $p < 0.001$ ; ns = not significant); no correction for multiple comparisons was applied across the 10 tests presented. The dashed horizontal line indicates zero error.

–10.2 min,  $p < 0.001$ ;  $>3000$  m  $D+$ : mean –60.4 min,  $p < 0.001$ ). The XGBoost layer substantially corrected these biases across all significant gain ranges (500–1000 m  $D+$ : +0.6 min; 2000–3000 m  $D+$ : +0.7 min;  $>3000$  m  $D+$ : –12.9 min). A significant but small difference was also present in the 1000–2000 m  $D+$  bin ( $p = 0.032$ ). No significant difference was observed for the lowest gain bin (0–500 m  $D+$ ,  $n = 23$ ,  $p = 0.070$ ). No correction for multiple comparisons was applied across the 10 simultaneous tests (5 format categories, 5 elevation bins); under a Bonferroni-corrected threshold of  $\alpha = 0.005$ , the STL format comparison ( $p = 0.026$ ) and the 1000–2000 m  $D+$  elevation bin ( $p = 0.032$ ) would not reach significance and should therefore be treated as exploratory findings.

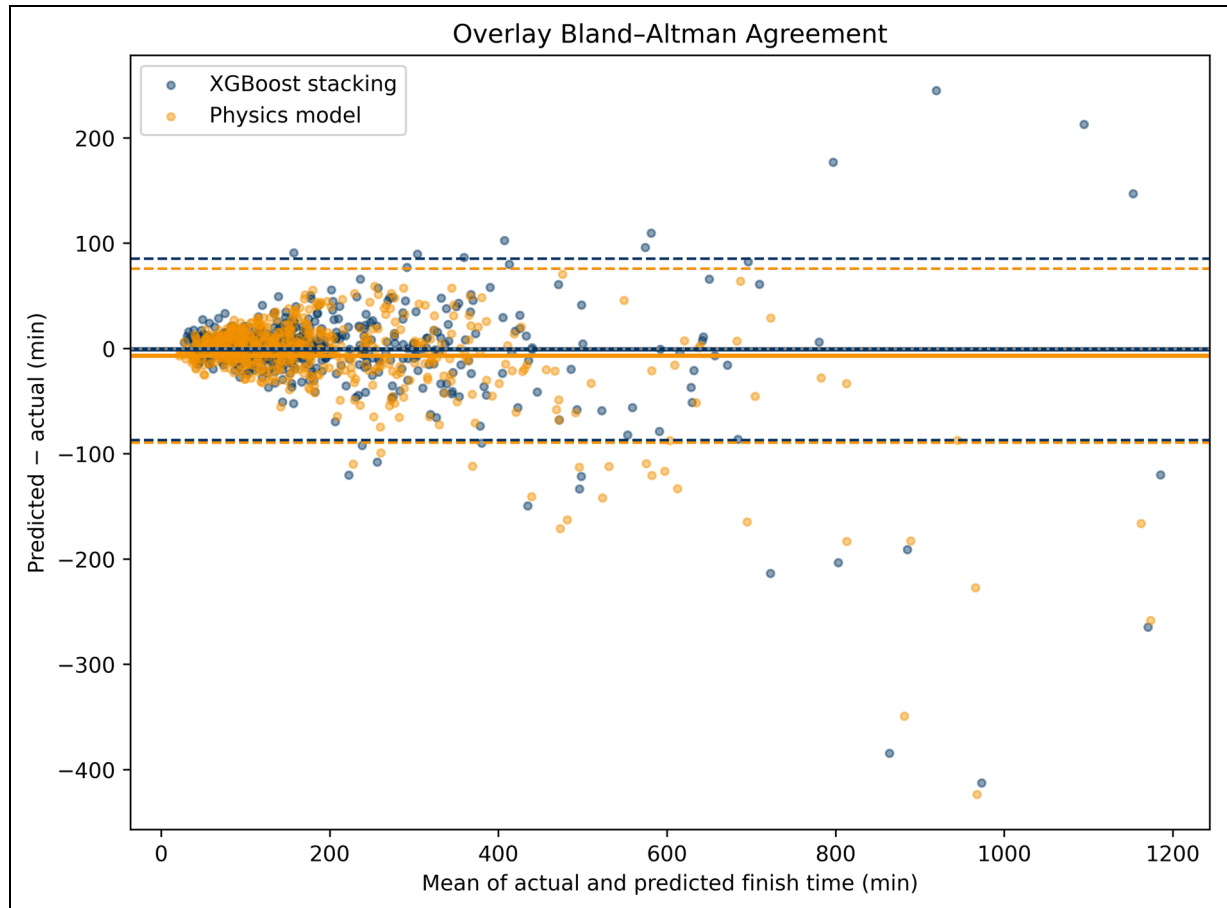
### Bland-Altman agreement analysis

Bland-Altman analysis provided a graphical and quantitative assessment of agreement between predicted and actual finish times for both layers (Figure 4). For the physics layer, the mean bias was –6.8 min (SD = 42.1 min), with 95% limits of agreement (LoA) of [–89.4, +75.7] min. For the XGBoost stacking layer, the mean bias was reduced to –0.9 min (SD = 43.9 min), with LoA of [–87.0, +85.2] min, indicating negligible systematic over- or underestimation. Proportional bias, a systematic relationship between prediction error and the magnitude of finish time, was present in both layers but markedly stronger for the physics model. The Pearson correlation between the mean of predicted and actual times and the signed difference was

$r = -0.564$  ( $p < 0.001$ ) for the physics layer, indicating that the physics model increasingly underestimates finish time as race duration grows. The XGBoost stacking layer reduced this proportional bias substantially ( $r = -0.206$ ,  $p < 0.001$ ), reflecting the capacity of the ML layer to partially correct the distance-dependent underestimation of the physics model.

### Discussion

This study demonstrates that a two-layer physics-informed ML pipeline (Figure 5) can accurately predict trail running finish times across a heterogeneous population of 62 competitive athletes and 579 races spanning the full spectrum of competitive trail formats. These results fulfil the first objective of this study: the two-layer pipeline achieved a MAPE of 11.2% versus 11.7% for the physics layer alone, a statistically significant improvement ( $W = 73.5$ ,  $p = 0.010$ ), with the XGBoost layer outperforming the physics layer in 38 of 62 athletes (61.3%). Under LOAO cross-validation, the most stringent evaluation protocol for population-level generalisation, in which predictions are made for athletes entirely absent from training, the pipeline achieved a MAPE of 11.2%, MAE of 22.0 min, and  $R^2 = 0.938$ , a significant improvement on laboratory based, classic endurance model anchored models in trail running,<sup>6,12</sup> which account for no more than 62–65% of performance variance in ultradistance trail races.<sup>27</sup> Feature importance analysis revealed that the physics-layer prediction alone accounted for 37.6% of the stacking model's explanatory



**Figure 4.** Overlay Bland–Altman plot comparing agreement between predicted and actual finish times for the XGBoost stacking layer and the physics model across 579 trail running races. Each point represents a single race. Solid lines indicate mean bias, dashed lines the 95% limits of agreement ( $\pm 1.96$  SD), and the dotted line zero difference. Positive values indicate overestimation and negative values indicate underestimation of finish time. The XGBoost stacking layer showed reduced bias and improved agreement compared with the physics model.

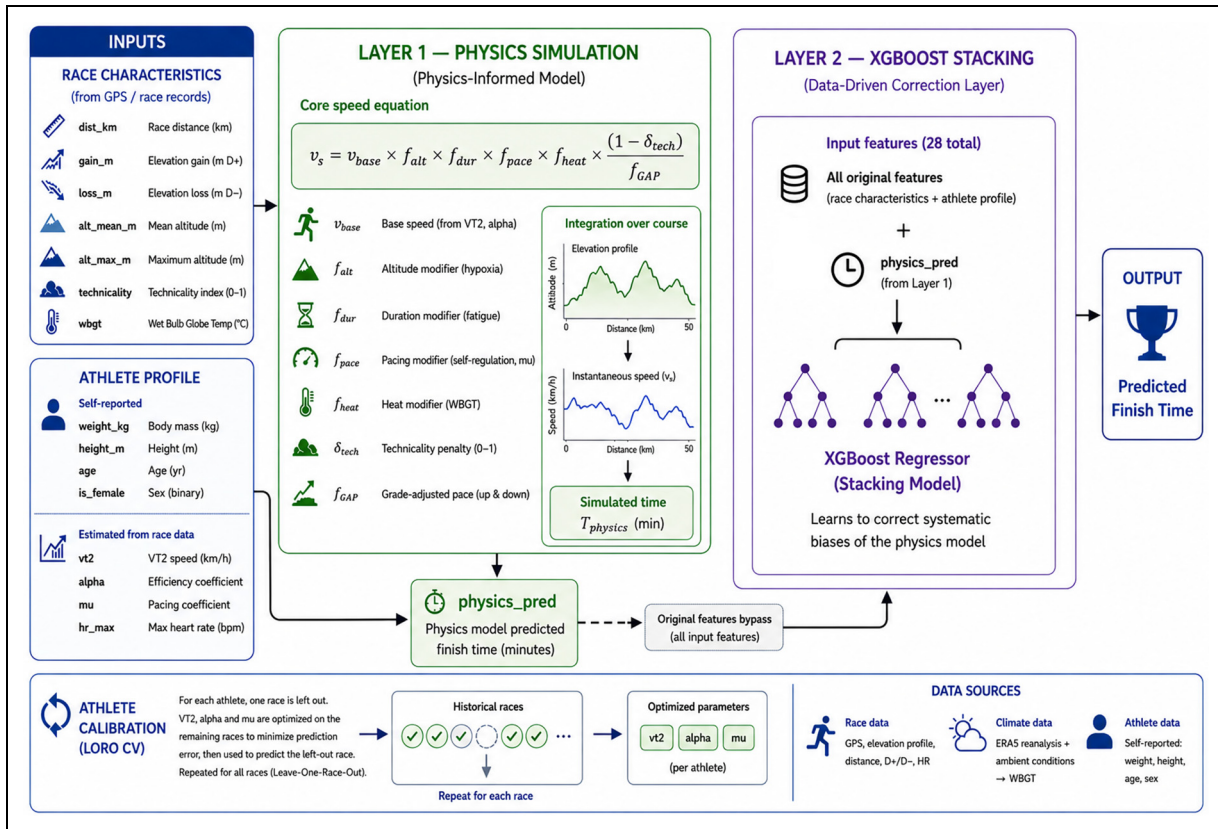
power, second only to log-transformed race distance (48.3%), confirming that the physics layer functions as a strong, structured informative prior that the XGBoost layer refines rather than replaces. The stacking layer corrects the systematic proportional bias inherent to the physics model, whereby prediction error grows with race duration (Bland-Altman  $r = -0.564$ ), reducing it substantially ( $r = -0.206$ ) without requiring any laboratory assessment, as all physiological parameters were estimated directly from athletes' GPS race records.

#### *Population-level validation of a laboratory-free digital twin*

The MAPE of 11.2% achieved across 62 athletes under LOAO cross-validation is particularly noteworthy when contextualised against the single-athlete proof-of-concept study,<sup>16</sup> which reported a MAPE of 11.1% under LORO

evaluation. The LOAO protocol is substantially more demanding, requiring the model to predict finish times for athletes with no representation in the training set, whereas LORO merely withholds individual races from athletes whose physiological parameters have already been estimated from the remaining data. Maintaining equivalent accuracy under this stricter protocol demonstrates that the GPS-derived physiological calibration procedure produces parameter estimates that generalise across individuals rather than overfitting to the idiosyncrasies of specific athletes. The physiological plausibility of the calibrated parameters further supports the validity of the GPS-based calibration procedure.

As shown in Figure 1, VT2 values ranged from 9.5 to 19.8 km·h<sup>-1</sup>, consistent with the wide competitive spectrum of the cohort. The pacing-decay coefficient was centred close to zero ( $\mu = 0.017 \pm 0.037$ ), indicating that systematic fatigue-induced deceleration was minimal on average across the cohort. This finding warrants careful



**Figure 5.** Architecture of the two-layer physics-ML framework. The proposed two-layer framework combines a physics-based simulation layer with an XGBoost stacking layer to predict trail running finish times. Athlete-specific parameters are calibrated from historical race data and integrated with race characteristics to generate an initial physics prediction, which is subsequently refined through ML to produce the final finish time estimate.

interpretation: unlike conventional pacing analyses, which track raw speed as a function of race progression, the present model expresses pace relative to each athlete's VT2 — a threshold that itself diminishes as fatigue accumulates over the course of a race. A value of  $\mu \approx 0$  therefore does not imply the maintenance of absolute speed, but rather a stable effort-to-capacity ratio across race duration whereby athletes slowed proportionally to their declining physiological ceiling. This dynamic, capacity-normalised characterisation of pacing represents a methodologically distinct contribution relative to previous race-analysis frameworks and suggests that, on average, the cohort regulated effort effectively rather than progressively exceeding their fatigue-adjusted capacity. The aerobic efficiency coefficient ( $\alpha = 0.71 \pm 0.10$ ) likely absorbs residual variance from factors not explicitly parameterised in the physics layer, including within-race biomechanical adaptations, terrain-specific movement economy, and individual nutritional strategy, rather than serving as a simple physiological scaling factor. However, because  $\alpha$  is estimated from GPS race outcomes alone, direct attribution of its variance to specific physiological or tactical constructs is not possible without concurrent laboratory or wearable measurement. Its role in the

present framework is therefore best understood as a GPS-calibrated composite parameter that improves physics-model accuracy at the individual level. This stands in contrast to previous group-level modelling approaches, such as the field-based GPS framework,<sup>11</sup> which estimated segment speed at population level ( $R^2 = 0.810$ ) but did not produce athlete-specific finish time predictions.

Integrative ML frameworks that aggregate physiological, biomechanical, and psychological inputs have pushed athletic performance  $R^2$  to 0.903,<sup>13</sup> yet such frameworks require multi-domain laboratory assessment pipelines logistically far beyond what GPS wearables provide. At the parameter level, the GPS-anchored VT2 calibration employed here offers a field-based alternative to direct CPET-derived threshold estimation. While a previous study demonstrated that ensemble ML models can predict anaerobic threshold from incremental exercise tests with  $R^2$  up to 0.803,<sup>23</sup> the present framework achieves VT2 estimation from historical GPS race records alone, requiring no controlled testing environment. The present framework requires only a GPS-enabled wearable device and a sufficient race history, removing the laboratory barrier to personalised performance prediction. From the perspective of digital twin design, this result confirms that athlete-specific, lab-free

calibration from naturalistic GPS records is a viable alternative to physiological profiling, consistent with the call for accessible sport digital twins.<sup>17,18</sup> The first objective of this study, to quantify overall predictive accuracy of the two-layer pipeline relative to the physics layer alone, is fulfilled: the stacking layer achieved a statistically significant though small improvement over the physics layer alone ( $W = 73.5$ ,  $p = 0.010$ ). As such, the novel contribution of the digital twin approach is that it is the first athlete-specific trail running performance model validated at population scale under LOAO cross-validation using exclusively naturalistic GPS race data and requiring no laboratory assessment.

### *The role of the physics model as an informative prior*

Feature importance analysis provided direct evidence that the physics layer functions as a structured informative prior for the ML component, fulfilling the third objective of this study. Log-transformed race distance (log\_dist, 48.3%) and the physics-layer finish time prediction (physics\_pred, 37.6%) together accounted for 85.9% of the XGBoost model's explanatory power, a concentration that indicates the stacking layer is predominantly learning a distance-calibrated correction to the physics prediction rather than independently reconstructing the underlying biomechanics from raw inputs. This interpretation is consistent with the Bland-Altman analysis: the physics layer exhibits substantial proportional bias ( $r = -0.564$ ), whereby underestimation of finish time grows with race duration, and the stacking layer substantially attenuates this bias ( $r = -0.206$ ) using primarily the physics prediction and race distance as inputs. The mechanistic structure of the physics layer, encoding the non-linear relationship between slope and metabolic cost established<sup>3</sup> and the pacing optimisation framework formalised previously,<sup>15</sup> constrains the ML search space to biologically plausible solutions, enabling efficient learning from the limited training data available in an LOAO protocol. This is consistent with the broader consensus in the digital twin and physics-informed ML literature that hybrid architectures outperform purely data-driven approaches when domain knowledge is available and sample size is modest.<sup>17</sup> Purely data-driven approaches applied to the same dataset would face the risk of encoding spurious correlations across the diverse race formats included here, but the physics backbone prevents this by anchoring predictions in known biomechanical constraints. This framework provides the first empirical demonstration, in a competitive trail running population, that a physics-based digital twin can serve as a biomechanically interpretable structured prior for ML, enabling the ML layer to correct systematic residuals without re-deriving physics from data.

### *Performance across trail running formats*

Predictive accuracy varied systematically by race format, fulfilling the second objective of this study. The highest

errors were observed in VK races (MAPE = 21.9%), possibly attributed to their unique uphill-only format and underrepresentation in the dataset (Table 1). The physics calibration procedure estimates VT2, aerobic efficiency, and pacing decay from a race pool dominated by mixed-terrain events, and parameters derived from this pool may therefore be suboptimal for predicting uphill-only efforts where downhill recovery is absent and a walk-run transition strategy is common.<sup>3</sup> MT races also showed elevated error (MAPE = 14.1%), and notably, the physics layer outperformed the XGBoost stacking layer in this format (12.9% vs. 14.1%), suggesting that short, technically demanding races introduce variance that the ML layer cannot readily correct given the limited number of examples in this category. Conversely, STL and STS races showed the lowest MAPE (9.9% and 10.3%), consistent with these formats providing the highest density of training examples and the most typical distance-to-elevation profiles. UT races achieved a MAPE of 10.3% despite their extreme duration, suggesting that physiology-constrained pacing over long distances yields a more predictable outcome than the technical variability of short, steep formats. No significant effect of distance category on group-level segment speed prediction has been previously reported,<sup>11</sup> but our athlete-specific finish time results reveal format-dependent accuracy gradients that only emerge when individual performance is modelled. Therefore, this provides the first systematic, athlete-level characterisation of prediction accuracy across all five ITRA race formats under a population-generalising cross-validation protocol, revealing that calibration-dataset format representation is as important a determinant of prediction quality as the biomechanical demands of the race category itself.

### *Elevation-dependent bias and future model development*

Analysis of signed prediction error by elevation gain profile exposed a systematic reversal in the physics layer's directional bias across the terrain severity spectrum. At moderate elevation (500–1000 m D+), the physics layer overestimated finish time by a mean of +5.9 min; at very high elevation (>3000 m D+), it underestimated by a mean of -60.4 min. This reversal is mechanistically interpretable: at moderate gradients, the physics model may overweight the metabolic penalty of uphills relative to athletes' actual pacing efficiency in their most-practised terrain, while at extreme elevation, the compounding effects of prolonged high-intensity effort, including glycogen depletion, thermoregulatory burden, and neuromuscular fatigue, exceed what a single linear pacing decay coefficient can capture. This finding aligns with the theoretical analysis previously published,<sup>15</sup> which demonstrated that nutrition availability and cumulative fatigue substantially alter the optimal pacing

strategy in ultra-distance trail events and cannot be neglected in accurate finish time modelling. The thermal correction factor, derived from WBGT,<sup>22</sup> partially accounts for environmental heat load, but does not capture race-specific microclimate variability or individual thermoregulatory responses. The XGBoost stacking layer substantially corrected these elevation-dependent biases across all significant gain ranges, reducing the worst-case underestimation ( $>3000$  m  $D+$ ) from  $-60.4$  min to  $-12.9$  min, a correction of 78.6%, without access to any fatigue-related physiological measurements. Incorporating an explicit fatigue-nutrition sub-model in the physics layer would address the mechanistic root cause of this elevation-dependent bias rather than relying on the ML layer for post-hoc correction. Candidate variables for such a sub-model include: (i) cumulative downhill metres ( $D-$ ) as a proxy for eccentric-loading-induced muscle damage, which disproportionately impairs performance in the latter stages of high-relief races; (ii) an estimated caloric deficit, derived from race-specific aid station density and athlete body mass applied to the physics model's cumulative energy expenditure estimate; (iii) a cumulative thermal load index, defined as the integral of WBGT over elapsed race time rather than the single race-level WBGT value used here; and (iv) altitude-adjusted hypoxic stress, combining the instantaneous altitude profile with elapsed time to capture progressive hypoxia-fatigue interaction beyond the static  $f_{alt}$  correction. Each of these variables is, in principle, estimable from GPS course files and publicly available climate data without laboratory assessment, preserving the laboratory-free character of the framework. This demonstrates that ML stacking over a physics model can detect and correct elevation-dependent systematic bias across the full competitive terrain range, while simultaneously identifying where the physics formulation requires mechanistic extension.

### Limitations and future directions

Several limitations should be considered when interpreting the present findings. First, the dataset largely represents a European competitive context, limiting its generalisability. Second, the framework relies on historical GPS race data, which is subject to signal degradation in mountain environments, and device-dependent recording intervals. Although quality filters were applied (Section 2.2), residual positional error may introduce noise into the estimated physiological parameters. The GPS-anchored recalibration of VT2 does not explicitly correct for segment-level trail technicality or thermal conditions; while pooling across multiple races reduces race-specific bias, athletes whose calibration pools are systematically more technical or thermally demanding may carry a small residual underestimate of VT2. Third, athletes with little or no race history cannot currently be calibrated reliably, and calibration accuracy may also

diminish when available records do not reflect current performance level (i.e., when data spans multiple seasons). Prioritising recent race data for calibration is therefore recommended. Future work should explore transfer-learning approaches and population-informed priors to address this cold-start constraint. Fourth, environmental conditions were represented using race-level WBGT estimates, which do not capture local microclimatic variability, altitude-dependent weather changes, or inter-individual thermoregulatory responses. The thermal component warrants caution when temperatures exceed  $18^{\circ}\text{C}$ , above which heat-induced pacing impairment becomes increasingly non-linear. Furthermore, the model does not account for within-race thermal accumulation that disproportionately affects performance in the latter stages of hot-weather events. In addition, the current physics layer does not explicitly account for nutrition, hydration, or cumulative fatigue. The residual underestimation observed in races with very large elevation gain suggests that incorporating dedicated fatigue and nutrition sub-models may further improve predictive accuracy. Then, the XGBoost stacking layer may be subject to format-imbalance overfitting: STS races constitute 51.5% of the training set, while VK (2.4%) and MT (18.8%) races are substantially underrepresented. Despite the inclusion of format-specific binary flags, the model's split-finding is disproportionately shaped by STS-typical distance–elevation profiles, which may partly explain the relatively higher MAPE observed in VK (21.9%) and MT (14.1%) formats. Future work should explore format-stratified sampling or cost-sensitive training to improve calibration across the full competitive format spectrum. Finally, the current framework provides pre-race finish-time predictions and does not update athlete state during competition. Consequently, its primary application is race planning rather than real-time decision support. Advances in wearable technology now allow the collection of biomechanical and physiological information in real-world environments through IMUs, heart-rate sensors, footpods, and muscle oxygenation devices. Emerging microsensor capabilities represent a plausible route to real-time, GPS-synchronised thermal data acquisition, which could substantially reduce reliance on reanalysis-based environmental estimates and address one of the key limitations identified above.<sup>27</sup>

### Conclusion

This study introduced and validated a two-layer physics-informed ML framework for predicting trail running finish times without laboratory assessment. Across 579 races from 62 competitive athletes evaluated under LOAO cross-validation, the model achieved a MAPE of 11.2%, demonstrating that athlete-specific performance prediction can be achieved using historical GPS race data alone. The findings show that GPS-derived physiological

calibration generalizes across athletes and race formats, while the physics-based layer provides a strong mechanistic foundation that can be further refined through ML. The stacking layer reduced systematic prediction bias, particularly in races with large elevation gain and longer durations. Collectively, these results establish this framework as the first population-validated, laboratory-free digital twin for trail running and provide a foundation for future GPS-calibrated digital twin applications in mountain endurance sports.

## Acknowledgements

The authors would like to thank all of the respective athletes included in the analysis for providing their race GPS data for analysis.

## ORCID iDs

Diego Jaén-Carrillo  <https://orcid.org/0000-0003-0588-0871>  
 Arcadi Margarit-Boscà  <https://orcid.org/0009-0006-0035-1427>  
 Simon Jake de Waal  <https://orcid.org/0000-0002-3841-3471>

## Ethical considerations

This study was approved by the Ethics Committee of the University of Innsbruck (Certificate 36/2023) and conducted in accordance with the Declaration of Helsinki; written informed consent was obtained from all participating athletes including permission to publish the (aggregated) findings of the research.

## Authorship contribution statement

[Diego Jaén-Carrillo] collected and curated the data, conducted the analyses, developed the machine-learning model, and contributed to substantially towards drafting the manuscript. [Arcadi Margarit-Boscà] and [Simon Jake de Waal] contributed substantially to the interpretation and narrative development of the work and to drafting and critically revising the manuscript. All authors reviewed and approved the final manuscript.

## Artificial intelligence disclosure

No artificial intelligence (AI) as used to generate content in this article, nor was AI used in any manner that might constitute prohibitive use of AI.

## Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

## Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

## Supplemental material

Supplemental material for this article is available online.

## References

1. Scheer V, Basset P, Giovanelli N, et al. Defining off-road running: a position statement from the ultra sports science foundation. *Int J Sports Med* 2020; 41: 275–284.
2. Viljoen C, Janse van Rensburg DC, Van Mechelen W, et al. Trail running injury risk factors: a living systematic review. *Br J Sports Med* 2022; 56: 577–587.
3. Minetti AE, Moia C, Roi GS, et al. Energy cost of walking and running at extreme uphill and downhill slopes. *J Appl Physiol* 2002; 93: 1039–1046.
4. Jaén-Carrillo D, Gobbo C, Girard O, et al. Acute hypoxia impairs uphill running performance and alters metabolic responses in adult trained trail runners: a pilot study. *Physiol Rep* 2026; 14: e70996.
5. Saunders PU, Pyne DB and Gore CJ. Endurance training at altitude. *High Alt Med Biol* 2009; 10: 135–148.
6. de Waal SJ, Gomez-Ezeiza J, Venter RE, et al. Physiological indicators of trail running performance: a systematic review. *Int J Sports Physiol Perform* 2021; 16: 325–332.
7. Andrade MT, Nunes-Leite MMS, Bruzzi RS, et al. Environmental, physiological, perceptual, and behavioral predictors of mean and final km paces during a laboratory-based self-paced 10 km run in hot conditions. *J Therm Biol* 2026; 136: 104374.
8. Knechtle B, Valero D, Villiger E, et al. Race course characteristics are the most important predictors in 48 h ultramarathon running. *Sci Rep* 2025; 15: 10901.
9. Knechtle B, Villiger E, Weiss K, et al. Analysis of the 72-h ultramarathon using a predictive XG boost model. *Sport Sci Health* 2025; 21: 181–194.
10. Bunker RP and Thabtah F. A machine learning framework for sport result prediction. *Appl Comput Inform* 2019; 15: 27–33.
11. Jidovtseff B and De Cock F. Field-based modelling of segment speed in trail running using GPS and course characteristics. *Int J Perform Anal Sport* 2026: 1–14.
12. Vercruyssen F, Morin J-B, Brisswalter J, et al. Short trail running race: beyond the classic model for endurance running performance. *Med Sci Sports Exerc* 2018; 50: 580–588.
13. Jianjun Q, Isleem HF, Almoghayer WJK, et al. Predictive athlete performance modeling with machine learning and biometric data integration. *Sci Rep* 2025; 15: 16365.
14. Fogliato R, Oliveira NL and Yurko R. TRAP: a predictive framework for the assessment of performance in trail running. *J Quant Anal Sports* 2021; 17: 129–143.
15. Jaszczak B and Płociniczak Ł. Optimal strategy for trail running with nutrition and fatigue factors. *SIAM J Appl Math* 2025; 85: 730–754. DOI:10.1137/24m1629936.
16. Jaén-Carrillo D and Patis D. A physics-based digital twin for trail running race performance prediction: a proof-of-concept study. *Sensors (Basel)* 2026; 26: 3731.

17. Hliš T, Fister I and Fister I Jr. Digital twins in sport: concepts, taxonomies, challenges and practical potentials. *Expert Syst Appl* 2024; 258: 125104.
18. Sédiri A, Barbaria S, Ceylan H, et al. Digital twins in sports science: applications for performance enhancement, injury prevention, and rehabilitation through advanced big data analytics and deep learning methodologies—a comprehensive narrative review. *Biol Sport* 2025; 43: 1271–1291.
19. Saxby DJ, Pizzolato C and Diamond LE. A digital twin framework for precision neuromusculoskeletal health care: extension upon industrial standards. *J Appl Biomech* 2023; 39: 347–354.
20. Kasiak PS, Wiecha S, Cieśliński I, et al. Validity of the maximal heart rate prediction models among runners and cyclists. *J Clin Med* 2023; 12: 2884.
21. Hersbach H, Bell B, Berrisford P, et al. The ERA5 global reanalysis. *Q J R Meteorol Soc* 2020; 146: 1999–2049.
22. Ely MR, Cheuvront SN, Roberts WO, et al. Impact of weather on marathon-running performance. *Med Sci Sports Exerc* 2007; 39: 487–493.
23. Tomaszewski M, Lukanova-Jakubowska A, Majorczyk E, et al. From data to decision: machine learning determination of aerobic and anaerobic thresholds in athletes. *PLoS One* 2024; 19: e0309427.
24. Chen T and Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, 13–17 August 2016. Association for Computing Machinery, 785–794. DOI:10.1145/2939672.2939785.
25. Cohen J. Statistical power analysis. *Curr Dir Psychol Sci* 1992; 1: 98–101.
26. Martin Bland J and Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; 327: 307–310.
27. Camomilla V, Bergamini E, Fantozzi S, et al. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: a systematic review. *Sensors (Basel)* 2018; 18: 873.