Positional drift in the sensor data of an open-field recording: detection from a fixed camera without a camera model

Technical report

October 30, 2026

Abstract

During the processing of an open-field session recorded simultaneously with a fixed video camera, an Xsens inertial suit and a Catapult vest, sensor positions agreed with the video over short stretches of a clip but not over long ones. We test whether this reflects a drift, using the video only as a record of image positions, with no camera calibration, field model or reprojection. For a static camera, the image displacement between two nearby events equals a local mapping **J** applied to the sensor's displacement; for planar motion **J** is exactly quadratic in the pixel position. **J** is fitted on each sensor's own displacements over 0.2 s to 0.5 s, where drift cannot act, and tested on revisits of the same places up to a minute apart, through the correlation and the misfit between image and predicted displacements as functions of time apart. For Xsens footfalls (five clips) the correlation falls from 0.97 to 0.77 and the median misfit grows from 14 to 36 px (growth 20.1 px, 95 % interval [9.0, 35.8]); for the Catapult neck (six clips, X/Y converted to metres by a parameter-free Earth-centred projection) from 0.91 to 0.53 and from 22 to 56 px (29.3 [15.6, 43.8] px). Every clip grows, some weakly. Across the camera ray this corresponds to about 13 cm to 20 cm (Xsens) and 23 cm to 31 cm (Catapult) over runs of about 100 m, while Catapult speed agrees with its positions to within 1 %. Synthetic data show no such growth without drift, and ablations of sync and clock rate, video timestamps, camera motion, non-planarity, lens distortion, keypoint bias and noise, footfall detection and an independent keypoint model do not reproduce it. Neither side is a gold standard; the evidence is consistent with sensor drift in this session, and the conclusions apply to these recordings only.

Contents

1 Introduction **3**

2 Previous work **3**

3 Setup **4**

4 Drift detection method **5**
4.1 Principle 5
4.2 The local image mapping **J** 6
4.3 The correlation criterion 7

5 Results **8**
5.1 Xsens 8
5.2 Catapult 9
5.3 Robustness of the intervals 9
5.4 Clips with both sensors 10
5.5 Drift in metres and comparison with previous work 10

6 Conclusion **12**

7 Ablation study **12**
7.1 A1: can a sync error look like drift? . . . 12
7.2 A2: are the video timestamps (PTS) drift-free? . . . 13
7.3 A3: non-planar motion and lens distortion . . . 14
7.4 A4: is the camera static? . . . 17
7.5 A5: keypoint bias . . . 17
7.6 A6: keypoint noise . . . 18
7.7 A7: footfalls, neck or other body points . . . 19
7.8 A8: pixel–metre (px–cm) against pixel–pixel (px–px) correlation . . . 20
7.9 A9: errors in detecting the footfalls . . . 21
7.10 A10: footfalls and keypoints from an independent model . . . 21
7.11 Summary of the ablations . . . 22

A What drift does to a pair **24**

B Validation on synthetic data **25**

C Catapult coordinates **25**

D The locality limit **26**

E Reproduction **27**

1 Introduction

An open-field training session was recorded at the same time by a fixed video camera, an inertial motion-capture suit (Xsens) and a Catapult tracking vest. During processing, the sensor positions agreed with the video over short stretches of a clip but not over long ones. A stretch aligned to the video left other stretches, seconds or tens of seconds away, visibly off, by an amount whose size and direction changed within a clip. This suggested a time-varying error, a *drift*, on top of short-term noise. Such errors are known for both kinds of device: an inertial suit drifts away from a camera track as a recording lasts longer [5], and a sport satellite receiver gives correctly shaped trajectories displaced by a slowly changing shift [16] (Section 2).

Scope. The report measures whether, and by how much, the agreement between each sensor's positions and the image of the same athlete degrades as the events compared lie further apart in time. Neither side is a gold standard: the sensors have their own errors and the image keypoints come from models. The test measures an inconsistency between the two; attributing it to the sensors is an inference supported by the ablations (Section 7), stated as evidence consistent with sensor drift. The report does not assess quantities derived from the sensors (speed, distance, biomechanical variables), and its evidence and conclusions apply to this session only: the devices as worn, the per-clip sync, the environment and the drills.

Comparing a sensor with a 3D reconstruction from the video would need a calibrated camera (intrinsics, lens model, pose), and errors in any of these can create, hide or imitate a disagreement. Reference systems in validation studies cover only a small area [3, 2], and markerless reconstructions from video depend strongly on the camera set-up [4]. The test of Section 4 instead uses the video only as a record of image positions: no calibration, no field coordinates, no conversion of pixels to positions.

2 Previous work

How positional trackers are validated. A scoping review of 48 validations of satellite (GNSS) and local positioning systems in team sports [1] found distance to be the parameter studied most (34 studies), then average and peak speed, mostly against tape measures, known courses, timing gates or radar guns. Few studies validated instantaneous position during movement. The authors argue that position is the quantity to validate, since distance and speed are derived from it through firmware-dependent filtering and a known course does not record the path actually run. The review excludes inertial sensors.

Inertial motion capture. An inertial suit obtains position by integrating acceleration twice. Unconstrained, a single unit's position error exceeds 150 m after a minute, mostly through gyroscope-driven orientation errors [12]; orientation accuracy degrades within minutes [13], and whole-body accuracy depends on task and duration [14]. Mavor et al. [3] compared an Xsens MVN Link suit with a Vicon system (both 240 Hz) during walking, running and transitions to kneeling and prone: joint-angle RMSE stayed below 9° for the lower limbs but reached 40.5° for the upper limbs, mostly through differing biomechanical models. Trials were cropped to one gait cycle or movement, normalised in time and offset to start at zero; the authors note that inertial systems drift in position over long recordings and realigned the avatar between trials, without measuring the drift. The manufacturer describes position drift as restrained by foot contacts and the biomechanical model, but validates joint angles only [6]. The current specification gives about 1 % of the distance travelled without external aiding, and lists 2D image correspondences among the aiding sources [7].

Boltes et al. [5] compared an Xsens head trajectory with a calibrated overhead camera during rectangular walks: the deviation grew with duration (maximum 0.35 m in 50 s, 0.87 m in 310 s), fusion needed a time-varying rotation, and a 20 ms sync offset raised the fused error by 19 %. In repeated walking between checkpoints, mean Xsens drift was 218.6 mm, and 56.3 mm with one

external tracker [8]. An early Moven test compared start and end positions of closed paths [9], the same return-to-the-same-place logic as here. At sprint speed, step length was underestimated by 0.20 m, more at higher speed [10]; while walking, trunk speed errors concentrated around stops and restarts [11].

Satellite tracking in sport. Ellens et al. [2] validated a recent Catapult device, the Vector S8 (10 Hz, worn between the shoulder blades), against a 22-camera Vicon system over a 10 m × 20 m area of an open field, with sprints, changes of direction, a 70 m circuit and a small-sided game inside the capture area. RMSE was 0.03 m for distance per sample, 0.14 m/s for (Doppler) speed and 0.29 m/s^2 for acceleration, with accumulated distance 1.42 % short on average, under near-ideal satellite conditions; the authors note that no reference system covers a full-sized pitch. Against a geodetic reference, a Catapult OptimEye S5 had a third-quartile horizontal error of 1.04 m with a speed precision of 0.072 m/s [15]. On a football pitch, a GPS unit had a position RMSE of 96 cm while total distance stayed within 2.2 % [17]; a 10 Hz unit in soccer drills had mean absolute errors of 41–48 cm per axis [18], although against a reference that was not an independent optical system. Jølstad et al. [16] describe a standalone sport receiver's error as a gradually changing random shift with little noise: the trajectory keeps its shape and the speed error is very low. Positional errors of 2 m or more can make tactical football metrics unreliable [19].

Correcting drift with external references. 2D pose detections fitted jointly with body-worn inertial units recover their heading drift [20]; calibrated cameras keep the global position of a sparse set of Xsens inertial units drift-free [21]; surveyed magnets reset inertial drift in alpine skiing [22]; one external tracker removes most of a suit's drift [8].

What is missing. These studies report a single error magnitude against a reference, or correct drift with a metric reference (calibrated cameras, surveyed anchors, a reference receiver). The validations closest to this report [2, 3] use a bounded capture area, trials of seconds to two minutes, and relative measures: joint angles, distance per sample, speed, accumulated distance. A slowly varying position error hardly affects these: at 10 Hz a drift of 1 cm/s changes distance per sample by at most 1 mm (against 3 cm RMSE [2]) and speed by about 0.01 m/s (against 0.14 m/s), and joint angles not at all, yet in our synthetic test it raises the median misfit by about 20 px (Appendix B), the size of the growth observed. Such drift is therefore compatible with those validations: it concerns whether a sensor reports the same position for the same place tens of seconds apart, the instantaneous position whose validation is rarely done [1]. In our (non-systematic) search, we found no study that measures how this inconsistency grows with the time between visits to the same place, or that does so without a reference system. This report does both, on a team-sport drill, for both devices on the same clips.

3 Setup

Camera. Sony ILCE-6700 with an E PZ 16–50 mm power-zoom lens on a tripod at the side of the pitch, recording 3840×2160 HEVC at 59.94 frames per second. Each frame carries a container presentation timestamp (PTS); each clip has a sidecar XML with frame count, timecode and wall-clock start.

Xsens. Inertial suit on one athlete: 3D positions of 23 body segments at 60 Hz in the system's own frame. The exported files record MVN 2023.2.0 with HD reprocessing; the suit hardware and the floor scenario are not recorded in them.

Catapult. Vest worn between the shoulder blades (model and firmware not recorded in the exported files): planar X/Y at 10 Hz, a Velocity channel (m/s) and acceleration-derived channels.

X/Y are undocumented and not metres: one unit is about 1.67 m along one direction and 1.0 m across it. This matches an Earth-centred frame with its third axis dropped, which foreshortens ground displacements along the direction from the frame's axis to the venue (Appendix C). One fixed conversion with no fitted parameter turns X/Y into metres for every clip; it is checked against the Velocity channel and against Xsens and used throughout. Samples are never interpolated across gaps and never mapped to the field.

Drill and clips. A cone course whose repetitions include sprints along the same diagonal, so each clip revisits the same places many times, at different times and in both directions. Eight clips, C1–C8 (Table 1); in Catapult clips the analysis starts where the tracked athlete's drill begins.

Table 1: Clips and sensors.

	C1	C2	C3	C4	C5	C6	C7	C8
Xsens	✓	✓			✓		✓	✓
Catapult	✓	✓	✓	✓	✓	✓		

Clocks and sync. The three clocks are independent. Video time is the PTS. Xsens time is the PTS plus one offset per clip; Catapult time is the clip's wall-clock start plus the PTS plus one offset per clip, with a second offset for the lagging Velocity channel. These syncs were set earlier and are kept fixed; A1 and A2 test the offsets, the clock rates and the PTS.

Image keypoints. The video was processed with FitWise AI, a markerless motion-analysis system. The tests use the per-frame 2D joint positions of the tracked athlete from its body-pose model and, in A10, the heel and toe keypoints and ground-contact flags from a second, independent FitWise AI model. Both are used only as image measurements; no calibration, field model or video-based 3D reconstruction enters the tests.

4 Drift detection method

A *displacement* is the difference between two positions of the same point, in the image ($\Delta\boldsymbol{p}$, pixels) or in a sensor ($\Delta\tilde{\boldsymbol{x}}$, sensor units); foot placements are *footfalls*.

4.1 Principle

With a static camera the projection $\pi : \mathbb{R}^3 \to \mathbb{R}^2$ never changes, so two events at the same place appear at the same pixel *however far apart in time*. The video is thus an independent record of position, without extracting positions from it: a drifting sensor reports two positions for one place, the video one pixel.

Since places are never revisited exactly, the test uses *nearby* places: for close events a and b, $\Delta\boldsymbol{p} = \boldsymbol{p}_b - \boldsymbol{p}_a$ is linear in the scene displacement through a local $2 \times d$ matrix $\mathbf{J}$ that depends only on the image position. This relation is learnt on displacements a fraction of a second long, where drift cannot act, and tested on pairs far apart in time. A drift-free sensor obeys it at any time apart; a drifting one departs from it increasingly.

Footfalls and running body points suit this: they are well-timed events that a model or a person can mark; a planted foot is momentarily at rest; and the drill provides both short-gap pairs, which define drift-free behaviour, and revisits far apart in time, from the same athlete, image regions and keypoint model. Other factors can still differ between short and long gaps (image regions revisited, mix of running directions, displacement size). Displacement size is controlled directly (Section 4.3); the others are tested in Section 7.

4.2 The local image mapping J

Planar motion. Let the point move on a plane with coordinates $\boldsymbol{x} \in \mathbb{R}^2$, $\bar{\boldsymbol{x}} = (\boldsymbol{x}, 1)^\top$. A pinhole camera maps the plane to the image by a homography $\mathbf{H}$ with rows $\boldsymbol{h}_i^\top$:

$$\boldsymbol{p}(\boldsymbol{x}) = \frac{1}{w(\boldsymbol{x})} \begin{pmatrix} \boldsymbol{h}_1^\top \bar{\boldsymbol{x}} \\ \boldsymbol{h}_2^\top \bar{\boldsymbol{x}} \end{pmatrix}, \qquad w(\boldsymbol{x}) = \boldsymbol{h}_3^\top \bar{\boldsymbol{x}}, \tag{1}$$

$$\mathbf{J}(\boldsymbol{x}) = \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{x}} = \frac{1}{w} \left(\mathbf{H}_{12} - \boldsymbol{p}\, \boldsymbol{c}^\top \right), \tag{2}$$

with $\mathbf{H}_{12}$ the upper-left 2×2 block of $\mathbf{H}$ and $\boldsymbol{c}^\top$ the first two entries of $\boldsymbol{h}_3^\top$.

J is exactly quadratic in the pixel position. With $\bar{\boldsymbol{p}} = (\boldsymbol{p}, 1)^\top$ and $\boldsymbol{g}_3^\top$ the third row of $\mathbf{H}^{-1}$, $\bar{\boldsymbol{x}} = \mathbf{H}^{-1}\bar{\boldsymbol{p}} / (\boldsymbol{g}_3^\top \bar{\boldsymbol{p}})$, so

$$w = \frac{\boldsymbol{h}_3^\top \mathbf{H}^{-1} \bar{\boldsymbol{p}}}{\boldsymbol{g}_3^\top \bar{\boldsymbol{p}}} = \frac{\boldsymbol{e}_3^\top \bar{\boldsymbol{p}}}{\boldsymbol{g}_3^\top \bar{\boldsymbol{p}}} = \frac{1}{\boldsymbol{g}_3^\top \bar{\boldsymbol{p}}}, \qquad \frac{1}{w} = g_{31}u + g_{32}v + g_{33}, \tag{3}$$

affine in $\boldsymbol{p} = (u, v)$. Since $\mathbf{H}_{12} - \boldsymbol{p}\, \boldsymbol{c}^\top$ is affine too, each entry of $\mathbf{J}$ is a polynomial of degree at most two:

$$\mathbf{J}(u, v) = \sum_{k=1}^{6} \phi_k(u, v)\, \mathbf{C}_k, \qquad \boldsymbol{\phi} = (1,\, u,\, v,\, u^2,\, uv,\, v^2), \tag{4}$$

with constant 2×2 matrices $\mathbf{C}_k$ (u, v normalised to $[-1, 1]$).

The sensor's frame is absorbed. A sensor reporting $\tilde{\boldsymbol{x}} = \mathbf{A}\boldsymbol{x} + \boldsymbol{t}$, with an unknown constant $\mathbf{A}$ (rotation, scale, or anisotropy as in Catapult's raw units), has $\partial \boldsymbol{p} / \partial \tilde{\boldsymbol{x}} = \mathbf{J}\mathbf{A}^{-1}$, again of the form (4). Fitting $\mathbf{C}_k$ against the sensor's own coordinates needs no metric, frame alignment or camera knowledge, and any *constant* frame error is absorbed into $\mathbf{J}$ rather than counted as drift.

Height. For a 3D point and camera matrix $\mathbf{P}$, $\mathbf{J} = (\mathbf{P}_{1:2,1:3} - \boldsymbol{p}\, \boldsymbol{P}_3^\top)/w$ with $w = \boldsymbol{P}_3^\top \bar{\boldsymbol{X}}$, and $1/w$ is affine in (u, v) only at fixed height. For Xsens a 2×3 $\mathbf{J}$ of the same form is used: exact at one height, first-order in the height spread (see Section 4.3 and A3).

Midpoint. Expanding π about the midpoint $\boldsymbol{x}_m$ of a pair gives $\pi(\boldsymbol{x}_b) - \pi(\boldsymbol{x}_a) = \mathbf{J}(\boldsymbol{x}_m)(\boldsymbol{x}_b - \boldsymbol{x}_a) + O(|\Delta \boldsymbol{x}|^3)$, because the second-order terms cancel; $\mathbf{J}$ at either event leaves an $O(|\Delta \boldsymbol{x}|^2)$ error that depends on the order of the events. $\mathbf{J}$ is evaluated at the observed pixel midpoint $(\boldsymbol{p}_a + \boldsymbol{p}_b)/2$, which adds another third-order term. The events themselves (e.g. two footfalls) are unchanged.

Fitting. $\mathbf{J}$ is fitted separately for each source, clip and tracked point, on samples taken every 0.05 s while the athlete runs. Every pair of samples (a, b) between 0.2 s and 0.5 s apart contributes two rows $(\boldsymbol{\phi}_{ab} \otimes \Delta \tilde{\boldsymbol{x}}_{ab})^\top$, with $\boldsymbol{\phi}_{ab} = \boldsymbol{\phi}(\boldsymbol{p}_m)$, to a design matrix $\mathbf{D}$, and the coefficients follow by linear least squares:

$$\hat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}} \sum_{(a,b)} \left\| \Delta \boldsymbol{p}_{ab} - \sum_k \phi_k(\boldsymbol{p}_m)\, \mathbf{C}_k\, \Delta \tilde{\boldsymbol{x}}_{ab} \right\|^2 = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \boldsymbol{y}, \tag{5}$$

with $\boldsymbol{c} = \mathrm{vec}(\mathbf{C}_1, \ldots, \mathbf{C}_6)$ (24 unknowns for 2×2, 36 for 2×3), estimated from tens of thousands of pairs per clip. The 0.2 s to 0.5 s window balances three effects. Drift adds only $\dot{\boldsymbol{d}}\, \Delta t$ (below 2 cm for a few cm/s, against 1 m to 3 m displacements). The third-order model error stays small (Appendix D). And running displacements are tens of times larger than sensor noise, which keeps the attenuation bias caused by noisy regressors small and the same for all pairs.

Foreshortening. The camera sees the pitch at a grazing angle (Figure 1). The singular values of the horizontal block of $\mathbf{J}$ give medians of 148 px/m across the camera ray and 16 px/m along it, a ratio of about 9 for both sensors. Displacements along the ray are nearly invisible, which is why pixel and metric distances cannot be compared directly (A8) and each pair is mapped through its own local $\mathbf{J}$.

Locality limit. Pairs are kept if $|\mathbf{J}\Delta\tilde{\boldsymbol{x}}| < 200$ px (about 1.4 m across the ray), trading model error against the number of revisits. On noise- and drift-free simulated data the median model error at 200 px is about 13 px (Xsens footfalls) and 2 px (Catapult neck); it depends on pair size, not time apart, and the simulated curves stay flat. Conclusions hold for limits from 100 to 300 px (Appendix D).

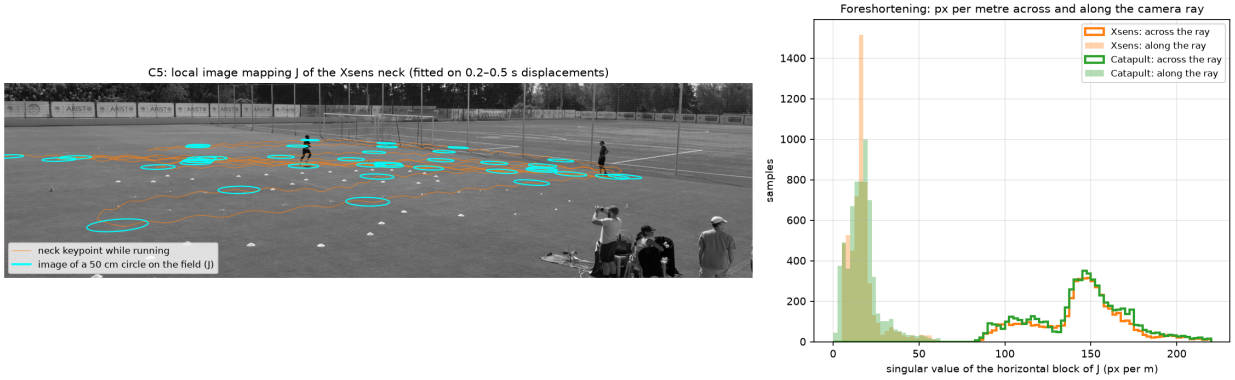


Figure 1: Left: clip C5, image of a 50 cm circle on the pitch under the $\mathbf{J}$ fitted from the Xsens neck, at places the athlete visits. Right: singular values of the horizontal block of $\mathbf{J}$ along the trajectories, Xsens (5 clips) and Catapult (6 clips, in metres).

4.3 The correlation criterion

Drift. The sensor is modelled as the true motion plus a slowly varying shift $\boldsymbol{d}(t)$, a slowly varying yaw about a point, and noise (Appendix A). Drift enters the misfit $\boldsymbol{m}_{ab} = \Delta\boldsymbol{p} - \mathbf{J}\Delta\tilde{\boldsymbol{x}}$ through its change between the two events; its growth with time apart is summarised by the *drift structure matrix*

$$\mathbf{S}(\Delta t) = \mathbb{E}\Big[\big(\boldsymbol{d}(t+\Delta t) - \boldsymbol{d}(t)\big)\big(\boldsymbol{d}(t+\Delta t) - \boldsymbol{d}(t)\big)^{\top}\Big], \tag{6}$$

where $\boldsymbol{d}$ includes the shift produced by yaw. No drift gives $\mathbf{S} = 0$, a random walk $\mathbf{S} \propto \Delta t$, a constant rate $\mathbf{S} \propto \Delta t^2$; a bounded drift levels off.

Expected correlation. Within a bin of time apart, the observed image distance $O = |\Delta\boldsymbol{p}|$ is correlated with the predicted one, $P = |\mathbf{J}\Delta\tilde{\boldsymbol{x}}|$; for an ideal sensor $O = P$ up to noise, whatever the direction or distance. With $\Delta\boldsymbol{p} = \mathbf{J}\Delta\tilde{\boldsymbol{x}} + \boldsymbol{\varepsilon}$ and $|\mathbf{J}\Delta\tilde{\boldsymbol{x}}| \gg |\boldsymbol{\varepsilon}|$,

$$O = P + \boldsymbol{u}^{\top}\boldsymbol{\varepsilon} + O\big(|\boldsymbol{\varepsilon}|^2/P\big), \qquad \boldsymbol{u} = \mathbf{J}\Delta\tilde{\boldsymbol{x}}/|\mathbf{J}\Delta\tilde{\boldsymbol{x}}|, \tag{7}$$

so with $\xi = \boldsymbol{u}^{\top}\boldsymbol{\varepsilon}$ uncorrelated with P,

$$r(\Delta t) = \frac{\sigma_P}{\sqrt{\sigma_P^2 + \sigma_\xi^2(\Delta t)}}, \qquad \sigma_\xi^2(\Delta t) = 2\sigma_n^2 + \sigma_\eta^2 + \boldsymbol{u}^{\top}\mathbf{J}\,\mathbf{S}(\Delta t)\,\mathbf{J}^{\top}\boldsymbol{u}, \tag{8}$$

with σ_n the keypoint noise per axis and σ_η^2 the sensor noise in pixels. Without drift r is the same in every bin; with drift it falls with Δt. Here $\sigma_P \approx 40$ px to 50 px and $\sigma_\xi \approx 5$ px to 15 px, so a drift-free source should give $r \approx 0.95$–0.99 at every time apart.

Range restriction. r also depends on σ_P, which differs between bins. We therefore also report r *matched* to the 1 s to 5 s bin: each pair in bin B is weighted by $w = f_{\mathrm{ref}}(P)/f_B(P)$ (histograms in 20 px bins), and $r_w = \sum w(P-\bar{P}_w)(O-\bar{O}_w)/\sqrt{\sum w(P-\bar{P}_w)^2 \sum w(O-\bar{O}_w)^2}$.

Misfit and margins. Correlating magnitudes misses rotated displacements of unchanged length; the median vector misfit $|\boldsymbol{m}|$ captures every direction, including along the ray. The signed residual $O-P$ shows whether displacements come out too long or too short. Its *margins* are the 2.5th and 97.5th percentiles in the 1 s to 5 s bin, so without drift each bin has about 2.5 % of pairs beyond each margin; the margins are conservative, since drift already present at 1 s to 5 s widens them.

Events and points.

- **Xsens footfalls:** same-foot onsets of Xsens ground contact (foot or toe segment within 3 cm of its floor level and slower than 1 m/s) while the body moves faster than 2 m/s; ankle keypoint against the Xsens foot segment, 2×3 **J** per foot.
- **Neck:** samples every 0.05 s while running; neck keypoint against the mean of the two Xsens shoulder segments, or the Catapult position in metres (2×2 **J**).
- **All tests:** pairs at least 1 s apart with $|\mathbf{J}\Delta\tilde{\boldsymbol{x}}| < 200$ px, in bins of 1 to 5, 5 to 10, 10 to 15, 15 to 25 and 25 s to 60 s; no direction filter.

The neck is sampled at every stride phase, so its height varies over about 20 cm (5th–95th percentile, s.d. 6 cm), with about 142 px per metre of height. The 2×3 **J** handles this to first order for Xsens; for Catapult, which has no height, it adds about 9 px typically and up to 28 px. This error depends on stride phase, not time apart, so it should raise every bin rather than create growth (tested in A3).

Statistics. Pairs share samples and neighbouring samples are correlated, so intervals come from a block bootstrap: blocks of 10 s of one clip, drawn with replacement, each pair weighted by the product of its two blocks' multiplicities (U-statistic bootstrap); 300 replicates for the main results and 100 for the ablations; 95 % percentile intervals. The intervals are set by the blocks (35 for the Xsens footfalls, 39 for Catapult), not by the thousands to hundreds of thousands of pairs. The blocks are not fully independent either: a drift that wanders over tens of seconds links neighbouring blocks of the same clip, so the effective sample size lies between the number of blocks and the number of clips (five and six). Section 5.3 therefore repeats the bootstrap with longer blocks and whole clips and adds a clip-level test. The growth statistics are $G_m = \mathrm{med}\,|\boldsymbol{m}|_{\Delta t \geq 15\,\mathrm{s}} - \mathrm{med}\,|\boldsymbol{m}|_{1 \leq \Delta t < 5\,\mathrm{s}}$ and $G_r = r_{1 \leq \Delta t < 5\,\mathrm{s}} - r_{\Delta t \geq 15\,\mathrm{s}}$, with one-sided bootstrap p-values against no growth. On simulated data (Appendix B) the curves are flat without drift, and a random walk of 1 cm/$\sqrt{\mathrm{s}}$ (about 8 cm per minute) is already detected.

5 Results

5.1 Xsens

For the footfalls (Figure 2, Table 2), r falls steadily from 0.97 at 1 s to 5 s to 0.77 at 25 s to 60 s, and the median misfit rises from 14 px to 36 px: $G_m = 20.1$ px [9.0, 35.8], $G_r = 0.18$ [0.08, 0.28], both $p < 0.003$. The matched correlation gives $G_r = 0.17$ [0.07, 0.27], so this is not a range effect. The neck agrees: 14 px to 39 px, $G_m = 20.6$ px [7.9, 35.8]. At 25 s to 60 s, 21 % of footfall pairs lie above the upper margin and 19 % below the lower one (margins −32 and +22 px, 2.5 % expected): displacements both too long and too short, as expected for a shift in a varying direction rather than a scale error.

The curves grow with time apart without a plateau within a minute, while simulated drift-free data with the same motion, noise and pairs stay flat (Appendix B). Per clip (Figure 4) the growth varies widely: C2 and C8 reach 55 px to 72 px beyond 15 s, C5 grows late to 51 px, and C1 and C7 are nearly flat (13 px to 19 px, 15 px to 29 px).

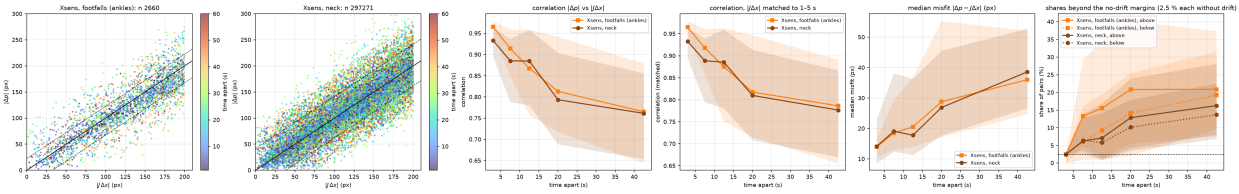


Figure 2: Xsens. Left: $|\Delta \boldsymbol{p}|$ against $|\mathbf{J}\Delta\tilde{\boldsymbol{x}}|$ for footfalls and neck, coloured by time apart (random subsample; solid: identity; dashed: no-drift margins). Right: correlation (plain and matched), median misfit and shares beyond the margins, with 95 % block-bootstrap bands.

5.2 Catapult

Against the neck keypoint (Figure 3), r falls from 0.91 to 0.53 and the median misfit rises from 22 px to 56 px: $G_m = 29.3$ px $[15.6, 43.8]$, $G_r = 0.31$ $[0.20, 0.45]$, both $p < 0.003$; matched $G_r = 0.30$ $[0.19, 0.44]$. At 25 s to 60 s, 24 % of pairs lie above the upper margin and 13 % below the lower one. The higher short-gap level (22 px against 14 px for Xsens) reflects noisier positions and the missing height; it does not depend on time apart. All six clips grow (Figure 4): C5 and C2 to 62 px to 72 px, C1 least, to about 33 px.

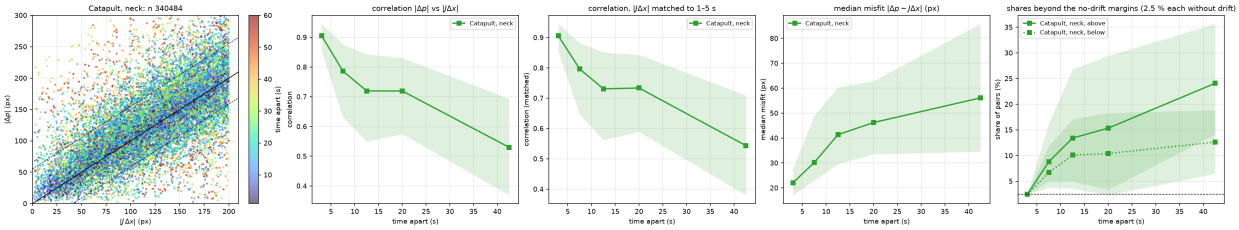


Figure 3: Catapult (metres, 2×2 **J**) against the neck keypoint; panels as in Figure 2.

Table 2: Main results by time apart: correlation r and median misfit (px), with growth statistics and 95 % block-bootstrap intervals (one-sided $p < 1/300$ in all three cases).

		1–5 s	5–10 s	10–15 s	15–25 s	25–60 s
Xsens footfalls	r	0.97	0.91	0.87	0.81	0.77
	misfit	14	19	21	29	36
	pairs	281	362	559	466	972
Xsens neck	r	0.93	0.89	0.88	0.79	0.76
	misfit	14	19	18	27	39
Catapult neck	r	0.91	0.79	0.72	0.72	0.53
	misfit	22	30	41	46	56

G_m (px): footfalls 20.1 [9.0, 35.8]; Xsens neck 20.6 [7.9, 35.8]; Catapult 29.3 [15.6, 43.8]
G_r: footfalls 0.18 [0.08, 0.28]; Xsens neck 0.16 [0.07, 0.25]; Catapult 0.31 [0.20, 0.45]

5.3 Robustness of the intervals

Table 3 repeats the bootstrap of G_m with blocks of 20 and 30 s and with whole clips as the resampling unit, and adds a clip-level sign test: the one-sided binomial probability that at least the observed number of clips grow if growth and no growth were equally likely in each clip. The intervals barely change with the block length and stay above zero. At clip level the evidence is strongest for Catapult, where all six clips grow ($p = 1/64 \approx 0.016$). For Xsens all five clips grow ($p = 1/32 \approx 0.031$), but two of them only weakly: C1 by 3 px and C7 by 13 px for the footfalls (Figure 4). With five or six clips, a whole-clip bootstrap is coarse, and the sign test is the more honest summary of the clip-level evidence.

Table 3: Growth G_m (px): 95 % bootstrap intervals with blocks of 10, 20, 30 s or whole clips, and clip-level sign test (per-clip G_m in clip order).

	10 s	20 s	30 s	clips	clips growing (one-sided p)
Xsens footfalls	[9.0, 35.8]	[8.0, 39.1]	[6.6, 38.8]	[5.9, 40.5]	5/5 (0.031): 3, 65, 35, 13, 25
Xsens neck	[7.9, 35.8]	[6.1, 39.0]	[7.8, 33.9]	[8.6, 34.1]	5/5 (0.031): 6, 41, 34, 15, 22
Catapult neck	[15.6, 43.8]	[14.5, 43.9]	[14.5, 42.6]	[15.9, 37.1]	6/6 (0.016): 12, 34, 30, 29, 37, 15

5.4 Clips with both sensors

In C1, C2 and C5 both sensors are compared with the same frames, so an image-side problem (camera motion, video timing, keypoint artefacts) would affect them alike. Instead, in C1 the Xsens footfalls are nearly flat (13 px to 19 px) while Catapult grows (21 px to 33 px); in C2 both grow strongly; in C5 Xsens grows late while Catapult is already high at 10 s to 15 s. This is consistent with the drift coming from the sensors rather than the video.

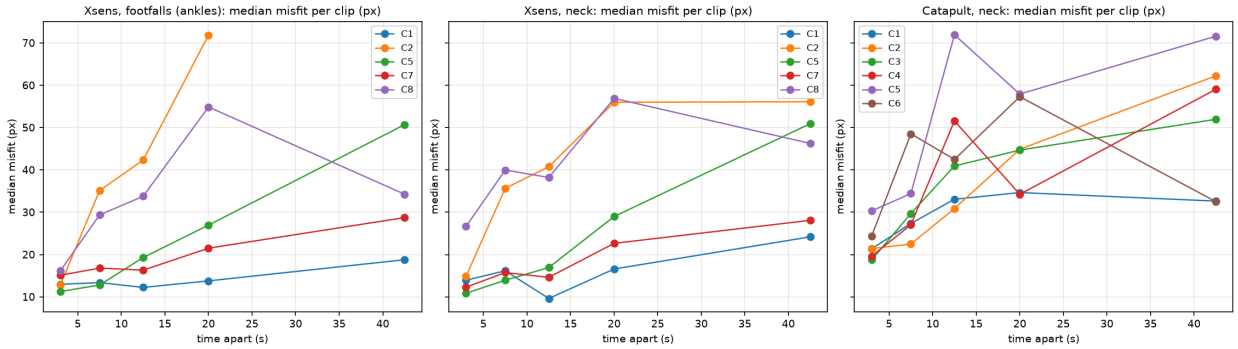


Figure 4: Median misfit by time apart for each clip.

5.5 Drift in metres and comparison with previous work

To compare with the literature, each misfit is converted to the drift component across the camera ray. With the singular value decomposition $\mathbf{J} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^\top$ of the horizontal block at the pair's midpoint (left and right singular vectors $\boldsymbol{\ell}_1, \boldsymbol{r}_1$, largest singular value $\sigma_1 \approx 148$ px/m) and $\boldsymbol{m} \approx -\mathbf{J}\boldsymbol{D}$ for a drift change $\boldsymbol{D}$, the component across the ray is $\boldsymbol{r}_1^\top \boldsymbol{D} = -\boldsymbol{\ell}_1^\top \boldsymbol{m}/\sigma_1$, in metres. No camera parameter is involved: $\mathbf{J}$ is fitted against the sensor's own coordinates, so these are the sensor's own metres. The across-ray value is approximate in both directions: noise biases the fitted $\mathbf{J}$ low, which makes it slightly high, while other effects below make it low. Only the magnitude of the total drift is bounded from below, because the along-ray component, nine times less visible, is left out.

The 1 s to 5 s median is treated as a noise floor and removed in two ways: by subtracting the medians, which needs no assumption, and in quadrature, $\sqrt{\text{far}^2 - \text{floor}^2}$, which assumes independent, roughly Gaussian floor and drift terms. This is why the metric values exceed a direct conversion of G_m: for example, 20.1 px/148 px/m ≈ 14 cm. The floor is not pure noise: the 1 s to 5 s pairs all run in opposite directions, so lag and heading terms inflate it (A5, A9), and it already contains 1–5 s of drift. Both effects make the removal conservative. Each pair is also given the distance run between its events along the sensor's own path (Xsens neck; Catapult position). At sprint speed Xsens tends to underestimate step length [10], so its distances may be short. Expressing drift as a percentage of distance also assumes that it grows with distance, which is not the case for a random walk (growth like $\sqrt{\Delta t}$); the percentages are for comparison only (Figure 5, Table 4).

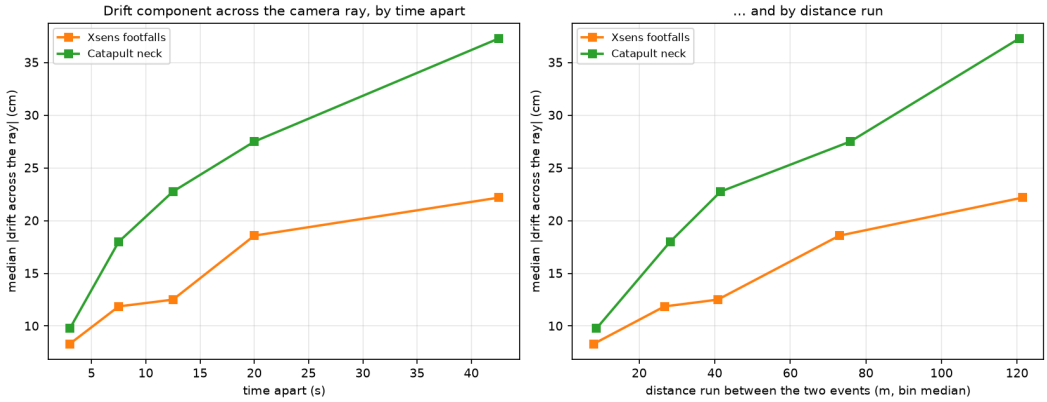


Figure 5: Median drift component across the camera ray, by time apart and by distance run between the two events (bin medians).

Table 4: Drift-related numbers from this recording and from previous work. The statistics differ, so the comparison is of orders of magnitude only. Ours: change of drift between revisits, one horizontal component, noise floor removed by subtraction / in quadrature.

Source	Setting	Statistic	Result
This report, Xsens	running drill, revisits ≥ 15 s apart	median, one component of a change	13 cm to 20 cm over a median 97 m run (0.13–0.20 %)
This report, Catapult	same drill	same	23 cm to 31 cm over 99 m (0.23–0.32 %)
Xsens specification [7]	no external aiding	not defined	∼1 % error in travelled distance
Boltes et al. [5]	Xsens, walking, 50 s / 310 s	maximum deviation	0.35 m / 0.87 m, growing with time
Chebel et al. [8]	Xsens, repeated walking	mean drift	218.6 mm
Gløersen et al. [15]	Catapult S5, skiing, geodetic reference	third quartile of absolute error	1.04 m; speed IQR 0.072 m/s
Linke et al. [17]	GPS, football pitch, Vicon	RMSE of absolute error	96 cm; distance within 2.2 %
Jølstad et al. [16]	sport GNSS, skiing	qualitative	slowly changing shift, low noise and speed error

Xsens. Between revisits at least 15 s apart the across-ray drift is about 13 cm to 20 cm (subtraction to quadrature) over a median 97 m run, 0.13–0.20 % of the distance for one component. This is small compared with the ∼1 % error figure in the specification [7], whose exact definition the manual does not give, and of the same order as the walking results of Boltes et al. [5] and Chebel et al. [8]. As in Boltes et al., the deviation grows with time and the rotation of the sensor frame wanders (A3, Appendix A); unlike them, we need no calibrated camera, and the athlete runs. Possible mechanisms, not tested here, include speed-dependent displacement errors at sprint speed [10], errors around stops and turns [11] and heading errors [12, 13].

Catapult. About 23 cm to 31 cm across the ray over a median 99 m. Published sport-GNSS errors are around a metre [15, 17], but they are a different statistic: absolute errors include the session's constant offset, which **J** absorbs, whereas ours is one component of the change within a minute. Our observations match the pattern described by Jølstad et al. [16]: positions drift while path shape and speed stay accurate. Here the Velocity channel agrees with the speed of the converted positions to within 1 % in every direction (Appendix C) while positions drift by tens of centimetres, so validations of speed and distance alone, such as Ellens et al. [2], would not reveal this drift. Changing satellite geometry is a possible mechanism [15], not tested here.

6 Conclusion

Findings. In this session, the evidence is consistent with drift in the positions of both sensors. Their agreement with the image of the athlete is very good for events a few seconds apart and degrades steadily with time apart: the median misfit grows by about 20 px (Xsens) and 29 px (Catapult) between revisits 1 s to 5 s apart and more than 15 s apart. The growth is significant, differs between clips and between the two sensors on the same clip, is absent in drift-free simulation, and is not reproduced by any alternative tested in the ablations. In metres, the across-ray drift between revisits at least 15 s apart is about 13 cm to 20 cm (Xsens) and 23 cm to 31 cm (Catapult) over runs of about 100 m, depending on how the noise floor is removed; the along-ray component is not measured, so the total drift is larger. For Xsens this is small compared with the ∼1 % figure of the specification, whose exact definition the manual does not give.

What this means in practice. On this recording, positions tens of seconds apart cannot be compared at the few-centimetre level: their difference typically carries tens of centimetres of drift, and much more in some clips (C2, C8). Over a few seconds positions match the image within keypoint noise, and Catapult's speed agrees with its positions to within 1 % (Appendix C). Quantities built from short spans of motion are therefore little affected; quantities comparing positions over longer spans, or aligning a whole clip with one fixed transform, are affected.

Limitations. One session; five Xsens and six Catapult clips of one to two minutes, hence few independent units and wide intervals; at clip level the Xsens evidence rests partly on clips that grow only weakly (C1, C7). Only the across-ray component is measured in metres. The keypoints are model outputs; their influence was tested (A5, A6, A10) but cannot be fully excluded.

Method. The test uses the video only as a record of image positions: no intrinsics, lens model, camera pose, field model or reprojection. The image–sensor relation is learnt from the sensor's own short displacements, absorbing every constant property of its frame (origin, orientation, scale, anisotropy), and the metric scale of the drift estimate is the sensor's own. Its assumptions (static camera, locally smooth mapping, correct relative timing) were tested in Section 7 together with the keypoint models and event detection. Since nothing estimated from the video enters, the test adds little bias of its own: it measures a sensor's inconsistency with itself over time, as seen by a fixed camera. It complements validations against a reference over short trials in a small area by testing the consistency of instantaneous position that such validations rarely address [1].

7 Ablation study

7.1 A1: can a sync error look like drift?

A constant lag δ moves each sample along its velocity, $\tilde{\boldsymbol{x}}(t) \approx \boldsymbol{x}(t) + \delta\, \boldsymbol{v}(t)$, adding $\mathbf{J}\delta(\boldsymbol{v}_b - \boldsymbol{v}_a)$ to the misfit: dependent on the two velocities, not on the time between them. A clock-rate error ε gives $\delta(t) = \varepsilon(t - t_0)$ and adds $\mathbf{J}\varepsilon[(t_b - t_0)\boldsymbol{v}_b - (t_a - t_0)\boldsymbol{v}_a]$, i.e. $\mathbf{J}\varepsilon\Delta t\, \boldsymbol{v}$ for same-speed, same-direction revisits, which does grow with time apart.

The tests were rerun with each sensor's clock shifted by −80 ms, −40 ms, 40 ms and 80 ms and with rate errors of 200 ppm and 1000 ppm (Figure 6). On recorded data the pinned sync gives the lowest misfit in every bin; 40 ms raises the curves (footfalls at 1 s to 5 s: 14 px to 19 px to 24 px) but keeps the growth. On simulated drift-free data lags raise the level without a trend, lowest at 10 s to 15 s because the lag term cancels for same-direction pairs, about half of that bin (Table 6). 1000 ppm, far more than measured in A2, barely changes the recorded curves (Catapult 22 px to 56 px becomes 29 px to 57 px) and adds no growth in simulation. In these tests, sync errors did not produce growth of the kind observed.

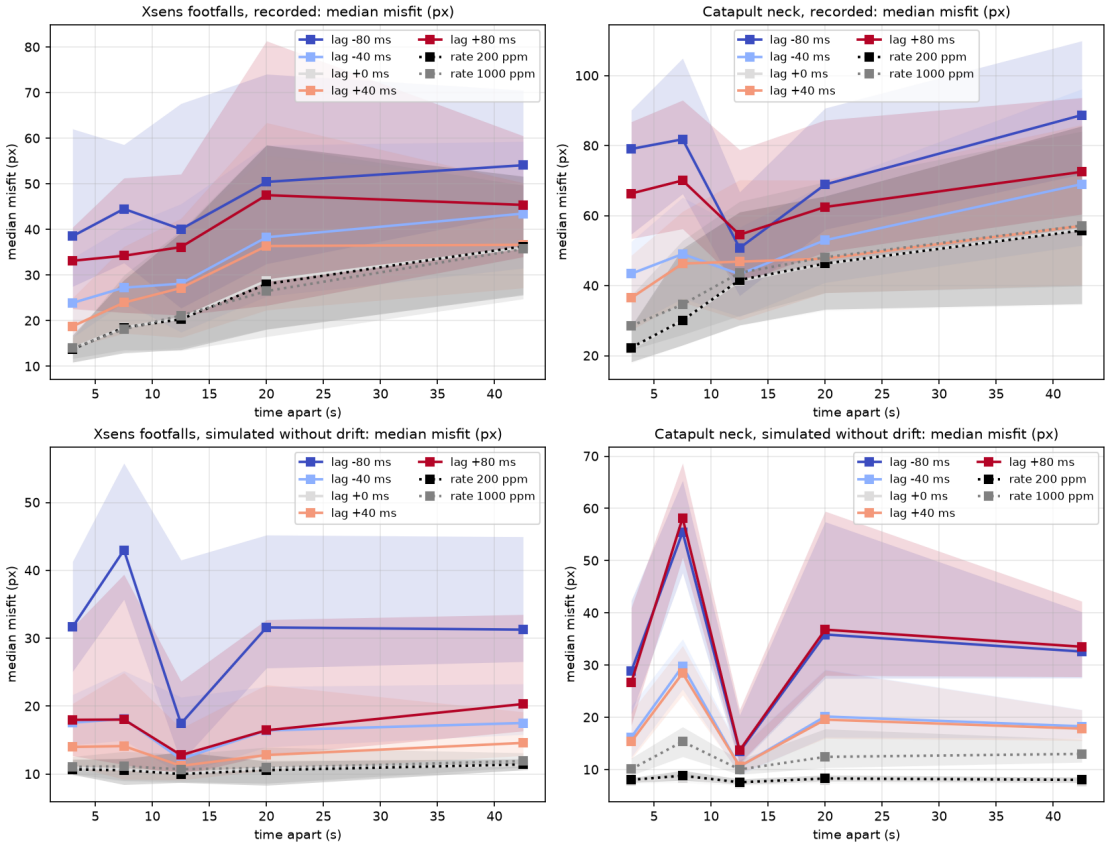


Figure 6: A1. Median misfit with the sensor clock shifted (solid) or at a wrong rate (dotted). Top: recorded data. Bottom: simulated without drift.

7.2 A2: are the video timestamps (PTS) drift-free?

The file reports every frame exactly at $k \cdot 1001/60000\,\mathrm{s}$ and a decoded frame count equal to the XML duration; these claims are tested with the keypoints.

Local defects. A dropped, repeated or mistimed frame shifts every joint along its motion. Around each running frame k, the torso keypoints (neck, hips, shoulders) over ± 6 frames are fitted by a cubic in the frame index n with one shared timing parameter d: $\boldsymbol{p}(n + d\,H(n)) \approx \boldsymbol{p}(n) + d\,H(n)\,\boldsymbol{p}'(n)$, with H the indicator of $n = 0$ (one mistimed frame) or $n \geq 0$ (a drop or repeat). A frame is flagged when $|\hat{d}| \geq 0.5$ frame and at least 4 standard errors. Injected defects are detected at 99 % (drops) and 100 % (repeats). Of 29 476 running frames, 83 (0.3 %) are flagged, but their $\hat{d}$ are scattered rather than near ± 1, and none coincides with a repeated image: the decoded grey-level difference is normal at every flagged frame, and no clip has a frame below half its local median difference. They are isolated keypoint outliers, inconsistent with timing defects.

Rate. A uniform PTS rate error is invisible within the video; it appears only as a lag growing against another clock. Keypoint events are timed against the two independent sensor clocks: Xsens footfalls against video stance onsets (ankle keypoint speed below 30 % of its local peak, sub-frame); Catapult turns (extrema along the run axis) against neck-keypoint turns; and, per 10 s window, the lag best matching the neck pixel velocity with $\mathbf{J}$ times the sensor velocity, which slow position drift hardly affects. The Huber slope of lag against time, pooled within clips and optionally adjusted for running speed, is the relative clock rate (Table 5, Figure 7).

Table 5: A2: relative clock rate between video PTS and each sensor (ppm, 95 % block-bootstrap interval).

Estimator	rate (ppm)	lag scatter
Xsens footfalls	−153 [−218, −98]	11 ms
Xsens footfalls, speed-adjusted	−71 [−147, −2]	10 ms
Xsens velocity windows, speed-adjusted	−5 [−41, +15]	1.2 ms
Catapult turns	+5627 [+4573, +7033]	98 ms
Catapult velocity windows, speed-adjusted	−44 [−292, +72]	7 ms

The most precise estimator (velocity windows, 1 ms scatter) puts the video clock within −41 ppm to 15 ppm of Xsens. The footfall trend halves once running speed is accounted for, so it is mostly a speed dependence of the footfall definitions (athletes speed up over a clip). Against Catapult the velocity windows give −44 ppm [−292, +72], containing zero, so both independent clocks are consistent with drift-free PTS; Catapult turns are too few and noisy (98 ms) to inform. The worst allowed rate, 292 ppm, moves a sample by at most 14 cm (about 21 px) over 60 s at 8 m/s, a worst case requiring full speed in the same direction at both events; A1 shows that even 1000 ppm adds no growth.

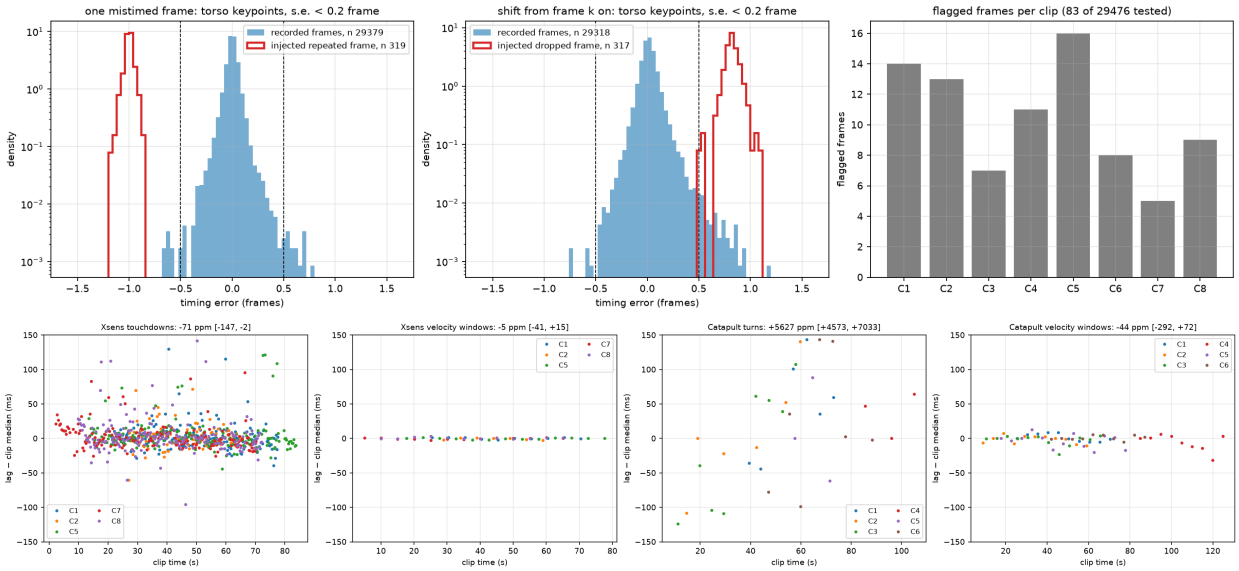


Figure 7: A2. Top: per-frame timing error from the torso keypoints (one mistimed frame; shift from frame k on) for recorded frames and injected defects, and flagged frames per clip. Bottom: lag of keypoint events against each sensor clock (relative to the clip median); titles give the pooled rate.

7.3 A3: non-planar motion and lens distortion

An uneven pitch, a point changing height, or radial lens distortion all change **J** as a function of position or stride phase, not of time: they raise the misfit level, and can create growth only if the places or phases sampled differ systematically between short and long gaps (Figure 8).

- **Simulated, no drift.** A pitch bent by ±10 or ±20 cm (15 by 11 m waves) and distortion $k_1 = \pm 0.15$ raise the level without growth; all G_m intervals contain zero (largest 1.5 $[-3.2, 4.2]$ px, bending plus distortion). The bent pitch dips at 10 s to 15 s, because same-direction revisits a lap apart cross the bumps along the same line and their errors cancel (Table 6).
- **Recorded, other forms of J.** Affine, cubic, or Xsens without height (2×2) shift the level by a few pixels and keep the growth: footfalls 14 px to 36 px (main), 15 px to 37 px (affine),

15 px to 38 px (cubic), 17 px to 37 px (no height). A **J** fitted on the first half of each clip reproduces the main result.

- **Neck height.** The Xsens neck test with height ignored (2×2, as for Catapult) and with only samples within ± 1.5 cm of the median running height (22 % of samples) gives 14 to 39 px (main), 19 to 41 px (height ignored) and 16 to 43 px (constant height): ignoring height adds about 5 px, and growth is unchanged ($G_m = 20.6$, 18.8, 22.3 px).

Fitting **J** on the second half only, or per 20 s window, does worse, even on short gaps and within the fitted image region: a window's **J** rotates relative to the clip's by several degrees over a clip (mostly up to 5° to 9° for Catapult; third row of Figure 8). A static camera (A4) cannot rotate the mapping, so this is the short-displacement signature of a slowly rotating sensor frame (Appendix A); such partial fits carry the drift they are meant to test. The conclusions do not depend on how **J** is modelled, and non-planarity, neck height and lens distortion did not produce the growth in these tests.

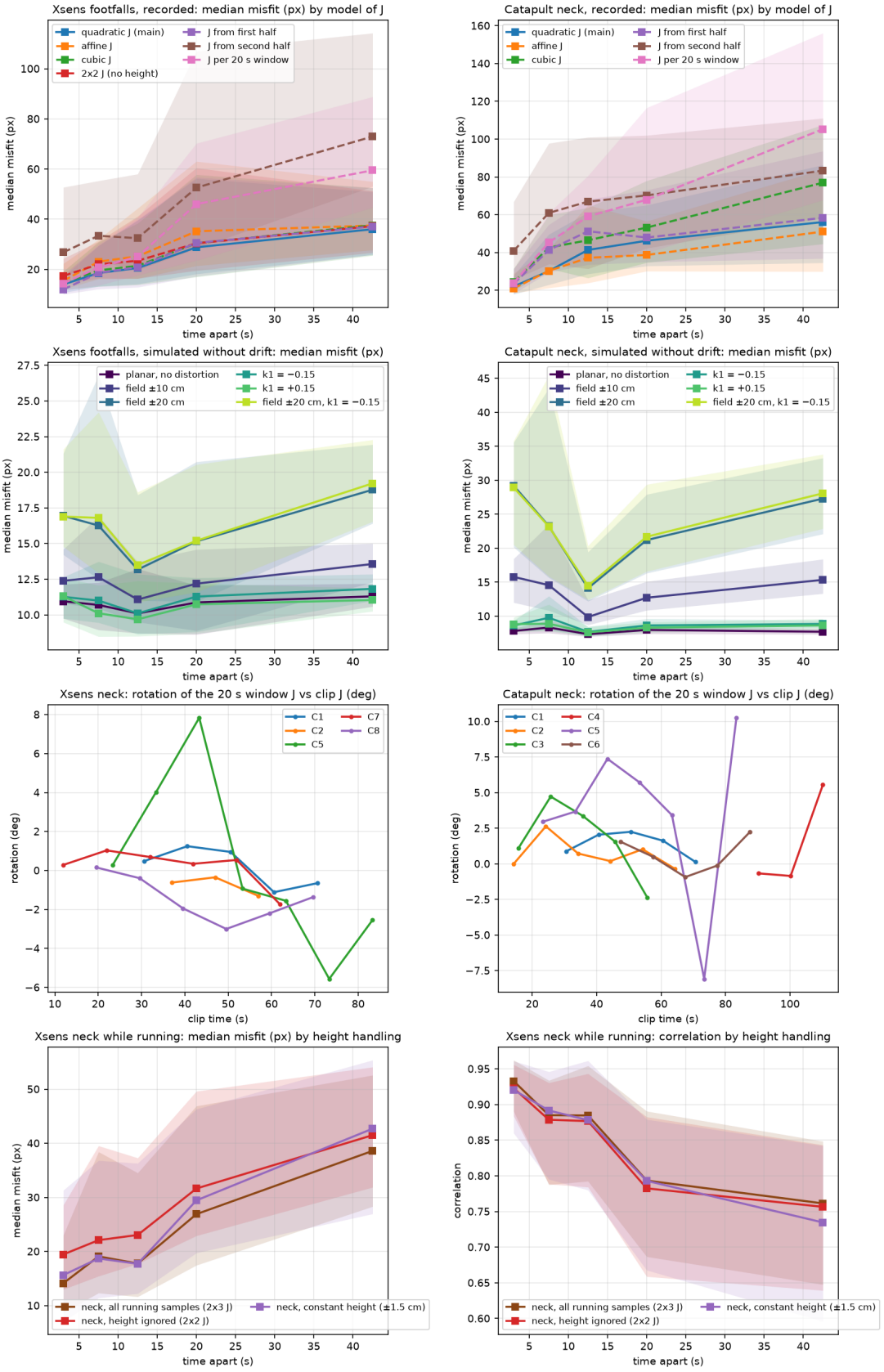


Figure 8: A3. Row 1: recorded data with different forms of **J**. Row 2: simulated without drift, uneven pitch and lens distortion. Row 3: rotation of each 20 s window's **J** relative to the clip's (neck). Row 4: Xsens neck with height modelled, ignored or held constant.

7.4 A4: is the camera static?

Keyframes (about one per second) were registered by SIFT features below the tree line (fences, banners, goals, markings, cones) with a RANSAC affine fit (1 px threshold); displacements over the drill area are measured in pixels of the 3840-wide image against the median over the analysed span (Figure 9). The camera jumps by up to 15 px in the first and last second, when the record button is pressed. In between, the largest displacement is 1.0 px to 3.3 px per clip, scale changes stay below 0.15 % (no zoom) and rotation below 0.04°: an order of magnitude below the 20 px to 30 px growth. Camera motion also could not make the two sensors on one clip differ, as in C1; it is too small to account for the drop in correlation.

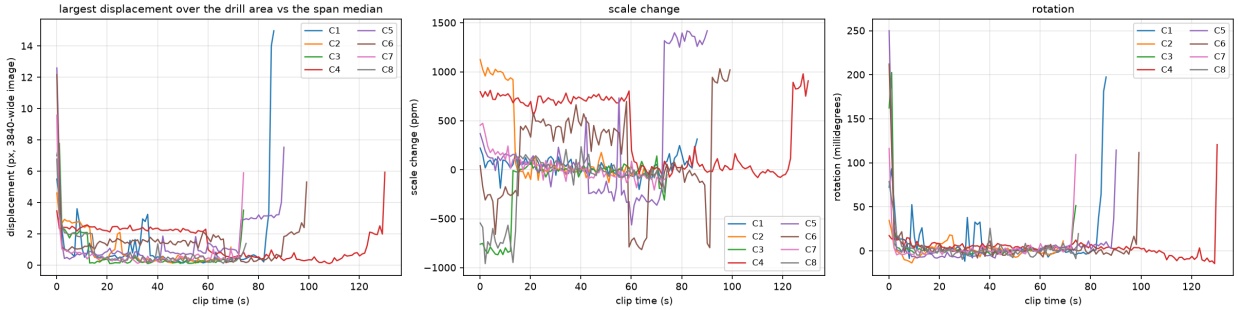


Figure 9: A4. Camera motion from keyframe registration: largest displacement over the drill area (px, 3840-wide image), scale and rotation, relative to the median over the analysed span. Jumps at start and end come from pressing the record button.

7.5 A5: keypoint bias

A constant keypoint offset cancels in $\Delta \boldsymbol{p}$. A pose- or heading-dependent offset $\boldsymbol{b}(\psi)$ adds $\boldsymbol{b}(\psi_b) - \boldsymbol{b}(\psi_a)$, zero for same-direction pairs and largest for opposite ones, depending on the headings, not the time between them.

Experiment. Splitting recorded pairs by the angle between running directions (Figure 10), same-direction pairs (within 30°), immune to heading bias, still grow: Xsens footfalls from 18 px to 36 px and Catapult from 29 px to 55 px, from 5 s to 10 s to beyond 25 s (no 1 s to 5 s pairs exist, since a quick revisit needs a turn). Their growth from 5 s to 10 s, 14.8 $[-8.3, 45.2]$ and 17.2 $[-7.7, 41.0]$ px, matches all pairs (15.6 and 21.1) but is less precise; opposite-direction pairs grow by 20.4 $[6.1, 34.0]$ and 20.2 $[1.1, 47.5]$ px.

Direction mix of the bins. The share of same-direction pairs varies with time apart (Table 6): none at 1 s to 5 s, about half at 10 s to 15 s, roughly one lap, when the athlete passes the same place in the same direction. Any error that cancels for same-direction pairs and adds for opposite ones, such as heading bias, a sync lag (A1), the stride phase over a bumped pitch (A3) or an image-only timing error (A9), therefore produces a non-monotonic level with a dip at 10 s to 15 s, not the steady growth observed; and the same-direction pairs, free of such errors, still grow.

Table 6: Share of same-direction pairs (directions within 30°) by time apart.

	1–5 s	5–10 s	10–15 s	15–25 s	25–60 s
Xsens footfalls	0 %	10 %	50 %	16 %	28 %
Catapult neck	0 %	19 %	45 %	19 %	22 %

Simulation. A simulated bias of 10 or 20 px along the direction of motion raises the level with a dip at 10 s to 15 s, as the direction mix predicts, but no growth: $G_m = -1.2$ $[-3.5, 1.5]$ and -3.0

$[-6.3, 0.2]$ px (footfalls). A 20 px bias also exceeds what the data allow, since the recorded 1 s to 5 s misfit is only 14 px. Bias of plausible size does not reproduce the growth; however, a large bias can mask a small drift ($G_m = 2.0$ $[-3.9, 8.1]$ px for 20 px plus a 2 cm/$\sqrt{\text{s}}$ random walk), so the test is conservative.

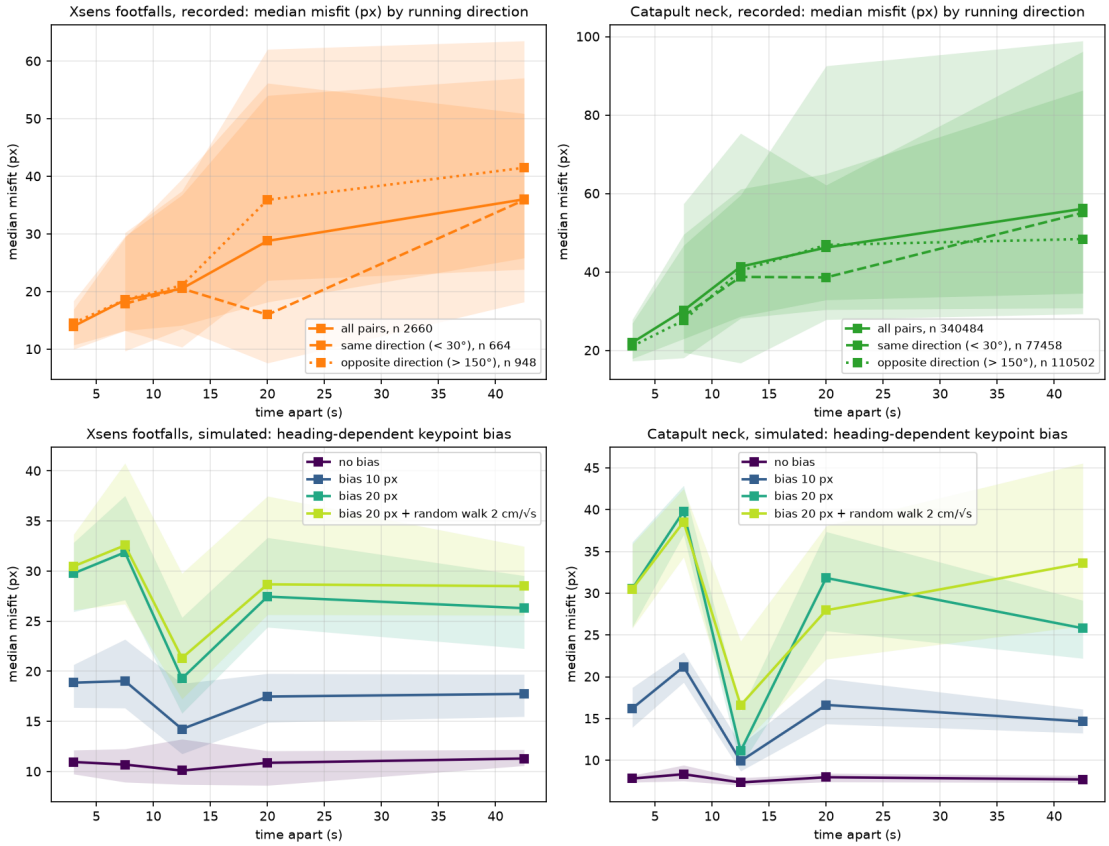


Figure 10: A5. Top: recorded pairs split by running direction. Bottom: simulated keypoint bias along the direction of motion.

7.6 A6: keypoint noise

By (8), independent noise σ adds about $2\sigma^2$ to $\operatorname{Var}|\Delta \boldsymbol{p}|$ and nothing to its covariance with $|\mathbf{J}\Delta\tilde{\boldsymbol{x}}|$, so $r' = r\sqrt{\operatorname{Var} O/(\operatorname{Var} O + 2\sigma^2)}$. Adding 2, 5 and 10 px of Gaussian noise to the recorded keypoints follows this to within 0.01–0.03 (Figure 11; with 10 px, footfall r at 1 s to 5 s is 0.94 against 0.93 predicted). Noise lowers all bins together and raises the misfit level; it dilutes the growth slightly but never creates it ($G_m = 15.1$ $[5.1, 27.1]$ px for footfalls and 26.9 $[17.4, 39.8]$ px for Catapult with 10 px added). The real keypoint noise, 1.6 px (neck) and 4.7 px (ankles) per axis on short displacements, is far below the level where the test loses power.

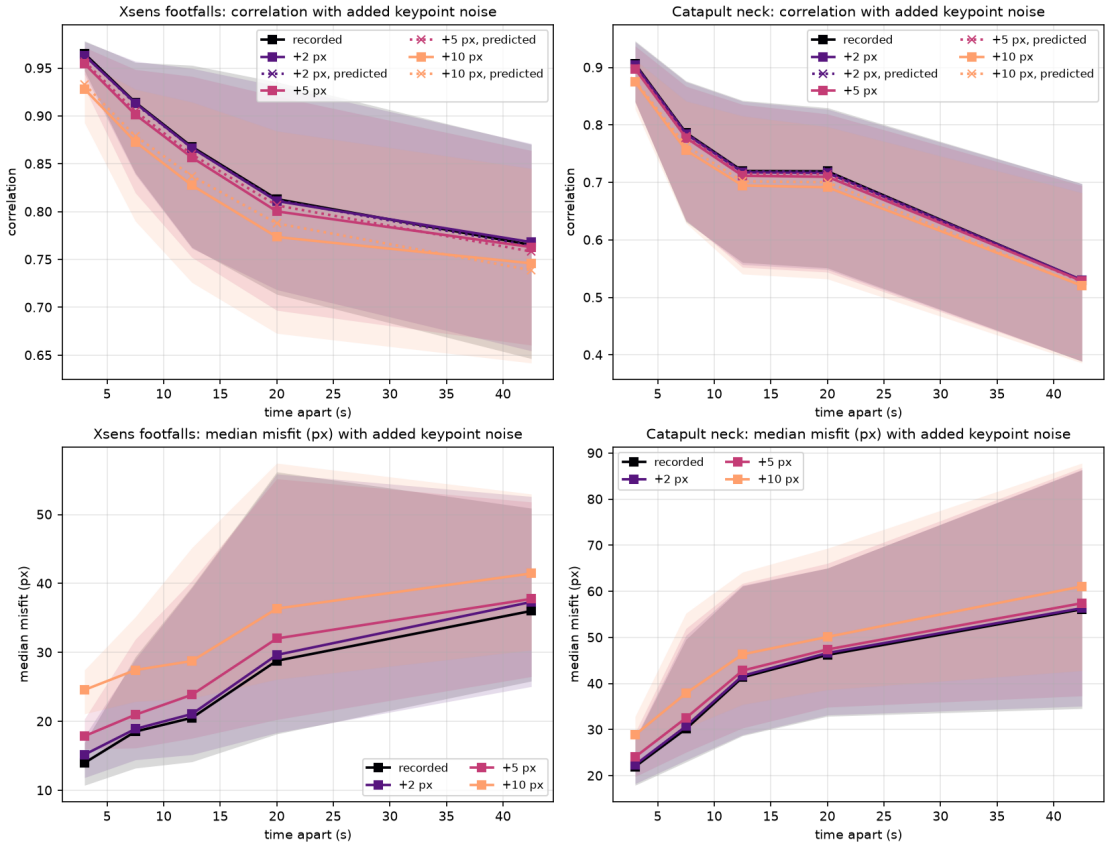


Figure 11: A6. Recorded data with added keypoint noise. Top: correlation, measured (squares) and predicted (crosses). Bottom: median misfit.

7.7 A7: footfalls, neck or other body points

For Xsens every body point agrees (Figure 12): footfalls, ankles, knees, mid-hip, neck and shoulders all go from 12 px to 17 px to 36 px to 39 px, with G_m of 18 px to 21 px and all intervals above zero. For Catapult, a single point on the upper back, the neck gives the strongest result and the shoulders the same growth ($G_m = 25.3$ [14.3, 38.9] px); the mid-hip, which moves differently from the vest, starts higher and grows less (14.1 [0.4, 33.7] px). Points close to the sensor give the cleanest measurement; footfalls add a well-defined instant and position and are easy to annotate by hand.

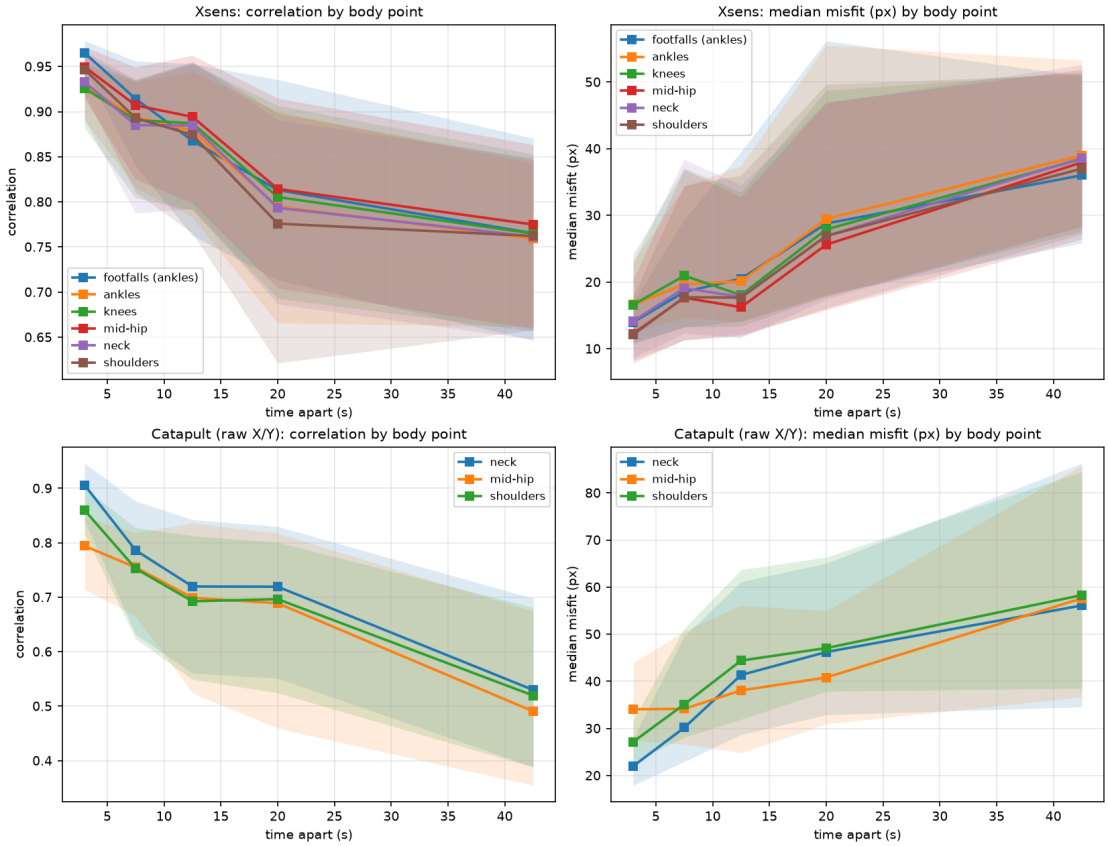


Figure 12: A7. Correlation and median misfit by body point. Top: Xsens. Bottom: Catapult against different image keypoints.

7.8 A8: pixel–metre (px–cm) against pixel–pixel (px–px) correlation

Correlating $|\Delta \boldsymbol{p}|$ directly with metric distance $|\Delta \boldsymbol{x}|$ is simpler, but $|\Delta \boldsymbol{p}| = |\mathbf{J}\Delta \boldsymbol{x}|$ ranges from about 16 to 148 px/m with direction and range, so without selection the scatter is a fan of lines (Figure 13), the correlation is low and shows no trend (Xsens 0.47, 0.57, 0.66, 0.58, 0.43): direction hides the drift. Keeping nearly horizontal image displacements (vertical share ≤ 0.15) that also lie across the ray in the sensor (along-ray share ≤ 0.3, from the sensor's own 2×2 map) restores one line, and the px–cm correlation then falls with time apart (Xsens 0.76 to 0.58, Catapult 0.79 to 0.43), with growing shares beyond its margins. The simpler test thus agrees, but only after discarding about 80 % of the pairs, at a lower level and relying on an estimated ray direction; the J-normalised test uses every pair and shows the trend more cleanly (Xsens 0.93 to 0.76, Catapult 0.91 to 0.53).

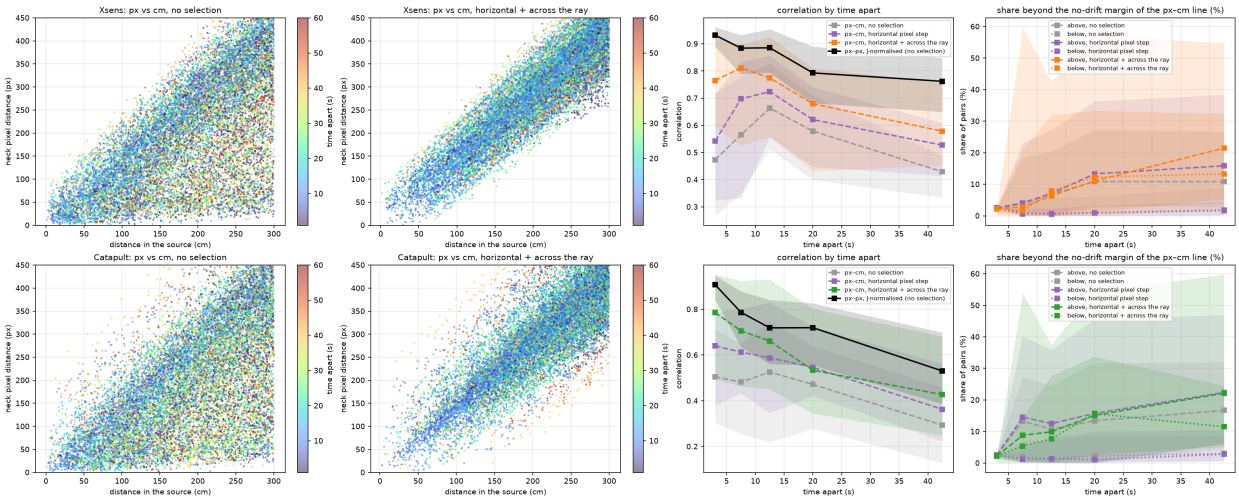


Figure 13: A8. Left: image against metric distance of the neck, without selection and with the horizontal and across-ray selection, coloured by time apart. Right: correlation by time apart for px–cm (three selections) and J-normalised px–px, and shares beyond the px–cm margins.

7.9 A9: errors in detecting the footfalls

A *common* error detects the footfall too early or late but reads image and sensor at the same instant: they stay consistent, revisits just match less closely in space. An *image-only* error detects the image event δ away from the sensor event, adding $\mathbf{J}\boldsymbol{v}_{\text{foot}}\delta$ per event, which depends on stride phase, not time apart.

The Xsens footfall instants were moved (Figure 14).

- **Common:** shifts of -6, -3, $+3$, $+6$ frames (up to ± 100 ms) or jitter of 2 or 4 frames s.d. Early detection raises the level (foot still in flight; 14 to 23 px at -6 frames), late detection hardly matters (foot planted); growth remains (G_m 14.9 to 21.6 px, all intervals above zero).
- **Image-only:** ± 3 frames or 2 frames jitter change the level by a few pixels and keep the growth (G_m 19.5 to 22.3 px). -5 frames (83 ms, the independent foot model's offset, A10) raises the 1 s to 5 s level from 14 to 29 px and hides most of the growth ($G_m = 4.1$ $[-7.1, 20.5]$ px).

The image-only term $\mathbf{J}\boldsymbol{v}_{\text{foot}}\delta$ cancels for same-direction footfalls and adds up for opposite ones; since the 1 s to 5 s bin is all opposite-direction pairs (Table 6), timing errors inflate the reference bin most. They can mask growth but not create it.

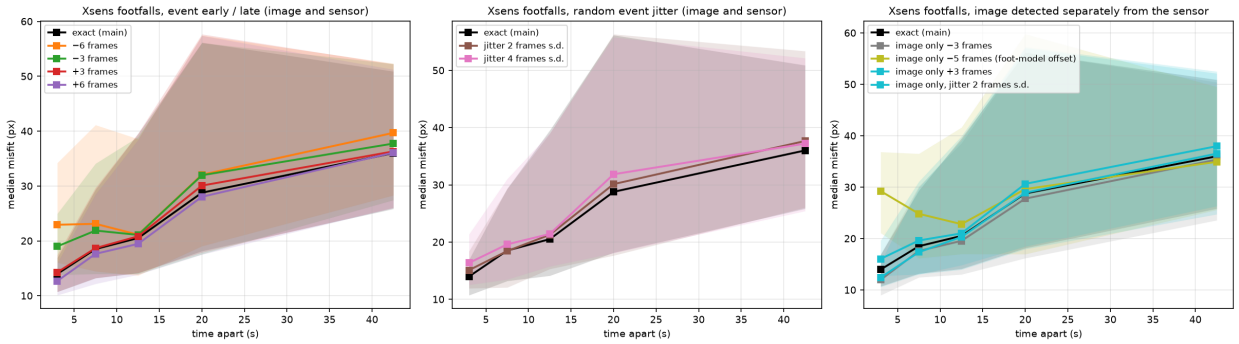


Figure 14: A9. Median misfit of the Xsens footfalls with moved instants: early or late for both image and sensor, random jitter for both, and moved in the image only.

7.10 A10: footfalls and keypoints from an independent model

The body-pose keypoints and the Xsens contacts could share a bias with the test. The second FitWise AI image model predicts, from its own crop, heel and toe keypoints and a ground-contact

flag per foot. Its contact onsets (flag rising and held for 3 frames) and mid-foot point (mean of heel and toe, against the mean of the Xsens foot and toe segments) replace the Xsens footfalls and the body-pose ankles, one at a time and together (Figure 15). Its onsets come 83 ms before Xsens's (median; robust s.d. 25 ms; 745 footfalls), a difference of definition (a foot about to land versus a foot low and slow); used for both image and sensor, this is a common error in the sense of A9.

Every combination grows: Xsens footfalls with body-pose ankles (main) $G_m = 20.1$ [9.3, 35.5] px; with model mid-foot 17.5 [0.4, 29.3]; model footfalls with body-pose ankles 18.2 [4.5, 37.8]; model footfalls with model mid-foot, using nothing from the body-pose model or the Xsens contacts, 15.4 [2.9, 23.6] px. The noisier mid-foot keypoints raise the level (29 px to 32 px at 1 s to 5 s). For Catapult, model footfalls and mid-foot against the vest grow by 22.1 [9.3, 38.7] px from a high level (52 px, the vest is not the foot). The growth is therefore unlikely to be an artefact of the body-pose model or of footfall detection.

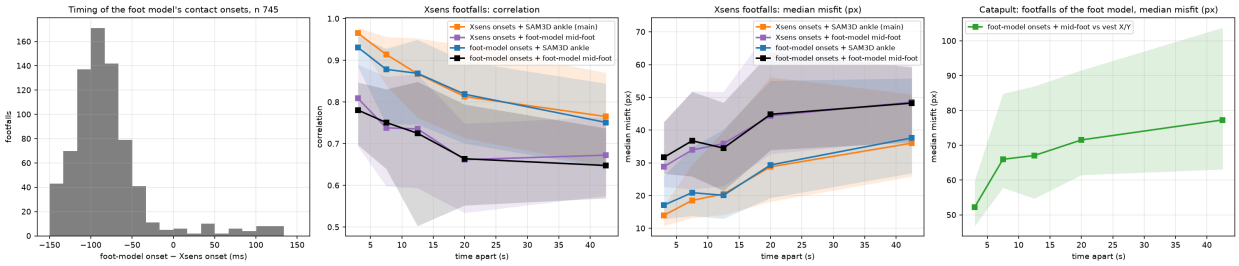


Figure 15: A10. Left: timing of the independent foot model's contact onsets against Xsens. Middle: Xsens footfall test with events and image points from either model. Right: Catapult with the foot model's footfalls and mid-foot keypoints.

7.11 Summary of the ablations

Question	Finding
A1 sync lag	A constant lag raises every bin, adds no growth; rate errors up to 1000 ppm have no effect.
A2 video PTS	No dropped or repeated frames; rate within −41 ppm to 15 ppm of Xsens; worst-case bound about 21 px over 60 s.
A3 non-planarity, height, lens	Bumps of ±20 cm and $k_1 = \pm 0.15$ add no growth; neck height changes the level only; any form of $\mathbf{J}$ gives the same result.
A4 static camera	Within 1 px to 3 px and 0.15 % scale during the analysed span.
A5 keypoint bias	Same-direction pairs still grow; simulated heading bias adds none.
A6 keypoint noise	Lowers r in every bin as predicted; adds no growth.
A7 body point	All Xsens points agree; Catapult is clearest with keypoints near the vest.
A8 px–cm	Same conclusion after direction selections, with fewer pairs and lower correlation.
A9 footfall detection	Timing errors change the level, not the growth; an image-only offset can mask growth, never create it.
A10 independent model	Footfalls and keypoints of a second image model show the same growth.

References

[1] L. S. Luteberget and M. Gilgien. Validation methods for global and local positioning-based athlete monitoring systems in team sports: a scoping review. *BMJ Open Sport & Exercise Medicine*, 6(1):e000794, 2020. `https://doi.org/10.1136/bmjsem-2020-000794`.

[2] S. Ellens, C. Moran and M. C. Varley. Concurrent validity and between-device reliability of the Catapult Vector S8 GNSS device. *PLoS ONE*, 20(10):e0333792, 2025. https://doi.org/10.1371/journal.pone.0333792.

[3] M. P. Mavor, G. B. Ross, A. L. Clouthier, T. Karakolis and R. B. Graham. Validation of an IMU suit for military-based tasks. *Sensors*, 20(15):4280, 2020. https://doi.org/10.3390/s20154280.

[4] M. W. Lenhoff, S. Zanini, G. Chingas, S. Patel, R. DiGiacomo and H. J. Hillstrom. Intra and inter-system reliability of a markerless motion capture system in different configurations during functional tasks: implications for utility. *Medical Engineering & Physics*, 142:104375, 2025. https://doi.org/10.1016/j.medengphy.2025.104375.

[5] M. Boltes, J. Adrian and A.-K. Raytarowski. A hybrid tracking system of full-body motion inside crowds. *Sensors*, 21(6):2108, 2021. https://doi.org/10.3390/s21062108.

[6] M. Schepers, M. Giuberti and G. Bellusci. Xsens MVN: Consistent tracking of human motion using inertial sensing. White paper, Xsens Technologies B.V., Enschede, 2018.

[7] Xsens Technologies B.V. *Xsens Link User Manual*, revision B. Enschede, 2026.

[8] E. Chebel, Y. Corky, O. Kassem et al. Quantifying the impact of integrating the HTC Vive tracker with the Awinda motion capture system in minimizing positional drift. In *2024 Medical Technologies Congress (TIPTEKNO)*, IEEE, 2024. https://doi.org/10.1109/TIPTEKNO63488.2024.10755488.

[9] R. G. J. Damgrave and D. Lutters. The drift of the Xsens Moven motion capturing suit during common movements in a working environment. In *Proceedings of the 19th CIRP Design Conference*, pages 338–342, Cranfield, 2009.

[10] R. van den Tillaar, R. Nagahara, S. Gleadhill and P. Jiménez-Reyes. Step-to-step kinematic validation between an inertial measurement unit (IMU) 3D system, a combined laser+IMU system and force plates during a 50 m sprint in a cohort of sprinters. *Sensors*, 21:6560, 2021. https://doi.org/10.3390/s21196560.

[11] M. K. Fleron, N. C. H. Ubbesen, F. Battistella, D. L. Dejtiar and A. S. C. Oliveira. On the use of inertial sensors to determine trunk displacement during walking: a validation study. In *Proceedings of the 35th Conference of the International Society of Biomechanics in Sports*, pages 496–499, 2017.

[12] O. J. Woodman. An introduction to inertial navigation. Technical Report UCAM-CL-TR-696, University of Cambridge, Computer Laboratory, 2007.

[13] K. Lebel, P. Boissy, M. Hamel and C. Duval. Inertial measures of motion for clinical biomechanics: comparative assessment of accuracy under controlled conditions – changes in accuracy over time. *PLoS ONE*, 10(3):e0118361, 2015.

[14] X. Robert-Lachaine, H. Mecheri, C. Larue and A. Plamondon. Validation of inertial measurement units with an optoelectronic system for whole-body motion analysis. *Medical & Biological Engineering & Computing*, 55:609–619, 2017.

[15] Ø. Gløersen, J. Kocbach and M. Gilgien. Tracking performance in endurance racing sports: evaluation of the accuracy offered by three commercial GNSS receivers aimed at the sports market. *Frontiers in Physiology*, 9:1425, 2018. https://doi.org/10.3389/fphys.2018.01425.

[16] P. A. H. Jølstad, R. C. Reid, J. G. O. Gjevestad and M. Gilgien. Validity of the AdMos, Advanced Sport Instruments, GNSS sensor for use in alpine skiing. *Remote Sensing*, 14(1):22, 2022. https://doi.org/10.3390/rs14010022.

[17] D. Linke, D. Link and M. Lames. Validation of electronic performance and tracking systems EPTS under field conditions. *PLoS ONE*, 13(7):e0199519, 2018. https://doi.org/10.1371/journal.pone.0199519.

[18] A. Bastida-Castillo, C. D. Gómez-Carmona, E. De La Cruz Sánchez and J. Pino-Ortega. Comparing accuracy between global positioning systems and ultra-wideband-based position tracking systems used for tactical analyses in soccer. *European Journal of Sport Science*, 19(9):1157–1165, 2019. https://doi.org/10.1080/17461391.2019.1584248.

[19] M. Corsie and P. A. Swinton. Reliability of spatial-temporal metrics used to assess collective behaviours in football: an in-silico experiment. *Science and Medicine in Football*, 7(3):297–305, 2023.

[20] T. von Marcard, R. Henschel, M. J. Black, B. Rosenhahn and G. Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, 2018. https://doi.org/10.1007/978-3-030-01249-6_37.

[21] C. Malleson, M. Volino, A. Gilbert, M. Trumble, J. Collomosse and A. Hilton. Real-time full-body motion capture from video and IMUs. In *International Conference on 3D Vision (3DV)*, 2017.

[22] B. Fasel, M. Gilgien et al. A new training assessment method for alpine ski racing: estimating center of mass trajectory by fusing inertial sensors with periodically available position anchor points. *Frontiers in Physiology*, 2018.

A What drift does to a pair

The sensor is modelled as the true motion plus a slowly varying shift $\boldsymbol{d}(t)$, as described for sport receivers [16], a slowly varying yaw $\theta(t)$ about a point $\boldsymbol{c}$, as needed to merge an Xsens trajectory with a camera's [5], and white noise $\boldsymbol{\eta}$:

$$\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t) + \boldsymbol{d}(t) + \theta(t)\,\mathbf{R}_\perp(\boldsymbol{x}(t) - \boldsymbol{c}) + \boldsymbol{\eta}(t), \qquad \mathbf{R}_\perp = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{9}$$

Constant parts are absorbed into $\mathbf{J}$, so $\boldsymbol{d}$ and θ are relative to their clip means. With keypoint noise $\boldsymbol{n}$, the misfit of a pair is

$$\begin{aligned} \boldsymbol{m}_{ab} = \Delta\boldsymbol{p} - \mathbf{J}\Delta\tilde{\boldsymbol{x}} = {} & \underbrace{\epsilon_{\text{model}}}_{O(|\Delta x|^3)} + (\boldsymbol{n}_b - \boldsymbol{n}_a) - \mathbf{J}(\boldsymbol{\eta}_b - \boldsymbol{\eta}_a) \\ & - \mathbf{J}\big[\boldsymbol{d}(t_b) - \boldsymbol{d}(t_a)\big] - \mathbf{J}\big[\theta_b \mathbf{R}_\perp(\boldsymbol{x}_b - \boldsymbol{c}) - \theta_a \mathbf{R}_\perp(\boldsymbol{x}_a - \boldsymbol{c})\big]. \end{aligned} \tag{10}$$

For a revisit ($\boldsymbol{x}_b \approx \boldsymbol{x}_a$) the yaw term is $\mathbf{J}(\theta_b - \theta_a)\mathbf{R}_\perp(\boldsymbol{x}_a - \boldsymbol{c})$: a shift scaled by a lever arm of several to tens of metres. On a short displacement the same yaw only rotates it by θ, a few per cent, which is what the windowed fits of $\mathbf{J}$ pick up in A3.

The noise and model terms depend on displacement size and image position, not on Δt; the drift terms depend on Δt through $\mathbf{S}$ (6). Without drift the misfit distribution is the same at every Δt (flat curves); a random walk of diffusion q gives $\operatorname{tr}\mathbf{S} = 2q\Delta t$ and growth like $\sqrt{\Delta t}$; a constant rate $\boldsymbol{v}_d$ gives $\mathbf{S} = \boldsymbol{v}_d\boldsymbol{v}_d^\top\Delta t^2$ and linear growth; a bounded drift (e.g. heading-dependent) gives $\mathbf{S} \to 2\operatorname{Cov}\boldsymbol{d}$, early growth then a plateau.

B Validation on synthetic data

The full test was rerun on simulated data with known truth: the recorded trajectories (Xsens joints; Catapult in metres), so motion, timing and revisits are those of this recording; a generic pinhole camera 20 m from the scene, 2.5 m high, $f = 2740$ px, reproducing only the real foreshortening; Gaussian keypoint noise at the level measured on real short displacements (1.6 px per axis for the neck, 4.7 px for the ankles); and a sensor equal to the truth plus injected drift. Without drift the curves are flat ($G_m = 0.4$ px $[-1.0, 1.8]$ for footfalls, -0.1 px $[-0.4, 0.4]$ for the neck; $r \approx 0.98$–0.99 in every bin). Every injected drift produces growth; a 1 cm/$\sqrt{\text{s}}$ random walk (about 8 cm per minute) is already detected, and 1 cm/s gives $G_m \approx 20$ px for both, about the size observed (Table 7, Figure 16).

Table 7: Synthetic validation: growth of the median misfit G_m (px), with 95 % intervals.

Injected drift	Xsens footfalls	Catapult neck
none	0.4 [−1.0, 1.8]	−0.1 [−0.4, 0.4]
random walk 1 cm/$\sqrt{\text{s}}$	1.8 [0.2, 3.5]	5.8 [1.5, 11.6]
random walk 2 cm/$\sqrt{\text{s}}$	6.3 [3.5, 9.7]	12.8 [3.1, 23.6]
random walk 4 cm/$\sqrt{\text{s}}$	19.1 [10.4, 28.6]	26.2 [6.6, 48.9]
constant shift 1 cm/s	19.8 [8.7, 29.7]	19.8 [12.9, 31.4]
yaw 0.05 °/s	5.6 [2.9, 9.1]	8.1 [6.7, 11.2]

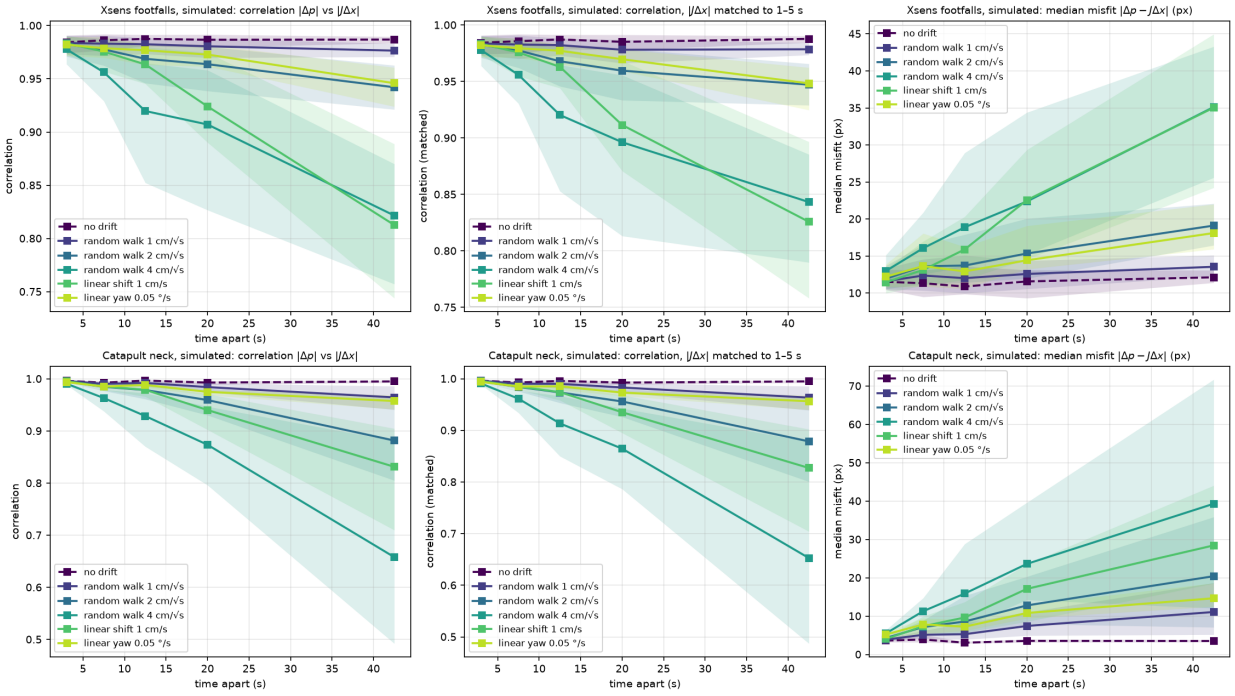


Figure 16: Synthetic validation: correlation (plain and matched) and median misfit by time apart for known injected drifts, with 95 % block-bootstrap bands.

C Catapult coordinates

Geometry. If X/Y are two components of an Earth-centred frame with the third dropped, a ground point at distance d from the third axis lies at angle α from the X/Y plane, $\sin\alpha = \sqrt{1-(d/R)^2}$. A small displacement across the radial direction $\boldsymbol{e}_r = \boldsymbol{p}_0/|\boldsymbol{p}_0|$ from the axis to the

point (along $\boldsymbol{e}_t \perp \boldsymbol{e}_r$) stays in the plane at full length; one along $\boldsymbol{e}_r$ tilts out of it and keeps a factor $\sqrt{1-(d/R)^2}$. Hence

$$\boldsymbol{x}_{\mathrm{m}} = \begin{pmatrix} \boldsymbol{e}_r^{\top}/\sqrt{1-(d/R)^2} \\ \pm\boldsymbol{e}_t^{\top} \end{pmatrix} (\tilde{\boldsymbol{x}} - \boldsymbol{p}_0), \qquad d = |\boldsymbol{p}_0|, \tag{11}$$

with one reference point $\boldsymbol{p}_0$ for all clips (the mean position over the six clips) and the sign of the second row setting the handedness; over a pitch the map is linear to about 10^{-5}. With $\boldsymbol{p}_0 = (-4\,846\,881.0,\ 1\,602\,373.7)$, $d = 5\,104\,886$ and the WGS84 radius at the venue, $R = 6\,371\,222$ m, the parameter-free prediction is a stretch of 1.671 along $\operatorname{atan2}(p_{0,y}, p_{0,x}) = 161.7°$ and 1.000 across. The frame is not standard WGS84 ECEF (about 1.75 along a different direction at the venue; implied latitude 36.75°) but a rotated Earth-centred frame of unknown definition. The drift test does not depend on it, since **J** absorbs any constant linear map.

Checks (Figure 17).

- **Speed:** the per-clip symmetric 2×2 map from raw velocities to the Velocity channel has scales of 1.666–1.707 and 0.977–0.999 m per unit along 160.2–161.8° over 0.2 s windows; longer windows inflate both, since a chord is shorter than the curved path (1.79–1.84 and 1.06–1.09 at 2 s).
- **Isotropy:** Velocity over the speed of converted positions is 0.99–1.01 in every running direction.
- **Against Xsens** (C1, C2, C5; raw sync, no correction): one rigid transform per clip maps the converted vest onto the Xsens mid-shoulders with median errors of 34, 40 and 41 cm; a similarity finds scales of 0.987, 0.956 and 1.001; the unmirrored frame fails (3–4 m), so the frame is mirrored relative to Xsens.

The converted positions are thus metres, consistent with an independent sensor without rescaling.

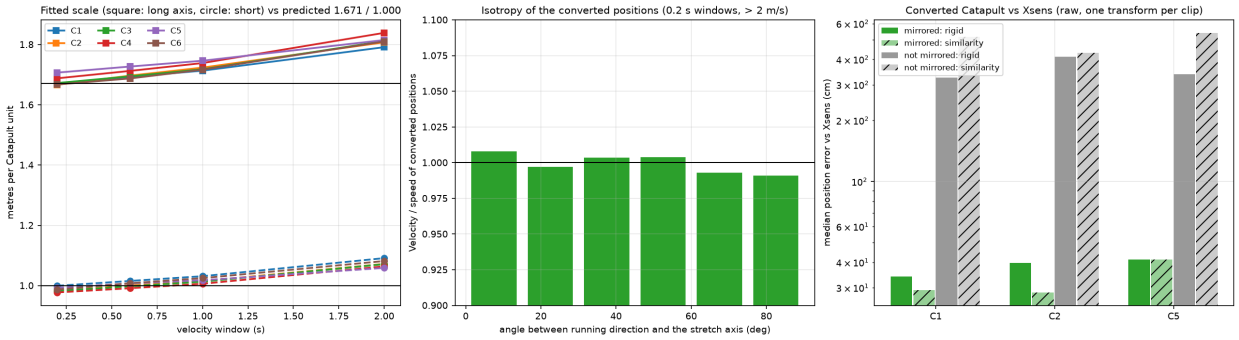


Figure 17: Catapult coordinates. Left: scales fitted from the Velocity channel against window length, with the parameter-free prediction (black lines). Middle: Velocity over converted speed by running direction. Right: converted Catapult against Xsens, median error after a rigid transform or a similarity, both handednesses (log scale).

D The locality limit

On simulated data without noise or drift, the misfit is the model error of the quadratic **J** at the midpoint; it grows with pair size as the third-order term predicts (Figure 18, left): for Xsens footfalls 4 px at 60 px, 13 px at 200 px and 29 px at 400 px, for the Catapult neck 2 px at 200 px and 3 px at 400 px. Ankles are harder to map with one **J** because its fit samples span a range of foot heights. The error depends on pair size, not time apart, and the simulated test containing it is flat. 100 px keeps a third of the pairs of 200 px; 300 px doubles the model error. On recorded data the growth barely depends on the limit (middle and right): G_m is 21.5, 18.7, 20.1 and 22.1 px (Xsens footfalls) and 26.3, 28.2, 29.3 and 30.5 px (Catapult neck) at 100, 150, 200 and 300 px. G_r

is larger at small limits, where σ_P is smaller (8), and positive throughout. The 200 px value comes from earlier exploratory analysis; the conclusions do not depend on it.

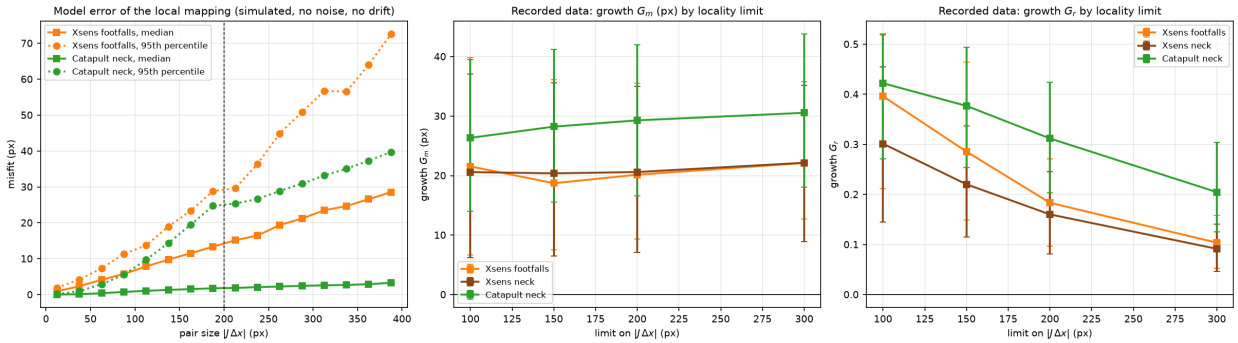


Figure 18: Left: model error of the local mapping by pair size (simulated, no noise, no drift). Middle and right: growth statistics of the recorded data for different locality limits.

E Reproduction

All numbers and figures come from `scripts/`: `make data` extracts the inputs, `make results` runs the analyses, `make pdf` builds this document; results are in `data/results/*.json`.