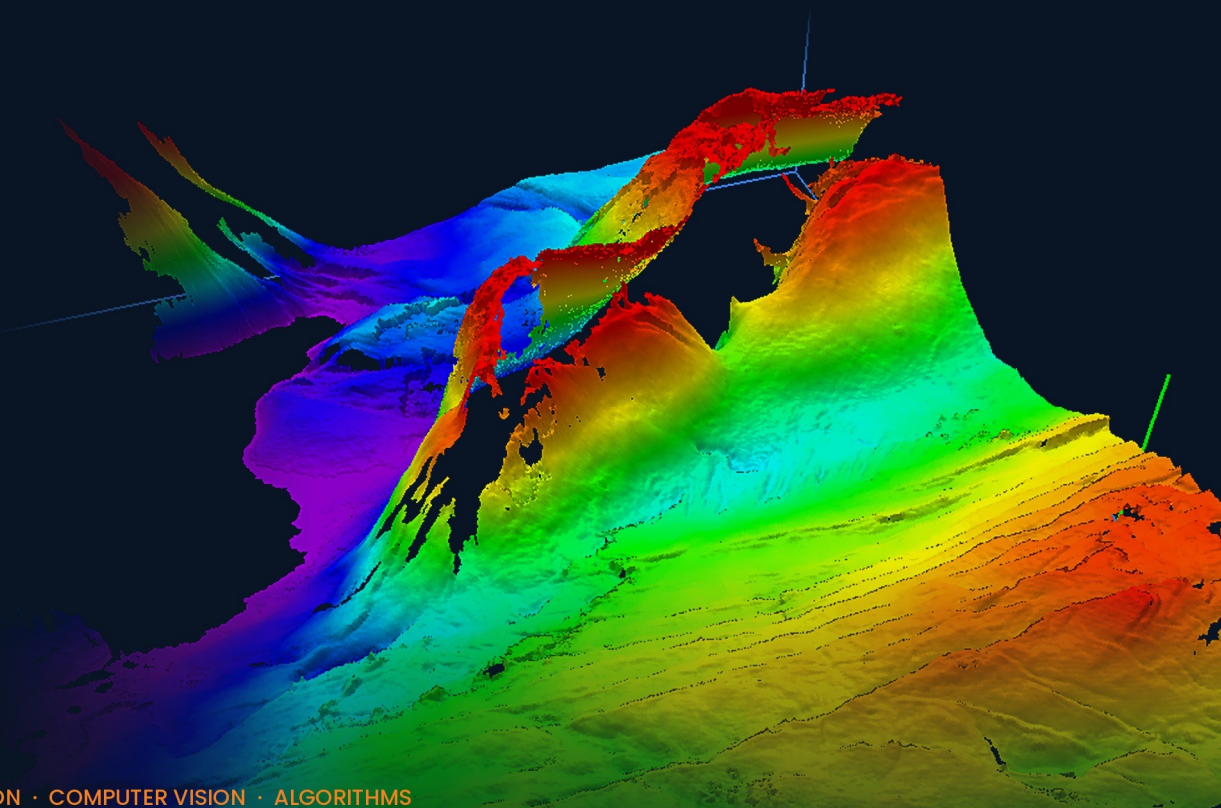


SEISMIC INTERPRETATION · COMPUTER VISION · ALGORITHMS

99.9 % locally right — and why *that is not enough*

The silent cycle skip in seeded 3-D seismic horizon tracking: comparison reference, growth order and access locality

99.9 %
of cells at local extrema — on a suspect surface

511 / 276
ms: surface span versus the span of the picks

600 → 3 s
the cost of fixing the order, recovered by a bucket queue

91.5 → 38 %
coverage when compared with the seed, not the parent

Edmilson Delfim Praia
COPIA Seis

Luanda · October 2026

Data: public F3 survey, North Sea (dGB Earth Sciences)

RESEARCH ARTICLE Seismic interpretation · Computer vision · Algorithms

99.9 % Locally Right — and Why That Is Not Enough

The silent cycle skip in seeded 3-D seismic horizon tracking: comparison reference, growth order and access locality

Edmilson Delfim Praia
COPIA Seis, Luanda, Angola

Abstract — Seeded automatic horizon tracking fails chiefly through cycle skipping: the resulting surface is smooth and plausible, and every cell satisfies the local acceptance criteria. On the public F3 survey (619,101 traces), a surface with **99.9 %** of its cells at local extrema of the correct polarity spanned **511 ms**, against **276 ms** for the 43 seed picks, all on one inline. The surface is suspect, not refuted — a horizon dipping away from that line may legitimately span more time — but the central point does not depend on it: **local tests cannot tell the two hypotheses apart**. We model drift as a systematic term proportional to path length plus a random walk on the growth tree, and verify four consequences in code and data. (i) Comparing each candidate with the **seed** trace, rather than the parent trace, removes the path from the acceptance decision: coverage fell from **91.5 % to 38.0 %** and the time span from 511 to 396 ms. (ii) A tolerance band around the trend of the picks cannot reject one-loop skips when its half-width exceeds half a period (≈ 16 ms on F3); at ±60 ms it only reduced coverage, and it is off by default. (iii) The correlation window and the growth order control anomalous steps (> 40 ms): from 15 to 0 when the window grows from 4 to 12 samples, and from 40 to 0 when amplitude-first priority is replaced by proximity-first. (iv) Correcting the growth order took full-horizon tracking from 5 s to over 600 s; a bucket queue aligned with the bricks of the storage format brought it to **3.0 s** with the same arithmetic. We identify the missing experiment — comparison with published, manually interpreted F3 horizons — and propose, without results, a machine-learning integration.

Keywords: *seismic interpretation; horizon tracking; cycle skipping; structure tensor; region growing; locality of reference; weak supervision.*

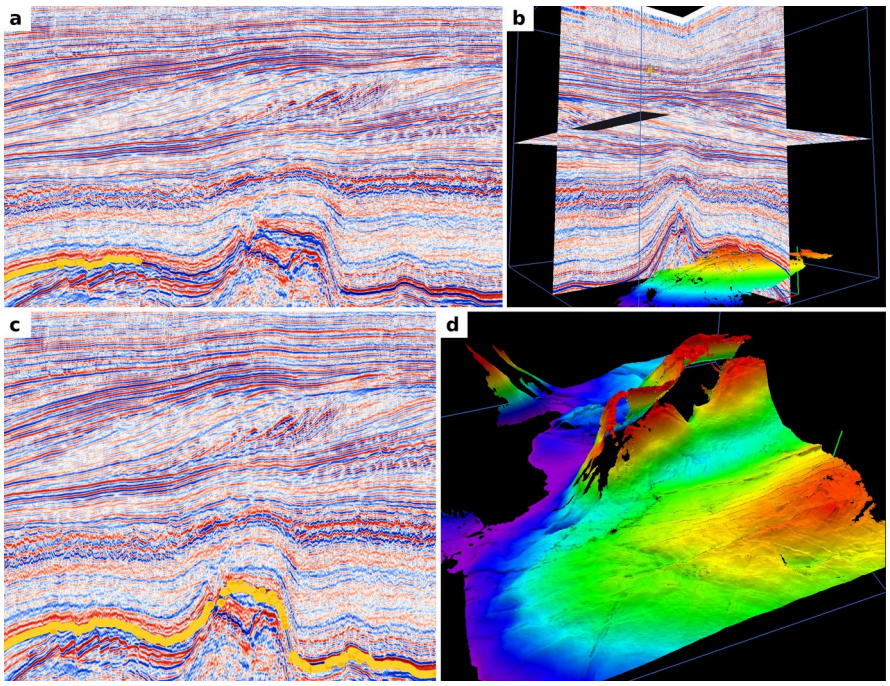


Figure 1. Seeded tracking in volume mode on F3, in the implementation described here. (a, b) Three seeds on a single section: small coverage, confined to the picked flank. (c, d) Twenty-two seeds distributed over several sections, including the crest and the steep flank of the central structure: the application reports **75.9 %** coverage. Colour scale in (b, d): two-way time of the horizon. Coverage is as reported by the application in an interactive session, not a controlled experiment. The holes in (d) are deliberate rejections (faults, low similarity, round-trip failures). The near-vertical walls by the central structure may be real fault throw or the blades of Section 4.6; telling them apart requires reference horizons (Section 7.2).

1 INTRODUCTION

A three-dimensional seismic survey is a sampled function $a(i, c, s)$ of two map coordinates — inline i and crossline c — and one vertical coordinate s in two-way travel time. Interpreting it starts with identifying **horizons**: surfaces $z(i, c)$ that represent continuous chronostratigraphic boundaries. On the F3 dataset used throughout this work this means $651 \times 951 = 619{,}101$ map positions over 462 samples, or 2.86×10^8 samples. Manual picking is infeasible; standard practice is **seeded tracking**, in which the interpreter picks a few points and an algorithm propagates the surface (Hoyes and Cheret, 2011).

The difficulty is not finding a local extremum. A seismic trace is quasi-periodic, and the loop next to the correct one closely resembles it. Following the wrong loop — **cycle skipping** — has three properties that organise this paper: it is *locally consistent* (the wrong position is a genuine extremum of the correct polarity, well correlated); it is *cumulative* (a slip of one loop every two hundred traces is imperceptible at each step); and it is *observable only globally*. A methodological thesis follows and runs through the paper: **visual inspection is not verification**. A wrong surface is as smooth as a correct one.

Our contributions are as follows:

(i) **Quantification of the silent failure mode**, and a drift model — a systematic term plus a random walk on the growth tree — that explains why no local criterion can detect it (Sections 3 and 6.1).

(ii) **One non-local constraint that works and one that does not, both explained by the model**: comparing with the seed trace removes the path from the acceptance decision; trend tolerance cannot reject one-loop skips above half a period. The round-trip check detects asymmetric skips (Sections 3.2, 4.4 and 4.5).

(iii) **Growth order as an algorithm parameter**, with a measured effect on geometry (Sections 4.6 and 6.2).

(iv) **The cost of fixing the order, and its recovery**: the correct order multiplied time by more than 100; a bucket queue aligned with the brick layout brought it below the original (Sections 4.8 and 6.4).

(v) Validated engineering contributions — Scharr kernels with inner scale, a map-view fault barrier, a convex and dip-corrected cost for guided picking — and a **proposal**, explicitly without results, for machine-learning integration (Section 8).

(vi) **An explicit statement of what remains to be shown**, and of the experiment that settles it with public data (Section 7.2).

2 RELATED WORK

Local tracking. Following events by normalised correlation between adjacent traces, organised as region growing from seeds (Adams and Bischof, 1994), is the basis of commercial trackers. The open reference implementation was read as literature and not reused, owing to licence incompatibility. The literature describes acceptance criteria in detail but rarely treats visit *order* as an experimental variable, which is the subject of Section 4.6.

Global methods. An alternative line avoids step-by-step propagation: flattening without picking solves a least-squares problem over the dip field (Lomask *et al.*, 2006), relative geologic time volumes assign an age to every sample (Stark, 2004), and interpreter constraints can be imposed on the global field (Wu and Hale, 2015). These methods do not accumulate drift in the same way, but they give up the central interactive operation — "this event, from here" — and depend on the quality of the dip field.

Orientation. The structure tensor (Bigün and Granlund, 1987; Kass and Witkin, 1987) and plane-wave destruction filters (Fomel, 2002) are the standard tools for dip estimation. Using them as a predictor for the next step exposes sensitivities — bias of the derivative operator and aliasing at steep dips — that Section 4.3 quantifies.

Consistency checks. The round-trip check is the seismic analogue of the left–right check in stereo matching and of the forward–backward error in video object tracking (Kalal *et al.*, 2010). We have not found it described explicitly as an acceptance criterion in horizon tracking, but we have not systematically reviewed bidirectional and commercial trackers, whose documentation is often incomplete; the novelty claim should be read with that reservation.

Machine learning. Convolutional networks have focused on fault segmentation trained on synthetic data (Wu *et al.*, 2019) and on regression of relative geologic time (Geng *et al.*, 2020), and F3 is an established benchmark (Alaudah *et al.*, 2019). These approaches replace the tracker; Section 8 instead proposes informing it component by component.

3 FORMULATION

3.1 Problem

Let Ω be the set of map positions $x = (i, c)$ and $S = \{(x_k, z_k)\}$ the set of seeds. We seek z: $\Omega' \to \mathbb{R}$, with $\Omega' \subseteq \Omega$ as large as possible, such that **every determined position belongs to the same geological event**. Without this clause any local extremum will do; with it, the problem is one of identity, not detection. Vertical indices are fractional and all times are two-way times.

3.2 Drift as a random walk on the growth tree

A seeded tracker builds a forest: every accepted cell y has a parent x and inherits its predicted position from it. Let $\pi(x)$ be the path from the seed to x, of length $L(x)$, and $\Delta_x = z(x) - z^*(x)$ the error with respect to the true event z^*. We split the error of each step as $\varepsilon_e = b + \eta_e$: a systematic term b — the bias of the dip predictor — and a zero-mean term η_e (residual quantisation, noise). Assuming b approximately constant along the path and η_e with short-range correlation and long-run variance σ^2, we have $\Delta_x = b\,L(x) + \Sigma\eta_e$ and, for two neighbouring cells,

$$E\left(\Delta_x - \Delta_y\right)^2 = b^2\left(L_x - L_y\right)^2 + \sigma^2\,\ell\left(x, y\right) \qquad (1)$$

where $\ell(x, y)$ is the length of the symmetric difference between $\pi(x)$ and $\pi(y)$. With short-range correlation the variance still grows linearly with ℓ, σ^2 being the sum of the autocovariances; only a persistent component, such as b, changes the scaling. Expression (1) has three consequences. First, error grows with path length — linearly through the systematic term, as the square root through the random one — and no test applied to an isolated step observes that sum: **drift is locally invisible by construction**. Second, the discontinuity between neighbours depends on $L_x - L_y$ and on ℓ, properties of the *shape of the tree* — that is, of visit order — and not of the acceptance rule. Third, a constraint $|z(x) - T(x)| \le \tau$ around a trend T with error $e_T = T - z^*$ guarantees $|\Delta_x| \le \tau + |e_T(x)|$ and $|\Delta_x - \Delta_y| \le 2\tau + |e_T(x) - e_T(y)|$, whatever the order. **The constraint bounds error relative to the trend, not geological error**: if the trend is wrong, the surface can satisfy it and follow another event.

Cycle skipping is the discrete case of the same mechanism: with probability q per step, η_e takes the value of one wavelet period, T_w, and the expected number of skips along a path of length L is qL. Reducing q (Sections 4.1, 4.4 and 4.5) delays failure. The trend tolerance rejects isolated skips only if $\tau <$

$T_w/2$; with a larger τ it stops only multi-loop drift. On F3, $T_w \approx$ 32 ms, so the threshold is τ ≈ 16 ms.

Prediction and decision. The path enters the tracker in two places: in the *prediction* of the position $\hat{z}$, and in the acceptance *decision*, when the candidate is compared with the parent trace. The second entry is avoidable: comparing with the seed trace makes the score independent of path length, although the candidate positions still depend on it. This is the fourth consequence of (1) — the effective non-local constraint is not a band around the trend but a fixed comparison reference. It does not eliminate cycle skipping, because the neighbouring loop also resembles the seed; it stops the reference from drifting with it.

3.3 Conditioning of the trend

The trend T is interpolated from the picks. If they lie on a single section, T is extrapolated perpendicular to it without information, and any constraint based on it loses meaning. The same holds for diagnostics: the ratio R of Section 6.1 compares the surface with picks from a single line. Figure 1 shows that coverage depends strongly on the seed distribution — three on one line against twenty-two over several sections — but it is an interactive session, not a controlled experiment.

4 METHOD

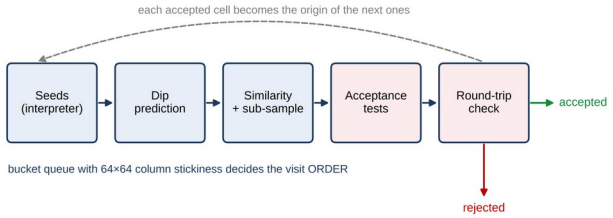


Figure 2. Tracking pipeline. Red: acceptance tests; the dashed arrow indicates that every accepted cell becomes the origin of the next ones. The bucket queue with column stickiness decides visit order (Sections 4.6 and 4.8).

Algorithm 1 summarises the tracker; the following subsections justify each step.

Algorithm 1 Seeded tracking with a bucket queue

```
input: a, S, barrier B, trend T, τ, W
 1 Q ← bucket queue; ∀(x,z)∈S: Q.push(x,0)
 2 while Q ≠ ∅:
 3   x ← Q.pop()      ▷ lowest band; same
                      ▷ 64×64 brick if possible
 4   for y ∈ N4(x), z(y) undefined:
 5     if B(y)=1: reject(fault); continue
 6     ẑ ← z(x) + p(x)            ▷ (5)
 7     d* ← argmax_{|d|≤W} ρ(d)   ▷ (2)
 8     z′ ← ẑ + d* + δ            ▷ (4)
 9     if ¬accept(ρ,pol,a,|z′−T(y)|≤τ):
10        reject(reason); continue
11     if back(y→x) ≠ z(x): reject(round-trip)
12     else z(y) ← z′; Q.push(y, band(y))
output: z, rejection counters per reason
```

4.1 Similarity measure and its threshold

Given a reference trace a centred at t_0 and a candidate b centred at t_1, we seek the shift d, $|d| \le W$, that maximises

$$\rho(d) = \frac{\sum_{k=-M}^{M} a(t_0+k)\,b(t_1+k+d)}{\sqrt{\sum_k a^2(t_0+k)}\,\sqrt{\sum_k b^2(t_1+k+d)}} \qquad (2)$$

Expression (2) does not subtract the mean: it is cosine similarity, which coincides with Pearson's coefficient only for zero-mean data. In processed data the DC component is removed and the distinction is irrelevant; after certain attributes, a common component inflates the numerator. This is a stated limitation.

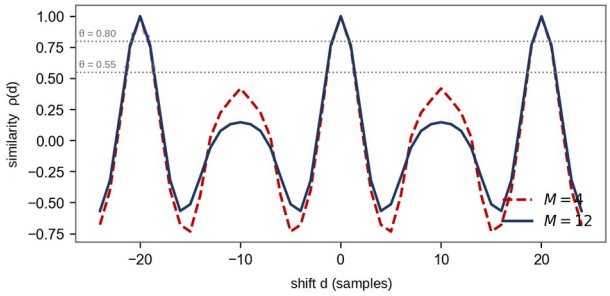


Figure 3. Similarity as a function of shift on a quasi-periodic trace. With M = 4 the adjacent loops score almost as well as the correct one; with M = 12 the correct peak stands out. Horizontal lines mark the 0.80 and 0.55 thresholds (Section 6.1).

Window. The correlation of a quasi-periodic signal over little more than one period is itself quasi-periodic (Figure 3); the window must span several cycles for their joint signature to distinguish one loop from its neighbour. We adopt M = 12 (±48 ms, a 100 ms window) and W = 5; Section 6.2 measures the effect of the window.

Threshold under the null hypothesis. For two independent segments of isotropic Gaussian noise with n effective samples, the statistic $t = \rho\sqrt{(n-2)}/\sqrt{(1-\rho^2)}$ follows a Student t distribution with $n-2$ degrees of freedom (Fisher, 1915). Because the tracker picks the best of $2W+1$ shifts, the probability of accepting noise is bounded by

$$P(\max \rho > \theta) \le (2W+1)\,P\left(t_{n-2} > \frac{\theta\sqrt{(n-2)}}{\sqrt{(1-\theta^2)}}\right) \qquad (3)$$

With n = 25 and $2W+1$ = 11, the bound is 2.4×10^{-2} for θ = 0.55 and 8.7×10^{-6} for θ = 0.80. Since seismic noise is band-limited, the effective number of independent samples is lower than $2M+1$; with n = 11 the values rise to 0.36 and 0.017. In both cases θ = 0.55 accepts noise at a rate that matters across 6×10^5 cells, whereas θ = 0.80 does not. The argument has a precise reach: it concerns the acceptance of **noise**, not cycle skipping. A neighbouring loop of the same event correlates above 0.80 (Figure 3), so no similarity threshold rejects it; that is the job of amplitude tolerance and the round trip.

4.2 Sub-sample refinement

An integer position carries the quantisation of the volume: ±2 ms at 4 ms sampling, about ±2.5 m at 2500 m/s. It shows up as terracing and propagates into derived maps — notably isochores, where the difference of two quantised surfaces produces a checkerboard with no geological meaning. A parabola is fitted to the values y_1, y_2, y_3 at the samples before, at and after the extremum:

$$\delta = \frac{1}{2}\,\frac{y_1 - y_3}{y_1 - 2y_2 + y_3}, \quad |\delta| \le ½ \qquad (4)$$

returning zero for a null denominator or a non-finite result; a vertex outside [−½, ½] indicates that the true extremum lies on another sample.

4.3 Orientation prediction

Without prediction, the search is centred on the previous position and, on a dipping reflector, the correct event may leave the window while the neighbouring loop enters it. Local dip p (samples/trace) is estimated from the structure tensor of the derivatives g_t (along the direction of advance) and g_z (vertical), smoothed by $\langle\cdot\rangle$:

$$p = -\frac{J_{tz}}{J_{zz}}, J_{uv} = \langle g_u g_v \rangle \tag{5}$$

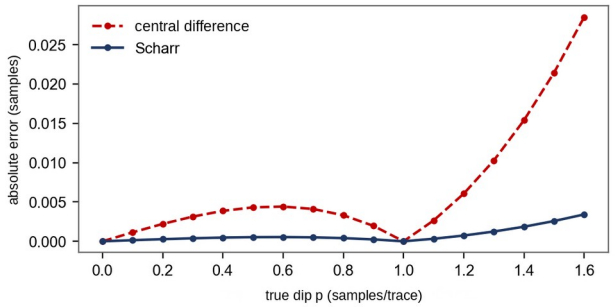


Figure 4. Absolute error of the estimated orientation on a synthetic field of known dip. The central difference is systematically biased, and the bias grows with dip — it is largest exactly where prediction matters most.

Derivative kernels. A central difference attenuates high frequencies unequally along the two axes, because the wavelet varies fast along the trace and slowly across traces; the ratio (5) is biased, always in the same direction (Figure 4). Scharr's kernels (Scharr, 2000), optimised for angular error, separate derivative and cross smoothing:

$$k_d = \langle -½, 0, ½ \rangle, k_s = \frac{1}{16}\langle 3, 10, 3 \rangle \tag{6}$$

Inner scale. A 3 × 3 stencil cannot resolve dips steeper than about one sample per trace: at dip 2.5 the gradients alias and the orientation collapses towards horizontal, with a measured error of 1.77 samples on the flank of a dome. Smoothing along the trace before differentiating, with a five-point binomial kernel, reduces the error to 0.56; since the error remained at 1.9 at dip 3, two passes are applied. This is an explicit trade-off between angular reach and vertical resolution.

Confidence. The eigenvalue contrast provides, at no extra cost, a measure of linearity

$$lin = \frac{\sqrt{(J_{tt} - J_{zz})^2 + 4J_{tz}^2}}{J_{tt} + J_{zz}} \tag{7}$$

which equals 1.000 on a synthetic layered field and 0.191 on pure Gaussian noise, with the same stencil and window.

4.4 Acceptance tests

Each candidate goes through the tests of Table 1, with one counter per rejection reason. **Amplitude tolerance** is the criterion that actually stops the cycle skip: the loops above and below are the same wavelet and pass the similarity threshold easily, because (2) is scale-invariant; what distinguishes them is amplitude. The **comparison reference** decides against which trace similarity and amplitude are measured — the parent or the seed (Section 3.2).

Table 1. Acceptance tests and default values, in increasing order of cost after the fault barrier.

Test	Default	Stops
Minimum similarity	$\rho \geq 0.80$	Low correlation
Polarity	that of the seed	Opposite event
Minimum amplitude	0.30 × seed	Dying reflector
Maximum step	2 sm./trace	Unrealistic dip
Amplitude tolerance	$\lvert a/a_{ref} - 1\rvert \leq 0.60$	Cycle skip
Trend tolerance	off ($\tau = 0$)	Skips only if $\tau < 16$ ms

An earlier version used the trend as a *guide*, centring the search on an inverse-distance interpolation, which tied the result to the seeds themselves: the trend is not evidence. As a *tolerance* it is coherent, but measurement shows its limit: with $\tau = \pm 60$ ms coverage fell from 20.8 % to 15.1 %, with the same worst step and no anomalous step in either case. The band rejected cells without rejecting skips — as Section 3.2 predicts, since 60 ms is far above $T_w/2 \approx 16$ ms. It is therefore off by default.

4.5 Round-trip check

All the tests above are statements *about the candidate*, and a tracker that slips one loop can satisfy all of them. The round-trip check looks the other way: having found z' at y starting from x, it searches for the event at x starting from z'; if the answer is not $z(x)$, the step is rejected. Because (2) is symmetric — for the same windows, $\rho_{AB}(d) = \rho_{BA}(-d)$ — the test accepts exactly the pairs that are a **mutual best match** within the two search windows. It therefore detects asymmetric skips, such as the one in Figure 5, where the chosen loop points back to a different loop; it **does not detect** a skip in which the wrong loop is the mutual best match. It is an inconsistency detector, not a proof of correctness, and its detection rate must be measured against a reference (Section 7.2). The cost is one extra search per accepted step.

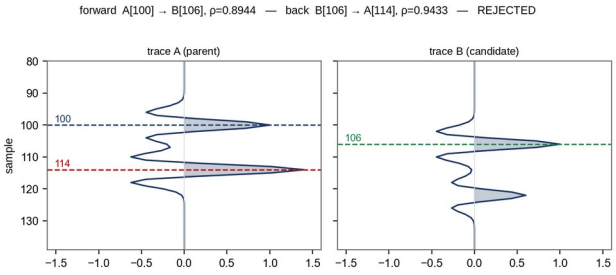


Figure 5. A step accepted by every forward test and rejected on the way back. Forward, from A(100), the search picks B(106) with ρ = 0.8944, above threshold; backward, from B(106), it picks A(114) with ρ = 0.9433. The two traces disagree about the identity of the event.

4.6 Growth order

Growth priority decides which cell is visited next and is independent of the acceptance rule. **Amplitude-first priority** always picks the best correlation and produces long, narrow fingers that move away from control; where two fingers meet, $\ell(x, y)$ in (1) is of the order of the finger length and the seam is a near-vertical blade. **Proximity-first priority** grows a wavefront — everything one step from the picks, then everything two steps away — and no branch outruns the front, which keeps ℓ small between neighbours. The defect is in no single step: every cell in a blade passes every test. It is in the order. In terms of (1), blades arise mainly from the systematic term: neighbours reached by paths of very different length accumulate $b(L_x - L_y)$, which grows linearly with the difference.

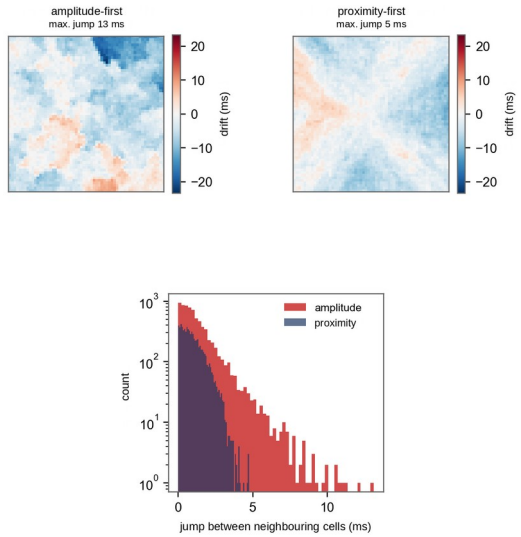


Figure 6. Simulation of the mechanism of equation (1): each accepted step adds an independent drift. Top: final drift under both priorities; bottom: distribution of jumps between neighbouring cells (log scale). Maximum jump 13.1 ms versus 4.8 ms. The simulation demonstrates the mechanism; magnitudes measured on F3 are in Section 6.2.

4.7 Map-view fault barrier

A tracker that ignores a fault crosses it and produces a smooth surface exactly where the structure should be broken. The conventional approach queries a coherence volume (Bahorich and Farmer, 1995; Marfurt *et al.*, 1998). For the question the tracker asks — *may I move from this map position to that one?* — the three-dimensional information is largely redundant, because a high-angle fault intersects the map along a line. The barrier is therefore one byte per map position, with a memory ratio

$$\frac{N_i N_c N_s \times 4}{N_i N_c \times 1} = 4 N_s = 1848 \qquad (8)$$

or 619 kB against 1.14 GB on F3. This ratio is an identity of construction, not a result; the real gain is that the test becomes one read from a contiguous vector, run before any correlation. The barrier does not remove the need to interpret faults — it must be derived from an attribute or from picks — and it is not equivalent to the volume: on a dipping fault the map trace changes with depth, so the barrier must be built at the depth of the horizon (for instance along the trend T) or it over-blocks. Low-angle faults are the extreme case.

4.8 Access locality

The samples of a trace are contiguous in memory, with linear index

$$idx(i, c, s) = i N_c N_s + c N_s + s \qquad (9)$$

A pure wavefront visits cells scattered across the whole map before returning to the neighbourhood of any of them — the worst possible pattern in a brick-compressed volume, where each access may force decompression of an entire brick. The history of the code shows it. The first version ran in 5 s, but by accident: a constant priority made the visit order arbitrary and cache-local — good locality, no guarantee of geometry. Correcting the ordering to the wavefront of Section 4.6 took the same tracking to over 600 s. We then replaced the priority queue with a **bucket queue** (Dial, 1969), with confidence bands instead of a total order and **stickiness to 64 × 64 columns**, matching the brick size of the ZGY format, so that each decompressed brick is exhausted before the next (Figure 7). The arithmetic is the same, but the result is not guaranteed to be identical: reordering visits within a band can change the parent of a cell and, with it, the acceptance decision.

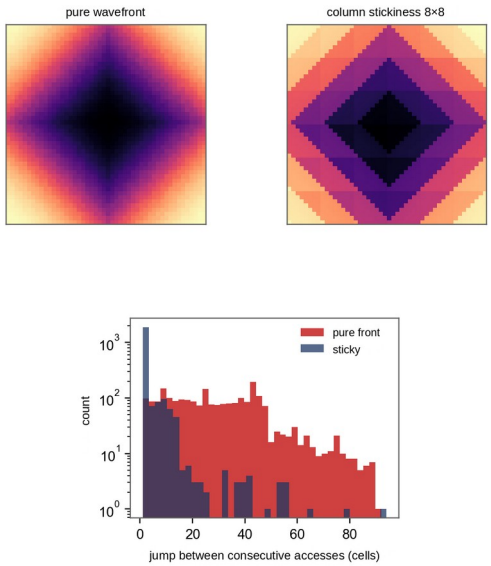


Figure 7. Visit order (top; colour = visit time) and distribution of the jump between consecutive accesses (bottom, log scale). The pure front jumps all over the map; stickiness exhausts one brick before the next. Illustrated on a reduced grid with 8 × 8 bricks; 64 × 64 in production.

4.9 Guided picking by dynamic programming

Between two known points on the same section, the optimal path is obtained by dynamic programming with backtracking (cf. Hale, 2013), with transition cost

$$C = \alpha(1 - \rho) + \beta \frac{(z - z' - p)^2}{\sigma^2} + \gamma \left| \ln(a/a') \right| \qquad (10)$$

Two flaws in the initial formulation, which we have not found documented, made it unusable. **Linear term:** with a cost proportional to $|z - z'|$, a twelve-sample jump costs the same as twelve one-sample steps; the quadratic term makes it twelve times more expensive — convexity is the mechanism. **Zero-dip reference:** without the $-p$ term, on the flank of a dome dipping 3 samples/trace, following the reflector cost $3^2 = 9$ and a horizontal path cost 0, so cutting across the structure was *cheaper* than following it. Subtracting the prediction, the costs swap exactly: 0 and 9.

5 DATA AND PROTOCOL

All measurements use the F3 survey from the Dutch continental shelf (dGB Earth Sciences): 651 inlines × 951 crosslines, 462 samples at 4 ms, 25 m bins, inline azimuth 358.4°. Times are medians of five runs, in an optimised build with a warm system cache; debug-build measurements are not reported. The

parameters — τ, κ, *W*, *M*, seed distribution and active constraints — depend on the experiment and are stated for each. Platform: Intel Core Ultra 9 185H, 31.4 GB of memory, g++ 13.1.0 (MinGW) with -O3, 256 MB column cache; growth runs on a single thread.

Given the silent failure mode, each component is verified against a case with a **known analytical solution**, not against the program's own output: a sinusoidal field of known dip (orientation), a wavelet evaluated off-grid (refinement), a synthetic fault at which tracking must stop (barrier), and a bit-exact round trip (persistence). The suite has 209 tests — 203 unit tests and 6 test programs, with 1080 assertions — and includes a third-party published reference set whose expected values do not come from this code.

Operational definitions. A cell sits at a *local extremum of the correct polarity* if the amplitude at *z* is the extremum of the seed polarity with respect to the adjacent samples (±1). This is the weakest possible test: it shows only that the surface rests on a peak or trough of the trace, and the 99.9 % must be read with that reach. A neighbour pair has an *anomalous step* if $|z(x) - z(y)| > 40$ ms. Neither definition uses a geological reference: both measure internal consistency.

6 RESULTS

Table 2. Summary of results measured on F3. Each row isolates one mechanism; all other conditions are held fixed. No value is measured against an independent geological reference (Section 7.2).

Mechanism	Before	After
Coverage, parent → seed[a]	91.5 %	38.0 %
Time span, same[a]	511 ms	396 ms
same, seed picks	276 ms	—
Coverage, trend off → ±60 ms	20.8 %	15.1 %
Steps > 40 ms, amplitude → proximity	40	0
Steps > 40 ms, $M = 4 \rightarrow 12$	15	0
Worst step, $M = 4 \rightarrow 12$	72.7 ms	34.5 ms
Orientation field error	0.094 sm.	0.004 sm.
Tracker drift (synthetic)	17.4 sm.	0.046 sm.
Time, priority queue → buckets[b]	> 600 s	3.0 s

[a] θ = 0.55, 43 seeds on one inline. [b] the original version, with degenerate ordering, ran in 5 s. sm. = samples.

6.1 The failure mode, quantified

This is the central result. With θ = 0.55 and comparison with the parent, the tracker filled **91.5 %** of the survey from **43 picks on a single inline**, rejecting only 2,863 of 619,101 cells, with trend tolerance off. 99.9 % of the cells sat at a local extremum of the correct polarity, and the surface spanned 511 ms where the picks spanned 276 ms.

Suspect, not refuted. The ratio 511/276 does not prove the error: with seeds on a single line, a horizon dipping away from it may legitimately span more time. What is shown is narrower and enough for the thesis: **local tests cannot tell the two surfaces apart** — the correct one and one that has changed loop would both pass with 99.9 %. Claiming more would commit the very error this paper describes. Changing only the comparison reference to the seed reduced coverage to 38.0 % and the span to 396 ms. Less coverage does not prove correctness — 396 ms is still above the picks — but it shows how much of the original surface depended on the decision accumulated along the path.

The practical consequence is counter-intuitive: **a permissive threshold does not produce visible errors — it produces coverage**. The disproportion between picks and coverage is itself the warning. We therefore propose the time-span ratio $R = \Delta z_{surface} / \Delta z_{picks}$ as a warning sign — not as proof; here $R = 1.85$. The value θ = 0.80 is adopted and is consistent with bound (3).

6.2 Growth order

Under amplitude-first priority, 40 pairs of neighbouring cells showed steps above 40 ms, up to 275 ms, on a surface whose mean step is 1.15 ms; under proximity-first priority, with the same criteria, none did. The correlation window has an effect of the same order: going from $M = 4$ to $M = 12$ reduced steps above 40 ms from 15 to 0 and the worst step from 72.7 to 34.5 ms. The two factors were measured one at a time and their interaction was not, so we do not attribute the removal of the steps to either alone. The mean step is blind to this defect: it is the tail that reveals it.

6.3 Orientation

Replacing the central difference with Scharr kernels reduced the orientation field error from 0.094 to 0.004 samples, and the final drift of the tracker on a synthetic case from 17.4 to 0.046 samples — two to three orders of magnitude, at the same computational cost.

6.4 Computational cost

Three times tell the story: 5 s in the original version, with degenerate ordering; over 600 s after correcting the order; 3.0 s with the bucket queue. The honest figure is not "200 × faster" but "the correct order cost over 100 × and the bucket queue recovered it". The total arithmetic — 25 samples, 11 shifts, 619,101 cells — is about 5×10^8 operations, under one second: **the time was in the accesses**. Read rates do not compare directly with times. The bucket queue reads about 4.1×10^6 samples/s; the priority queue started at 2.18×10^5 (19 × less) and fell to 8.9×10^4 at 200,000 cells (46 × less). Since the rate decays as the front grows, total time integrates the whole curve and the cumulative ratio exceeds both. The cell-by-cell difference between the two queues was not measured.

6.5 Coverage

Reported coverages compare only with their configuration: 91.5 % and 38.0 % (θ = 0.55, 43 seeds on one inline, parent or seed as reference); 20.8 % and 15.1 % (trend off or at ±60 ms, other parameters equal); 32.3 % with interpreter seeds and 5.1 % with automatic seeds; and 75.9 % reported by the application in the session of Figure 1. Coverage depends mostly on the seed distribution (Section 3.3) and **is not a quality measure**: more coverage can be worse, and less coverage does not prove it is better. Automatic placement does not replace the interpreter.

7 DISCUSSION

7.1 What generalises

Three results reach beyond this software. **Local criteria are insufficient for propagation problems with accumulation:** any algorithm that propagates state step by step, accepting each step by local criteria, obeys (1) — video object tracking, phase unwrapping and first-break picking share the structure. **Visit order is an algorithm parameter:** it determines the shape of the tree and hence where discontinuities appear. **In large volumes, locality dominates arithmetic:** the order that geometry requires is the worst for memory, and the same arithmetic ranged from 3 s to over 600 s through access order alone.

7.2 Threats to validity

Internal validity. Comparisons isolate one factor at a time, but no result has yet been measured against a geological reference. That reference exists: the F3 repository includes interpreted horizons, and Silva *et al.* (2019) published nine horizons

manually reinterpreted by two geoscientists, in open access. Tracking those horizons from seeds taken on them measures absolute error in milliseconds, the true skip rate, and the fraction detected by the round trip. **This is the missing experiment and the main gap in this work. External validity.** A single survey of relatively simple geology and good quality: the mechanisms are general; the magnitudes may not transfer. **Construct validity.** Model (1) is an explanatory hypothesis, not a statistical model fitted to the tracker: constancy of b along the path and short-range correlation of η have not been verified. **Ambiguity of Figure 1(d).** The near-vertical walls by the central structure may be real throw or growth blades; only comparison with the reference can tell them apart. **Design limitations.** Similarity without mean removal overestimates on data with a trend; the two-dimensional barrier must be built at the horizon depth; trend tolerance helps only below half a period and with distributed picks; comparing with the seed assumes the wavelet changes little laterally.

7.3 Reproducibility

Figures 3 to 7 are generated by code from the same expressions as the text, not drawn by hand: a computed figure would change if the algorithm changed. While preparing this manuscript, two discrepancies between text and generator were detected by this mechanism and corrected.

8 PROPOSAL: MACHINE-LEARNING INTEGRATION

> *This section is a proposal. It describes an architecture, a supervision scheme and an evaluation protocol; it presents no trained models and no measured performance.*

Training a network to output relative geologic time directly (Geng *et al.*, 2020) faces two objections. It requires fully interpreted volumes, which are scarce and of unknown quality — by Section 6.1, a published interpretation may contain cycle skips that nobody detected. And it replaces the tracker instead of informing it, losing interactivity. We propose instead to **replace components while keeping interfaces** (Table 3).

Table 3. Proposed integration points and label sources.

Point	Learned component	Labels
P1 Seeds	suitability field (3-D CNN) or policy	final coverage
P2 Dip	regression of the field p	exact synthetics
P3 Similarity	siamese network (Hadsell *et al.*, 2006)	round-trip pairs
P4 Acceptance	binary classifier	round-trip decisions
P5 Order	visit policy	coverage and cost

8.1 Weak supervision by the tracker itself

The round-trip check is an oracle that does not depend on human interpretation: running the tracker over uninterpreted volumes, every attempted step yields a labelled example ($y = 1$ if forward and backward agree, 0 otherwise), in the manner of programmatic weak-supervision schemes (Ratner *et al.*, 2017). A single F3 horizon yields hundreds of thousands of pairs. The central objection must be stated: **a model trained on the tracker's decisions learns the tracker, not the geology**. There are two partial answers — if the round trip is conservative, a hypothesis not yet measured, a model that reproduces it with higher coverage may be better; and the labels for P2 are synthetic and independent of the tracker — but neither removes the need for validation against the reference horizons of Section 7.2. By Section 4.5, these labels also inherit the round trip's blindness to mutually consistent skips.

8.2 Specific risks

P3 fixes a real limitation — the scale invariance of (2) makes it blind to the amplitude difference that betrays a cycle skip — but P4 is *less auditable* than five named thresholds and loses the explanation to the interpreter that rejection counters provide; we propose keeping the counters in parallel. P5 carries a hard, non-obvious constraint: any learned policy must include access cost in its objective, or it will lose the gain of Section 6.4.

8.3 Evaluation protocol

Comparing coverage is not enough: Section 6.1 shows that more coverage can be worse. We propose four metrics in order of precedence: (1) the time-span ratio R, close to 1; (2) the tail of the distribution of neighbour jumps; (3) **hold-out cross-validation** — on a reference horizon (on F3, those of Silva *et al.*, 2019), N lines are picked, tracking uses $N - k$, and the error in milliseconds is measured on the k held out; and (4) only then coverage. Metric 3 is the only honest comparison against a known answer and penalises cycle skipping directly, since a skipped surface errs by about one wavelet period on the held-out lines. Any claim that a learned component outperforms the method of Sections 4–6 must rest on it, with held-out lines chosen before training.

The greatest risk is **silent degradation**. The whole architecture of Sections 4–6 exists because this domain gives no signal when it is wrong; a learned component does not change that property — it amplifies it, because it is harder to inspect. Any integration must bring its own mechanism for detecting silent error; without one, it should not be integrated.

9 CONCLUSIONS

The main failure mode of seeded horizon tracking is silent and, by model (1), undetectable by local criteria by construction: a surface with 99.9 % of cells at local extrema and a time-span ratio of 1.85 passes every local test, whether right or wrong. The non-local constraint that worked was comparing with the seed trace rather than the parent; the band around the trend cannot reject one-loop skips above half a period and was switched off. The correlation window and growth order control anomalous steps, and the correct order, costly for memory, became viable with a bucket queue: from over 600 s to 3.0 s. The next step is defined and feasible with public data: compare with the manually interpreted F3 horizons and show, on seismic sections, the cases of skips and blades. Only then does the 511 ms surface move from suspect to refuted — or to correct.

DATA AVAILABILITY AND ACKNOWLEDGEMENTS

The F3 dataset is made publicly available by dGB Earth Sciences. Measurements on well data used the Roxar RMS training project, published independently. This work neither uses nor links against reciprocally licensed code; dependencies are permissive: OpenZGY (Apache-2.0), glm and spdlog (MIT) and Qt (LGPL-3.0, dynamically linked). The figure generator accompanies the work. Reference horizons: Silva *et al.* (2019), doi:10.5281/zenodo.1422787.

REFERENCES

Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6), 641–647.

Alaudah, Y., Michałowicz, P., Alfarraj, M. and AlRegib, G. (2019). A machine-learning benchmark for facies classification. *Interpretation*, 7(3), SE175–SE187.

Bahorich, M. and Farmer, S. (1995). 3-D seismic discontinuity for faults and stratigraphic features: The coherence cube. *The Leading Edge*, 14(10), 1053–1058.

Bigün, J. and Granlund, G. H. (1987). Optimal orientation detection of linear symmetry. *Proc. 1st ICCV*, 433–438.

dGB Earth Sciences. *F3 Demo dataset*, offshore Netherlands. Public data.

Dial, R. B. (1969). Algorithm 360: Shortest-path forest with topological ordering. *Communications of the ACM*, 12(11), 632–633.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10(4), 507–521.

Fomel, S. (2002). Applications of plane-wave destruction filters. *Geophysics*, 67(6), 1946–1960.

Geng, Z., Wu, X., Shi, Y. and Fomel, S. (2020). Deep learning for relative geologic time and seismic horizons. *Geophysics*, 85(4), WA87–WA100.

Hadsell, R., Chopra, S. and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. *Proc. IEEE CVPR*, 1735–1742.

Hale, D. (2013). Dynamic warping of seismic images. *Geophysics*, 78(2), S105–S115.

Hoyes, J. and Cheret, T. (2011). A review of "global" interpretation methods for automated 3D horizon picking. *The Leading Edge*, 30(1), 38–47.

Kalal, Z., Mikolajczyk, K. and Matas, J. (2010). Forward–backward error: Automatic detection of tracking failures. *Proc. 20th ICPR*, 2756–2759.

Kass, M. and Witkin, A. (1987). Analyzing oriented patterns. *Computer Vision, Graphics, and Image Processing*, 37(3), 362–385.

Lomask, J., Guitton, A., Fomel, S., Claerbout, J. and Valenciano, A. A. (2006). Flattening without picking. *Geophysics*, 71(4), P13–P20.

Marfurt, K. J., Kirlin, R. L., Farmer, S. L. and Bahorich, M. S. (1998). 3-D seismic attributes using a semblance-based coherency algorithm. *Geophysics*, 63(4), 1150–1165.

OpenZGY project. *Open source ZGY accessor library*. Apache-2.0 licence.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S. and Ré, C. (2017). Snorkel: Rapid training data creation with weak supervision. *Proc. VLDB Endowment*, 11(3), 269–282.

Scharr, H. (2000). *Optimale Operatoren in der digitalen Bildverarbeitung*. PhD thesis, University of Heidelberg.

Silva, R. M., Baroni, L., Ferreira, R. S., Civitarese, D., Szwarcman, D. and Vital Brazil, E. (2019). Netherlands dataset: A new public dataset for machine learning in seismic interpretation. *arXiv*:1904.00770.

Stark, T. J. (2004). Relative geologic time (age) volumes — Relating every seismic sample to a geologically reasonable horizon. *The Leading Edge*, 23(9), 928–932.

Wu, X. and Hale, D. (2015). Horizon volumes with interpreted constraints. *Geophysics*, 80(2), IM21–IM33.

Wu, X., Liang, L., Shi, Y. and Fomel, S. (2019). FaultSeg3D: Using synthetic data sets to train an end-to-end convolutional neural network for 3D seismic fault segmentation. *Geophysics*, 84(3), IM35–IM45.