

Limitations, and how far to trust the answer

BETA - testing stage. PVDCal is under testing. Results are for evaluation only and must not be used for construction design.

This is an EXTRACT from the PVD Design Tool user guide, not the guide. It carries the chapters that state where the model is valid, what it does not cover, and what a back-analysed forecast does and does not establish. They are published on their own because a reviewer should not have to take a 226-page download to read them.

Everything else - the method, the verification against eight closed-form solutions, the worked examples, the parameter reference and the validation campaign - is in the full guide, free and without sign-up, at pvdcal.geotechlogic.com/docs.

Cut from PVD_Design_Tool_User_Guide.pdf (customer edition, 2026-10-28). The page numbers below are the full guide's own, so any passage here can be found in the complete document.

WHAT IS IN THIS EXTRACT

1.9 Validity domain (pages 16 to 17 of the full guide)

the domain the model is valid in, as a table

Chapter 9 Limitations, and checking your own run (pages 88 to 89 of the full guide)

the limitations of the numerical method, and a check on your own run

Chapter 25 How far to trust the answer (pages 182 to 195 of the full guide)

what the back-analysis does and does not establish, and what is not modelled

It agrees with independent methods	T	Chapter 26 — Huang et al. (2025) BUS + Hypothesis B	Comparison is on the same site, and §26.5 concedes the reproduction attempt did not match the published result. §26.6 traces the mismatch to the inputs, not to an identified defect in either implementation: the paper's priors pin the settlement magnitude but leave the rate almost free, so they do not transfer as point estimates into a different forward model. A like-for-like comparison on matched inputs remains open
------------------------------------	----------	---	---

1.9 Validity domain

The conditions under which the tool's answers are meaningful, collected in one place (Table 1.8) so they can be lifted into the assumptions section of a report. Each is expanded where indicated.

Table 1.8. Validity domain — the assumptions a result stands on.

Assumption	What it means, and when it bites	Detail
Unit cell, no flow across the outer wall	Every drain is assumed to have an identical neighbourhood. At a block edge, a spacing change, or next to untreated ground, water does cross that surface and consolidation is over-predicted.	\$6.4
Small strain; the mesh does not deform	Drain spacing and layer thickness are held at their initial values. Large-strain sites are outside the model.	Chapter 9
Fully saturated, Darcy flow	No partial saturation, no threshold hydraulic gradient.	Chapter 9
Vertical permeability is derived from c_v and m_v	k_v is recovered per layer as $c_v \cdot m_v$ and harmonic-meant at layer interfaces, so a low- k_v band does restrict flow. The contrast is only as good as the m_v supplied for each layer.	—
Well resistance — modelled from a typed q_w ; D kept as a screening number	Since October 2026 a typed discharge capacity q_w is modelled in Design and Simulate on both routes — swept to convergence on the radial-only route, solved implicitly with each time step on the coupled route (any layer $c_v > 0$, or a drain stopping above the base). It slows consolidation and leaves the ultimate settlement unchanged; two-way discharge needs the drain to reach a draining base. The discharge factor $D = q_w / (k_h \cdot l_m^2) \geq 5$ (Mesri & Lo 1991; Hansbo 1997; Holtz et al. 1991; Rujikiatkamjorn & Indraratna 2007; Abuel-Naga et al. 2015) is still printed as a screening number and overstates the effect. The back-analysis does not model it: a fitted k_0 already includes the real drain's resistance, so a Design built on that profile must not apply q_w again. Watch the exceptions: very long drains, organic soils where q_w degrades with time (Miura & Chai 2000), kinked drains.	\$25.9
Minimum usable record for back-analysis	Below $U \approx 40\text{--}50\%$ the magnitude family is essentially unconstrained and the band collapses to the prior. $U \approx 50\%$ is typical, not a floor.	\$25.9
Smear is weakly identifiable	η trades off against k_0 almost exactly, so back-analysis holds it and the forward tasks range it. A fitted η should not be quoted as a measurement.	Chapter 22
Determinism and reproducibility	Design and Simulate: the same project file on the same build reproduces the same result, from the sampling seed saved in the file and printed on the report. Back-calculation: no random draws, so the same file on the same build and machine reproduces the answer (the built-in example to the millimetre); the last digits are platform-sensitive, and on one short-record audit case two runs settled 4 % apart in a flat valley — re-run and compare the search logs rather than assume.	\$20.4, \$25.9
Validation basis	One field case. Everything else is verification or self-consistency.	\$1.7

CHAPTER 9

Limitations, and checking your own run

9.1 Assumptions built into the numerics

Collected in one place, in Table 9.1, so a design report can state them.

Table 9.1. Numerical assumptions and their consequences.

Assumption	Consequence	When it matters
Unit cell with a no-flow outer wall	over-predicts rate where drains are not identical	block edges, spacing changes, adjacent untreated ground
Vertical permeability is derived, not typed	k_v is recovered per layer as $c_v \cdot m_v$, so a layered profile does carry a real contrast — but it is only as good as the m_v you supply.	profiles where m_v is left at a nominal value while c_v is varied per layer
Coefficients frozen within a step	coefficients lag the solution slightly	very rapid loading with strongly stress-dependent c_h ; shrinks with the step
Pore pressure clamped non-negative	strict conservation is broken while the clamp is active	only when it would otherwise go negative; disabled automatically for unloading and vacuum runs
Small strain; the mesh does not deform	geometry is fixed as settlement proceeds	very large strains, where drain spacing and layer thickness change materially
Fully saturated, Darcy flow	no partial saturation, no threshold gradient	desiccated crusts, very low gradients in some clays

9.2 A five-minute numerical check

For a design you are going to sign, the numerical quality of the run can be established quickly.

REPRODUCE IT

1. **Read the mesh back.** Open the *Numerical mesh* results card and check $N_r \times N_z$. Confirm the layer count and thicknesses are what you intended.
2. **Halve the vertical cell size** and re-run. If the answer moves by an amount you would care about, the mesh was too coarse for your profile — most likely because a layer is thin relative to the 0.5 m default.
3. **Halve the time step** and re-run. Expect a smaller change than step 2; if it is larger, your loading is rapid relative to the step.
4. **Check the degenerate limit.** Set $\eta = 1$ (no smear) and compare against the Hansbo ideal curve — the built-in benchmark cases do exactly this, and you can reproduce it with your own geometry.
5. **Sanity-check U.** Degree of consolidation is a degree of *primary* consolidation and must never exceed 100 %. Creep and immediate settlement belong to the settlement, not to U.

RECAP

The short version. The tool discretises the conservative form of the consolidation equation with cell-centred finite volumes on annular control volumes. The mesh is built around the physical boundaries, so every permeability jump falls on a cell face. Face conductivities are harmonic means, so each face carries one shared flux and the scheme conserves water exactly. Time integration is Crank–Nicolson, or Peaceman–Rachford ADI when the vertical direction is coupled; both are second order and unconditionally stable. Convergence has been measured, not assumed: second order in both directions, with the default mesh delivering around a tenth of a percentage point of error against exact solutions. Eight closed-form benchmarks verify the machinery. The remaining uncertainty in any real prediction lies in the soil parameters, not in the numerics — which is exactly where it should lie.

CHAPTER 25

How far to trust the answer

Coverage is being re-measured, and the figures below are withdrawn as current evidence.

This is the Part that leans on the campaign hardest, and every number in it was measured on an earlier version of the solver; the pre-declared shift test on the rebuilt solver was **material** (11 October 2026). Nothing here may be cited as a present-day reliability of this software, and the replacement campaign is under way. Take from the Part what does not turn on a number: how a reliability claim is built and tested, what a miss looks like on a chart, and which failure modes a wider range can and cannot repair.

This part states plainly what the guide's numbers do and do not establish. It positions the method against the prior literature, and it reports the empirical studies run to answer the questions two rounds of independent review raised. That includes the **band-coverage calibration** (§25.3), which a good first revision surfaced as the load-bearing issue. All drive the *real* forward solver.

A note on resolution. For replicate throughput these studies run on a deliberately coarse solver grid, and it matters which result that helps or hurts. Band *coverage* (§25.3) is a calibration property a too-tight interval exhibits at any resolution, so the N-1 finding is robust and if anything conservative. The speed-axis benefit (§25.4) and the structural-error signal (§25.6) instead depend on a *sharp* misfit landscape a coarse grid smooths — so null results there are reported as *unconfirmed at production resolution*, not refutations, with the production-grid re-run named as the next step.

25.1 What the evidence covers — and what it does not

WATCH OUT

Every synthetic result up to this point is a **solver self-consistency** check: the “truth” is generated by the same forward model the estimator scores against. Benchmark Project A, the three-knob evidence figures and the “ $\chi^2_r \approx 1$ noise floor” are all of this kind. That configuration (an *inverse crime*) can show the optimiser converges, that parameters are identifiable from a truncated record, and that the machinery is internally consistent — but it cannot, even in principle, expose *model inadequacy*, the dominant error in real back-analysis. That is why the one test against an instrumented case history, the Ballina embankment of §25.8, carries more weight than everything before it. It is also why a second such site remains the single most valuable addition to this tool's evidence base.

25.2 Where this sits in the literature (what is, and isn't, new)

The speed–magnitude trade-off the method “discovers” is the well-known rate–compressibility coupling of consolidation inversion, not a new phenomenon; freezing one family while sweeping the other is coordinate descent with a nuisance-parameter profile. The standard alternatives are Asaoka's observational method (1978), direct Hansbo unit-cell inversion, and Bayesian or Kalman-filter settlement updating. Against them, the contribution claimed here is narrow and specific: *robustness of*

the magnitude scan to a mis-anchored drainage rate under coupled, oppositely-directed input bias. It is obtained by carrying a bounded global speed axis through the scan and alternating (Chapter 22). That is a measurable design choice (§25.4), not a general claim of superior accuracy; establishing the latter needs head-to-head comparison against those baselines on common data. Chapter 26 is one such comparison, against a published Bayesian back-analysis of the same Ballina embankment. But it is a single site against a single alternative, which is a long way short of a general claim.

The method's ancestry, set out in Foundations at the opening of Chapter 21, bears on the evidence too. The coverage shortfall measured in §25.7 is not a private defect of this implementation. It is the documented property of behavioural bands, debated in print for twenty years: such bands are conditioned on the model structure and on the threshold, and they are not formal credible intervals (Beven & Binley 2014). The measurements in §25.7 calibrate the band empirically. That calibration is this part's answer to a limitation the family has always carried.

Naming the method precisely. The calibration is a deterministic, engineered search followed by a propagation step that can be read as quasi-Bayesian: the limits you typed act as the prior, the fit to your readings discounts each parameter set's vote, and the vote-weighted percentiles approximate the posterior forecast. What that reading does not carry on its label, and this one does, is a *measured* calibration — and the card prints beside the fit-narrowed row how many parameter sets effectively carry its edges and whether its two halves agree, in bold when they do not.

Calibration as a falsifiable contract. The deepest difference between this pipeline and a conventional Bayesian analysis is not algorithmic but epistemic: it lies in what kind of statement each makes about its own reliability, and in what it would take for that statement to fail. A posterior credible interval is exact *conditional on its premises* — the assumed error model, the prior, and the structural adequacy of the forward model. That conditional exactness is a genuine strength, and nothing here disputes it; but the analysis itself contains no trial of those premises, so when one fails the interval remains fluent and confident while its guarantee has quietly lapsed. This is not a defect of Bayesian reasoning — it is the price of deductive guarantees, and a rigorous Bayesian practitioner pays it down with posterior-predictive checks and sensitivity studies external to the inference proper. The reliability statements in this part are constructed the other way about. Every printed coverage figure is the outcome of a designed trial against known truths that was free to return a refuting answer, under acceptance criteria registered *before* any run stating what result would count as failure. The claim “this range contained the truth in 77 % of trials” is therefore falsifiable in the strict sense: a definite procedure existed that could have contradicted it, was executed in full, and did not.

The record shows this contract being exercised in both directions, which is what distinguishes it from ceremony. The $\Delta\chi^2$ region's nominal 90 % was measured at 58 % — and the finding was printed on the result card as a warning rather than absorbed into a footnote. The September 2026 construction drift was exposed precisely because a measured truth fell outside a range that claimed to contain it, and the repair was itself re-measured before shipping. The fit-narrowed row's own on-card diagnostics flagged, on their first outing, that its edges rested on too few effective recipes to be quoted as solid — and the card said so in bold. Three reliability claims exposed to refutation; three refutations found; three repairs, each re-verified by the same instrument that found the fault. A calibration that can fail out loud — and demonstrably has — is the only kind whose successes carry

evidential weight. That is this part's considered answer to the fair observation that behavioural methods lack the deductive guarantees of the formal Bayesian family: what this method forgoes in guarantee it repays in tested, repairable, and continuously re-measured performance, with the test protocol published alongside the number it produced.

25.3 How reliable is the reported range?

The result page reports two ranges around the forecast. They answer different questions, they have very different reliabilities, and the difference decides which one you design to.

Range	The question it answers	Typical width	Measured coverage
Including your input uncertainty	Given the Lo/Hi you typed for each layer, how far could the forecast move?	~18 % of the settlement	77 % [71–81]
Data-resolvable only (χ^2 region)	How tightly does the settlement record <i>alone</i> pin the answer?	~5–17 % (7.5 % pooled)	58 % [52–64]

Both are labelled P05–P95, so both nominally promise 90 %. Measured over 270 paired tests spanning five soil profiles, three record lengths, three noise levels and three grades of laboratory input, the data-resolvable range was measurably the narrower and the more over-confident of the two; both figures are withdrawn and are being re-measured on the current version of the solver. A second coverage round (September 2026; 315 cases, same star design, same seeds) re-measured every construction through the *shipped* code path rather than a stand-alone replica implementation, after an implementation drift had shown the difference matters. It replicated this study almost exactly — design range 79 % pooled against 77 % here, χ^2 region 60 % against 58 %, paired agreement on 90 % of matched cases — and added four findings. The pre-fix band construction measured 41 %, with 102 paired cases where the repaired construction covered and the old one did not, against one reversal. Omitting the calibration search entirely (the same propagation centred on the typed values) measured 37 % against 79 % — the deep search's centring is worth roughly forty points of coverage, and its centre error falls from 10 % of the settlement at quarter-records through 6 % at a third to below 3 % at half, while the typed guess stays near 14 % throughout. Under AR(1)-correlated noise ($\rho = 0.7$ — the realistic case of drifting instrument error) the χ^2 region's coverage fell further to 42 % while the design range held 73 %: the assumption-dependent band degrades when its noise assumption breaks; the typed-width band barely notices. The failure is repairable, and the repair is now shipped: type the drift into Observation accuracy and the score whitens the residuals and counts the effective readings honestly, recovering the χ^2 region to 83 % on the same 60 sites at more than double its width (\$27.5). And a dedicated one-third-record slice, run at the same breadth as each main truncation (105 cases: all profiles, all noise and laboratory-quality excursions), graded the truncation advice: the design range's coverage rose steeply with the length of the record — so “below about half, treat every range as provisional” stands, with one-third records usable only with caution. The figures are withdrawn and its coverage is being re-measured: the figures previously quoted here

were measured on an earlier version of the solver and are withdrawn, a campaign on the current version is in progress, and none is quoted until it reports. Broken down by how much settlement had been observed when monitoring stopped:

Record reaches	Data-resolvable range	Including your input uncertainty
a quarter of the settlement	40 %	43 %
half	68 %	96 % [89–98]
three quarters	66 %	91 % [83–95]

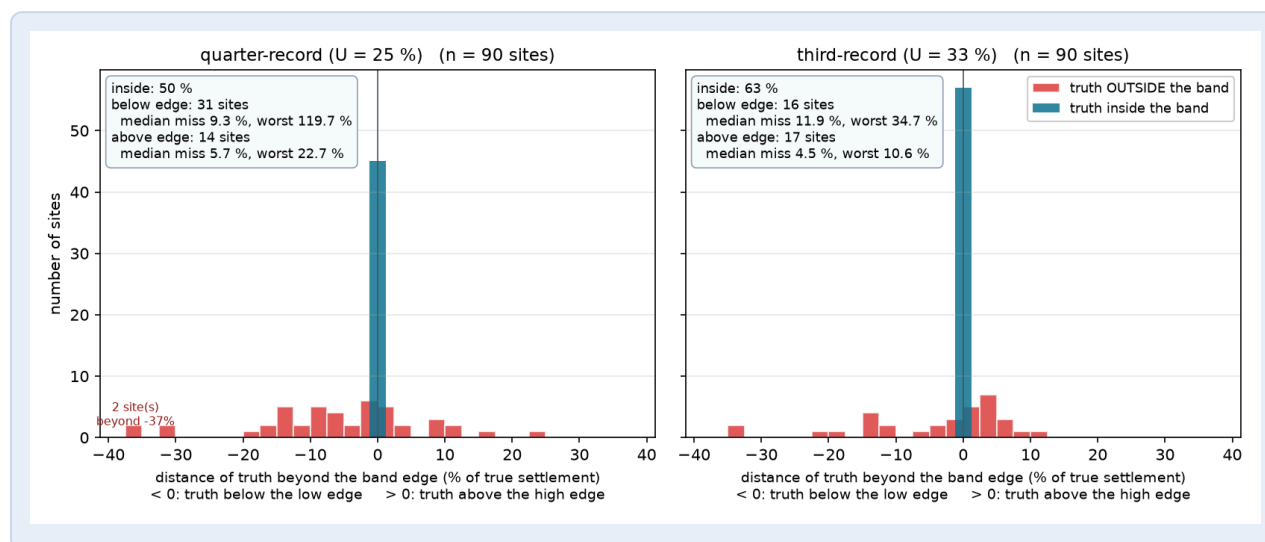


Figure 25.1. Where the truth sits relative to the design range at quarter-records (left) and third-records (right), 90 independent-noise cases each: teal = inside; each red bar a miss, placed by its distance beyond the nearer edge (negative = the range over-predicted). Correlated-noise cases excluded to match the 48 % and 63 % headline coverages.

The anatomy of a miss. Figure 25.1 answers the question the coverage percentages hide: when the range misses, by how much — and why. Four regularities stand out. *First, misses are mislocation events, not width events.* Among the cases the range contained, the winner's centre error is small — median 5.1 % of the settlement at quarter-records, 3.4 % at a third — while among the misses it is 19 % and 14 %: roughly twice the band's own half-width of 9–10 %. The band almost never misses because it is too narrow; it misses because the search, starved of record, parked it in the wrong place, and most misses are accordingly grazes of 5–12 % just beyond an edge with a thinner tail of larger errors. *Second, laboratory quality is the dominant controllable factor at short records.* At quarter-records the range contained the truth in 60 % of good-laboratory cases, 53 % of standard ones and only 27 % of poor-laboratory ones; by a third the same split has recovered to 80 % / 62 % / 53 %. With little record the search remains anchored near the typed values, so a biased laboratory drags the centre with it — and the study's single pathological case (a quarter-record, poor-laboratory profile whose range sat at more than double the truth, centre error +154 %) is exactly that mechanism at full strength. *Third, instrument precision matters far less than record length.* The low- and high-noise excursions differ from one another by no more than their small-sample uncertainty at either truncation, while a quarter-to-third increase in record length moves containment by thirteen points

and a quarter-to-half increase by forty-five: a longer record with noisy plates beats a shorter record with precise ones, which is worth remembering when a monitoring budget is being argued over. *Fourth, the miss directions are asymmetric in the direction that favours safety.* At quarter-records two thirds of the misses are over-predictions — the range sitting above the truth, the conservative error for settlement design — while the dangerous side, real settlement above the range's top, has both fewer cases and shorter reach: its worst case is 23 % beyond the edge at quarter-records and 11 % at a third. Lengthening the record cuts these tails before it repairs the averages — the worst over-prediction falls from 120 % to 35 % between a quarter and a third — and that tail-cutting, as much as the coverage staircase itself, is what the “wait until about half” advice buys.

KEY IDEA

Design to the input-uncertainty range. With a record reaching at least half the settlement it contains the truth about nine times in ten — close to the 90 % it claims. The data-resolvable range is genuinely informative about *what your measurements resolve*, and worth reading for that, but it is not a design envelope.

Why the narrow range fails, and why more data stops helping it

The data-resolvable range is a profile-likelihood region: the set of parameter combinations that fit the record about as well as the best one. It therefore measures how tightly *the measurements* constrain the answer — not how much you do not know. Those are different quantities, and the gap between them is where the error lives.

As monitoring accumulates the record pins the parameters ever more tightly, so this range narrows — from about 17 % of the settlement early on to about 5 % once three quarters has occurred. But the forecast error does *not* shrink as fast, because part of it comes from soil variability and model limitations that more settlement readings cannot resolve. The result is the pattern in the table above: **coverage stops improving as the record lengthens.** It is 68 % at half consolidation and 66 % at three quarters, while the interval keeps tightening. (Its lowest coverage, 40 %, is at the *shortest* record, where the centre itself is not yet located.) At three quarters consolidation it reports roughly ± 2.5 % and is right about two thirds of the time.

The input-uncertainty range does not have that failure mode. Its width comes from the laboratory scatter you typed, which does not shrink because a settlement plate was read for another six months. It stays near 18 % and its coverage rises with data rather than falling.

The failure is in the range, not in the search

One innocent explanation would be that the hidden truth simply lies outside the region the sweep can reach. It does not: across the replicate suite the true C_c and σ'_p both lay inside the expanded search walls in all 12 replicates of the calibration study (100 % [76–100]). **The truth is almost always inside the searched domain but outside the reported interval** — so the shortfall is a calibration fault of the interval, not a failure of the search to look in the right place. That is why the remedy was a different width, not a wider search.

WATCH OUT

Two limits that remain. Below about half consolidation neither range is reliable (43 % and 40 %): the central estimate itself is not yet located, and no width around a wrong centre recovers it — keep monitoring. And the *per-parameter* ranges (C_c , σ'_p and the rest) are narrower than their own errors justify; read them as diagnostic, never as a design envelope.

25.4 Does searching the speed axis actually help?

("Knob S" below is the search's bounded adjustment of the drainage rate while the magnitudes are explored; "frozen-speed" is the same search with that adjustment switched off.)

The headline "89.3 → 23.9 mm, 3.7× better" figure compares the method only against a crippled version of itself at two hand-picked bias corners. We re-ran the contrast over randomly-directed input bias and added a genuine non-ablation baseline ("use your lab values, just fit the rate"). Over **18** randomised bias directions, median held-out surface RMSE was **58 mm** (naive: typed centre + fitted rate), **33 mm** (frozen-speed scan) and **28 mm** (full three-knob); the full method beat its frozen-speed ablation in **7 / 18** runs, median improvement **-0.2 mm**.

Two distinct conclusions, and it matters which is which. First, the magnitude *sweep* is clearly worth doing. Both sweep variants roughly halve the median held-out error of the naive baseline. More importantly, they cut off its long tail: the naive forecast can be very wrong when the typed centre is biased. That is a real benefit of the core method against a non-ablation alternative. Second, and contrary to the headline, the *speed-axis / 3-D* refinement gives essentially **no average benefit** over the frozen-speed ablation across randomised bias — the median improvement is within noise of zero and the full method wins only about half the time. The large gains quoted in Chapter 22 are real but *confined to the specific opposite-direction-bias configuration* that maximises them. The honest reading: knob S is best understood as insurance against one particular failure mode (a mis-anchored rate under oppositely-biased inputs), not a general accuracy improvement.

Does the insurance actually fire when the covered event happens? The average benefit being ~zero is only defensible if the benefit concentrates in the regime knob S is for. So a second suite was run with the bias direction controlled — 8 runs in each regime, separate from the 18-run study above. In the **opposite-direction-bias** regime, where the settlement-raising group is typed low while the lowering group is typed high (the exact mode knob S targets), the full method beat its frozen ablation in **2/8** runs with a median gain of **-0.0 mm**. In the **coherent-bias** regime the win rate was **2/8** and the median gain **-0.1 mm**. So the safeguard is **not demonstrably firing** here, which on its face argues for making knob S opt-in rather than always-on.

WATCH OUT

One honest caveat before acting on that. The wrong-basin failure knob S exists to fix is a *sharp-landscape* phenomenon — two distinct misfit valleys that swap depth when the speed changes. The deliberately coarse study grid used here (chosen for replicate throughput) smooths the misfit landscape and may not reproduce those distinct basins. This null result is therefore *consistent with* knob S having nothing to fix at this resolution. It does not show the knob is useless at the production grid, where the original $89 \rightarrow 24$ mm gap was measured. The correct next step is to repeat this conditional test at the app's production solver grid before deciding opt-in vs always-on; until then knob S 's value is best called *unconfirmed*, not refuted.

25.5 Does the answer depend on the tool's internal settings?

Four load-bearing constants were previously asserted with a plausibility story but no sensitivity evidence. Varying each one-at-a-time on a representative noised case (others held at their shipped values) gives these swings in the delivered forecast and band width — the larger the swing, the more the constant is a genuine user-facing knob rather than a settled default:

Four internal settings — the interval threshold that closes the record-only region, the width of the rate adjustment the scan allows, the factor by which the typed walls are widened, and the discounting strength of the fit-narrowed row — were each varied over their tested ranges on a representative noised case. The largest effect on the delivered forecast was a 17 mm change in held-out error, and the largest on a range's width 53 mm; the discounting strength moved nothing measurable. None is a user setting, and their effect is small beside the soil uncertainty the ranges carry.

25.6 Testing against a different model — still an open question

The cleanest fix for the inverse crime is a field case history. Ballina (\$25.8) is one, but with a year or more of record its residual model error is absorbed. At the shortest cut, where it is not absorbed, a single site cannot isolate the mechanism. Absent a field case that clearly exercises a non-absorbable error, the second-best is a truth generated so that structural error is present by construction, then inverted with the default model. We tried two structural errors. Importantly, we **report this section as inconclusive**, because both proved largely *absorbable* by the free parameters for this radially-dominated PVD problem:

- **Unmodelled smear ($\eta=3$) + a 30 % void-ratio error.** With the truth matched to the model (inverse crime), the median over 12 replicates was $\chi^2_r \approx 1.20$, held-out RMSE **32 mm**, band coverage **58 %**. Regenerating the truth WITH a smear zone ($\eta=3$) and a 30 % void-ratio error, neither directly representable by the default inversion, gave $\chi^2_r \approx 1.19$, held-out RMSE **24 mm**, band coverage **83 %**. Did not separate: a smear zone is mostly a drainage-*rate* effect the speed family soaks up, and a void-ratio error is partly masked by the $C_c/(1+e_0)$ degeneracy.
- **A wrong drainage boundary (truth double-drained, model single-drained).** Over 12 replicates, the matched control fit to $\chi^2_r \approx 1.20$ with held-out RMSE **34 mm** and band coverage **83 %**; generating the truth with DOUBLE drainage moved these to $\chi^2_r \approx 1.19$, held-out RMSE

31 mm, band coverage **75 %**. Also did not separate — for a radially-dominated unit cell the vertical drainage boundary moves the settlement curve only slightly, so the radial speed knob absorbs the rest.

WATCH OUT

What this does and does not show. An unfalsified hypothesis from tests that could not falsify it is not evidence of robustness. Neither attempt exercised the failure mode it was built to probe, so **the cross-model question remains genuinely open**. It is quietly informative that several *plausible* structural errors turned out absorbable for a radial PVD unit cell. The test that would settle it is a decisively non-absorbable error: finite drain *well resistance*, which the inversion has no term for (the forward kernels model it from a typed q_w since October 2026, but the back-analysis does not pass q_w , so for the fit it remains an error the recovered k_0 must absorb); an unmodelled thin soft lens; or a second real instrumented field record. Ballina (§25.8) supplies the field data but does not exercise a non-absorbable error, so the decisive case remains outstanding. This section should be read as “not yet demonstrated either way,” not as reassurance.

KEY IDEA

One signal here *is* firm, and it is independent of the structural-error question. The reported band covered the truth in *well under* 90 % of runs **even in the matched inverse-crime case**. That independently corroborates §25.3: the band, at its shipped width, overstates its own coverage.

25.7 The reported range, measured properly

The band was measured on **270 paired runs** spanning five soil profiles, three truncations ($U = 25 / 50 / 75 \%$), three measurement-noise levels (5 / 10 / 20 mm) and three laboratory-quality levels ($\times 0.5 / \times 1 / \times 2$ disturbance of the typed values). Every run starts from *disturbed* inputs, never from the answer.

KEY IDEA

The band is over-confident, and more data does not fix it.

Condition	P05–P95 coverage of the true final settlement
U = 25 %	40 %
U = 50 %	68 %
U = 75 %	66 %
noise 5 / 10 / 20 mm	51 / 57 / 67 %
laboratory quality $\times 0.5$ / $\times 1$ / $\times 2$	71 / 56 / 51 %
pooled	58 % [52–64] against a nominal 90 %

Against a nominal 90 %. Note the third row: with the *most* monitoring the interval is narrowest (median width ~ 5 % of the settlement) yet covers the truth no better than at half consolidation (66 % against 68 %). The lowest coverage of all is at the *shortest* record (40 %), where the centre itself is not yet located. The point of this row is different: as data accumulate, the band narrows faster than the error shrinks, so it stops improving precisely where it looks most credible. **Report the P05–P95 range as indicative; do not use it as a design envelope.**

The range you are shown is the survivor of a systematic search for a better one. Eleven candidate constructions were built and measured against this suite; the two that held up across every noise level and every laboratory-quality level are what ship, and they are described below. The others improved matters in the conditions they were designed for and failed when those conditions changed.

That is worth one sentence of guidance: treat any uncertainty estimate, in this tool or any other, as unproven until it has been measured across different data quantities, noise levels and input quality. An interval that has only been checked under one set of conditions is a claim, not a measurement.

The band that replaced it — width from YOUR uncertainty, not from the data

The $\Delta\chi^2$ band answers “how tightly do my *data* pin this?” After three quarters of the settlement has occurred, the data pin the parameters very tightly, but to slightly the wrong place. The band therefore reports ± 2.5 % around an answer that is 4 % out. A second band was therefore built: **same winner at the centre, width taken from the Lo/Hi the engineer typed**. Laboratory scatter does not shrink because a settlement plate was read for another six months, so this band cannot collapse as data accumulate. Its width comes from the limits you typed, and from nothing else.

	$\Delta\chi^2$ band (data-resolvable)	Input-uncertainty band
U = 25 %	40 %	43 %
U = 50 %	68 %	96 % [89–98]
U = 75 %	66 %	91 % [83–95]
pooled	58 % [52–64]	77 % [71–81]
median width	7.5 %	18.5 % ($\times 2.5$)

KEY IDEA

The defect is removed. The $\Delta\chi^2$ band narrows 16.6 % $\rightarrow$ 5.1 % as data accumulate while its coverage *falls*; the input-uncertainty band holds ~18–20 % and its coverage *rises* to 91 % against a nominal 90 % — essentially calibrated. Paired McNemar over 270 runs: 61 runs covered where the $\Delta\chi^2$ band missed, 10 the other way, **p < 0.0001**. Better at every noise level and every laboratory-quality level tested, including deliberately poor laboratory data (67 % vs 51 %). It is now the headline range on the result page; the $\Delta\chi^2$ band is retained beneath it, labelled as what the settlement record alone can resolve.

WATCH OUT

It does not help below about half consolidation (43 % vs 40 %). There the winner itself is mislocated, and no width around a wrong centre recovers it. This strengthens rather than softens the rule that a usable back-analysis needs the record to reach $U \approx 50$ %.

Two things the tool tells you about itself

A value resting on a wall is flagged. Any recovered parameter that finishes at the edge of its search range is marked on the results page. This matters because such a number is not a measurement: the data were still pulling and the boundary stopped them, so what is reported is set by where the limit was placed rather than by the record. When you see the flag, widen that parameter's range and run again — and if it still rests on the wall, say so in the report rather than quoting the value as a result.

A layer the record cannot see is held to what you typed. A layer whose settlement has barely begun is not constrained by the monitoring record; what moves its parameters is the fit to the *other* layers. Where a layer has at least 1.5 times as much settlement still to come as the best-observed layer in the profile, the tool keeps its search inside your typed range instead of widening it, on the grounds that your laboratory value is then the better evidence. The results page lists every layer, how much of its settlement remains, and whether it was held, so the decision is visible rather than buried.

25.8 Ballina — the test against a real embankment

Everything in §25.3–§25.7 is an *inverse crime*: the same forward model both generated and inverted the data, so model error is zero by construction. The Ballina NTF trial embankment (Pineda et al. 2016; Kelly et al. 2018; Lim et al. 2018) is the one test with real soil, real construction and real measurement error. It is a 3 m embankment on PVD-improved marine clay, on a 1.2 m square grid, monitored for about three years. The tool was given the measured record up to a cut day and asked to forecast to day 1080.

Which measured record, and why it matters. Two published records exist for this embankment and they do not agree (§25.8.1). The results below are scored against **M0**, the top extensometer magnet, whose final reading is **1427 mm** at day 988 (the digitised tail wobbles by ± 16 mm — day 974 reads 1444 mm — vector-extraction noise, not real heave). M0 is used because its coordinates are extracted exactly from the vector artwork of Huang, Kelly et al. (2025) rather than traced by eye from a printed figure; because it is a single defined instrument rather than a mean over four plates at different positions; and because it is the record the comparator of Chapter 26 targets, which makes that comparison like-for-like.

Record given	Readings used	Forecast at day 1080	Error vs the final reading	$\Delta\chi^2$ band	Input-uncertainty band
to day 200	11	1293 mm	−9.4 %	1.7 % wide — misses	13.3 % wide — misses
to day 400	25	1406 mm	−1.5 %	7.4 % wide — covers	16.0 % wide — covers
to day 700	34	1416 mm	−0.8 %	4.3 % wide — covers	17.5 % wide — covers

How these errors are scored. The M0 record ends at day 988; the forecast runs to day 1080 and is compared against that last reading carried forward. Scored strictly at day 988, where a real measurement exists, the three errors are −10.3 %, −2.4 % and −1.9 % — about a percentage point worse than the table, and the honest figures to quote if you need a like-for-like epoch.

KEY IDEA

What the field test does show. With a year or more of monitoring the forecast lands within about 2 % of the measured settlement two years ahead (held-out RMSE 26 mm at the 400-day cut, 24 mm at 700 days; every held-out point inside ± 20 %). With six months it is about 10 % low — broadly the record-length dependence the synthetic work predicts, though on real soil, where model error is *not* zero, the two cannot be separated from a single site (see the caveat below).

WATCH OUT

What it does *not* show — and this cuts against an earlier claim.

- **Ballina does not corroborate the over-confidence finding.** An earlier version of this section reported that the $\Delta\chi^2$ band missed at all three cut-offs and cited that as the strongest evidence in the guide. Scored against M0 it covers two of the three, and its width does not narrow monotonically (1.7 → 7.4 → 4.3 %). With three cut-offs on one site this test was never able to resolve a coverage question either way. **The over-confidence result now rests on the 270-run study of §25.7 alone**, and that is where it should be judged.
- **Both ranges still fail at six months.** At the 200-day cut the central estimate is about 10 % low and neither range contains the truth — and note that this cut already covers **two thirds** of the settlement, so it is *not* the below-half regime the synthetic work identifies. A band cannot rescue a mislocated centre. What put the centre in the wrong place here cannot be settled from one cut-off: real model error is a candidate the synthetic suite cannot produce, but §25.7's under-observed-layer study shows a short record can do it too on a fast-shallow-over-slow-deep profile — which is what this site is. Read it as a warning that the half-settlement rule is necessary rather than sufficient, not as a diagnosis.
- **The un-inverted laboratory parameters over-predict by 8.4 %** (+121 mm at day 974). Measured at day 974 the inversion beats them at the 400- and 700-day cuts and is beaten by them at 200 days. So back-analysis earns its place here only once the record is long enough. (Over the four held-out checkpoints of the 200-day cut, days 398, 691, 900 and 1080, with the last against the final reading carried forward, the inversion is still ahead: 115 mm against the typed baseline's 129 mm, and still ahead on the three measured days alone.)
- **Recovered parameters are not physically clean.** k_0 spans several orders of magnitude across a fairly uniform marine clay, and layers sit on the C_c/C_r cap. Use the forecast; do not carry these values to another site. The bound-hit flag exists for exactly this.

25.8.1 Two records for one embankment

The settlement of this embankment has been published twice, and the two records disagree by far more than either paper suggests:

Day	SP1–SP4 plate mean	M0 extensometer	Difference
100	633 mm	719 mm	+86 mm (+13.6 %)
200	837 mm	942 mm	+106 mm (+12.6 %)
292	942 mm	1118 mm	+176 mm (+18.6 %)
500	1119 mm	1270 mm	+151 mm (+13.5 %)
700	1264 mm	1317 mm	+53 mm (+4.2 %)
974	1431 mm	1443 mm	+12 mm (+0.9 %)

The pattern is specific and worth reading carefully. The two records disagree about *how fast* the ground settles, by as much as 18.6 % in mid-record. They agree to within 1 % about *how much* it finally settles. They are not offset by a constant; the gap opens and then closes.

That is what lateral variability across a single footprint looks like. M0 sits at one point; SP1–SP4 are plates at different positions averaged together. Where the soil profile, the drain performance or the fill thickness varies across the site, different locations reach the same final settlement along different time paths. A single instrument is therefore a measurement of *that location*, and a plate average is a measurement of nothing in particular — which is the practical argument for scoring against a defined instrument.

WATCH OUT

Stated plainly, because it affects every number above: reading error in the hand-digitised plate record cannot be excluded as a contributor, and with two records we cannot separate real spatial variability from digitisation drift. What we can say is that *a back-analysis is only ever as good as the record it is fitted to*. Fitting the plate mean instead of M0 moves this tool's day-1080 forecast by 10 mm at the 200-day cut, 157 mm at 400 days and 12 mm at 700 days — and the 400-day figure alone is larger than any parameter effect measured anywhere else in this part. Before trusting a back-analysed forecast, be sure you know which instrument produced your settlement record and what it represents.

25.9 What this tool does not model

- **Well resistance** is modelled by the forward tasks and not by the inversion. Since October 2026 a typed discharge capacity q_w is modelled in Design and Simulate on both solution routes: on the radial-only route (every layer $c_v = 0$ and every drain reaching the base) the drain is solved as a quasi-steady conduit swept to convergence against the soil; on the coupled route — taken when any layer has $c_v > 0$ or a drain stops above the base — it is solved implicitly with each time step. A typed q_w slows consolidation and leaves the ultimate settlement unchanged: at full consolidation there is no flow, so there is no head loss. Two-way discharge needs the drain to reach a draining base; a drain stopping within the clay discharges upward only. The discharge factor $D = q_w / (k_h \cdot l_m^2)$ (Mesri & Lo 1991; $D \geq 5$ negligible, a requirement modern band drains exceed by an order of magnitude — Hansbo 1997; Holtz et al. 1991; Rujikiatkamjorn & Indraratna 2007; Abuel-Naga et al. 2015) is still printed beside the drain inputs, as a screening number: it is built on k_0 and on the fastest layer over the whole drain length and overstates the effect, so read the modelled result. **The back-analysis does not model it.** In the inversion the drain wall is held at zero excess pressure, so a fitted k_0 already absorbs whatever resistance the real drain imposed on the record; carrying that profile into a Design run that also applies the same q_w counts the resistance twice and over-predicts the delay. The back-analysis page and the exported report say so whenever a q_w is present — either clear q_w for the forward run, or treat the fitted k_0 as an apparent value that already stands in for both. A project saved with a q_w typed for the screening check alone reopens with the choice to keep it (the run is the new, slower one) or clear it (the project reproduces the result it was saved with). The cases where resistance bites — very long drains (the length enters squared), organic soils where discharge capacity degrades with time (Miura & Chai 2000 measured up to fivefold reduction over 100–

300 days) and kinked drains — are the cases to type q_w for in Design, and in a back-analysis a drainage rate the fit cannot explain by smear may be well resistance the fitted k_0 has absorbed.

- **Loading history.** Multi-stage loading is supported, and treated the same way in the fit and in the forecast: every stage in the loading schedule, whether it falls inside the monitored window or after the last reading, is applied at its typed time and ramp, and each stage restarts the dissipation in the model as it does in the ground. The schedule itself is never fitted. A fill height, a stage date or a ramp duration recorded wrongly is therefore a model error, and the search will absorb it into the soil parameters it is free to move; take stage timing from the construction records, not from the shape of the settlement curve. The validation campaign ran two- and three-stage schedules throughout, with stages on both sides of the cut (§27.2).
- **Minimum usable record.** $U \approx 50\%$ is quoted as typical, not a floor: below $U \approx 40\text{--}50\%$ the magnitude family is essentially unconstrained and the band collapses to the prior; the Asaoka cross-check is itself unreliable below $U \approx 60\%$.
- **Determinism and reproducibility.** The search draws no random numbers and the final ranges are built from a fixed sampler, so the same project file on the same build and the same machine reproduces the same result: the published example reproduces to the millimetre between runs. Two limits are measured rather than assumed. The last significant figures of a line search depend on floating-point order, so another machine may differ in the last digits. And on one audit case with a short record and a flat valley of near-equal fits, two runs of the same file settled on forecasts 4 % apart (866 and 902 mm, both inside every reported range; §20.7); the cause is being traced. Treat agreement between runs as something to check, not assume: re-run, and compare the exported search logs. Design and Simulate are reproducible by construction, from the sampling seed saved in the project file and printed on the report.