Online Resource 1

Supplementary material for "Calibration of Reported Uncertainty in State-of-Charge Estimation for Lithium Iron Phosphate Cells"

Alexandre Barbosa de Lima
Faculty of Exact Sciences and Technology, Pontifical Catholic University of São Paulo (PUC-SP), São Paulo, SP, Brazil
Artificial Intelligence and Quantum Technologies Research Group, Naval War College, Rio de Janeiro, RJ, Brazil
Corresponding author: ablima@pucsp.br

Journal of Control, Automation and Electrical Systems

Contents

Table S1: truncation bounds of the prior of the initial SOC. Fig. S1: slope of the OCV curves. Fig. S2: driving profiles. Table S2: coverage of the grid filter by stratum. Table S3: amendments to the preregistration. Section S1 and Table S4: numerical validation of the exact reference and of the grid filter. Section S2: power analysis. Section S3 and Table S5: exploratory sensitivity to the hysteresis rate. Fig. S3: example trajectory. Fig. S4: coverage of the grid filter against the exact reference. Section S4: computational cost. Section S5: aggregation of the RMSE ratio. Tables S6 and S7: coverage and RMSE of all estimators in the primary and in the robustness conditions.

The preregistration, its amendments, and the archived code and results use other labels for some conditions and estimators: the voltage-sensor conditions S1, S2 and S3 of the paper are labeled S1, S3 and S4, and the Fisher estimator, the IEKF, the UKF and the grid filter are labeled E1, E2, E2u and E6.

Table S1: Truncation bounds of the prior of the initial SOC by profile and segment.

Profile	Segment	μ_0	Bounds of z_0
DST	crossing	0.88	[0.680009, 0.996751]
US06	crossing	0.88	[0.679665, 0.996751]
FUDS	crossing	0.88	[0.677496, 0.996751]
DST	plateau	0.70	[0.399506, 0.996751]
US06	plateau	0.70	[0.400016, 0.996751]
FUDS	plateau	0.70	[0.400039, 0.996751]

The bounds are the same for both OCV curves and for both prior standard deviations. With $s_0 = 1$ pp the truncation is inactive; with $s_0 = 10$ pp it retains 85.6 to 85.7% of the prior mass on the crossing segments, where the upper bound lies 1.17 standard deviations above μ_0, and 99.7% on the plateau segments.

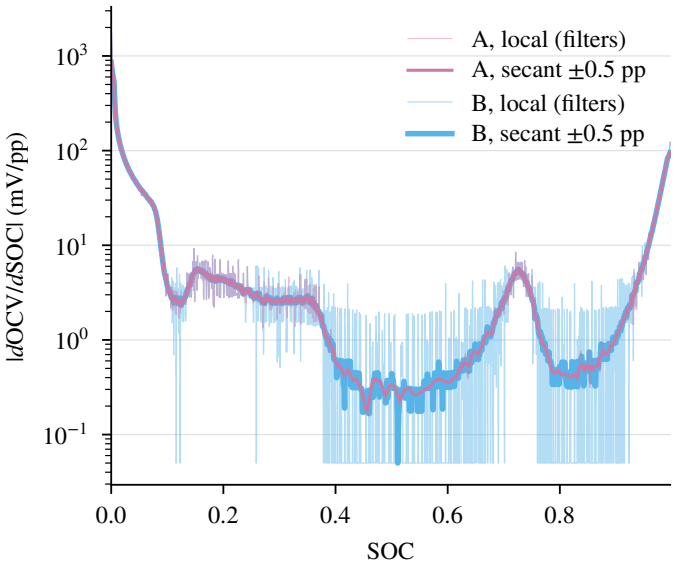

Figure S1: Local slope of the OCV curves A and B, computed as the secant over intervals of 0.025 percentage points of SOC.

Table S2: Coverage (%) of the grid filter by voltage sensor, RC process noise, and segment; 12 conditions and 600 trajectories per row (exploratory summary of the registered stratum results).

Sensor	q_{RC}	Segment	50%	80%	95%
S1	0	crossing	55.6	78.4	86.3
S1	0	plateau	60.0	84.9	93.7
S1	10^{-8}	crossing	50.3	79.6	94.9
S1	10^{-8}	plateau	50.5	80.4	95.4
S2	0	crossing	50.4	79.8	95.1
S2	0	plateau	51.8	81.4	95.0
S2	10^{-8}	crossing	50.2	80.6	96.1
S2	10^{-8}	plateau	50.7	80.4	95.5
S3	0	crossing	50.1	78.0	93.8
S3	0	plateau	47.9	76.9	92.8
S3	10^{-8}	crossing	49.2	80.7	95.2
S3	10^{-8}	plateau	51.4	79.1	94.3

Table S3: Amendments to the preregistration, in the order in which they were made, all before the confirmatory run started and before the examination of any of its outcomes.

Amendment	Content
01	Corrected the table of feasible initial SOC values to match the generator and extended the software tests.
02	Registered the sensitivity analysis of the IEKF with a limit of 50 iterations.
03–05	After a first execution stopped on a numerical overflow of the grid filter: added the numerical safeguards, the likelihood-driven refinement, and the recording of the grid resolution in every realization.
06	Replaced an earlier implementation of the exact reference after a uniform-grid evaluation of the same posterior showed it to be inaccurate.
07	Kept the grid filter unchanged after an exploratory diagnosis showed residual local deviations from the exact posterior that finer grids did not remove (Section S1).

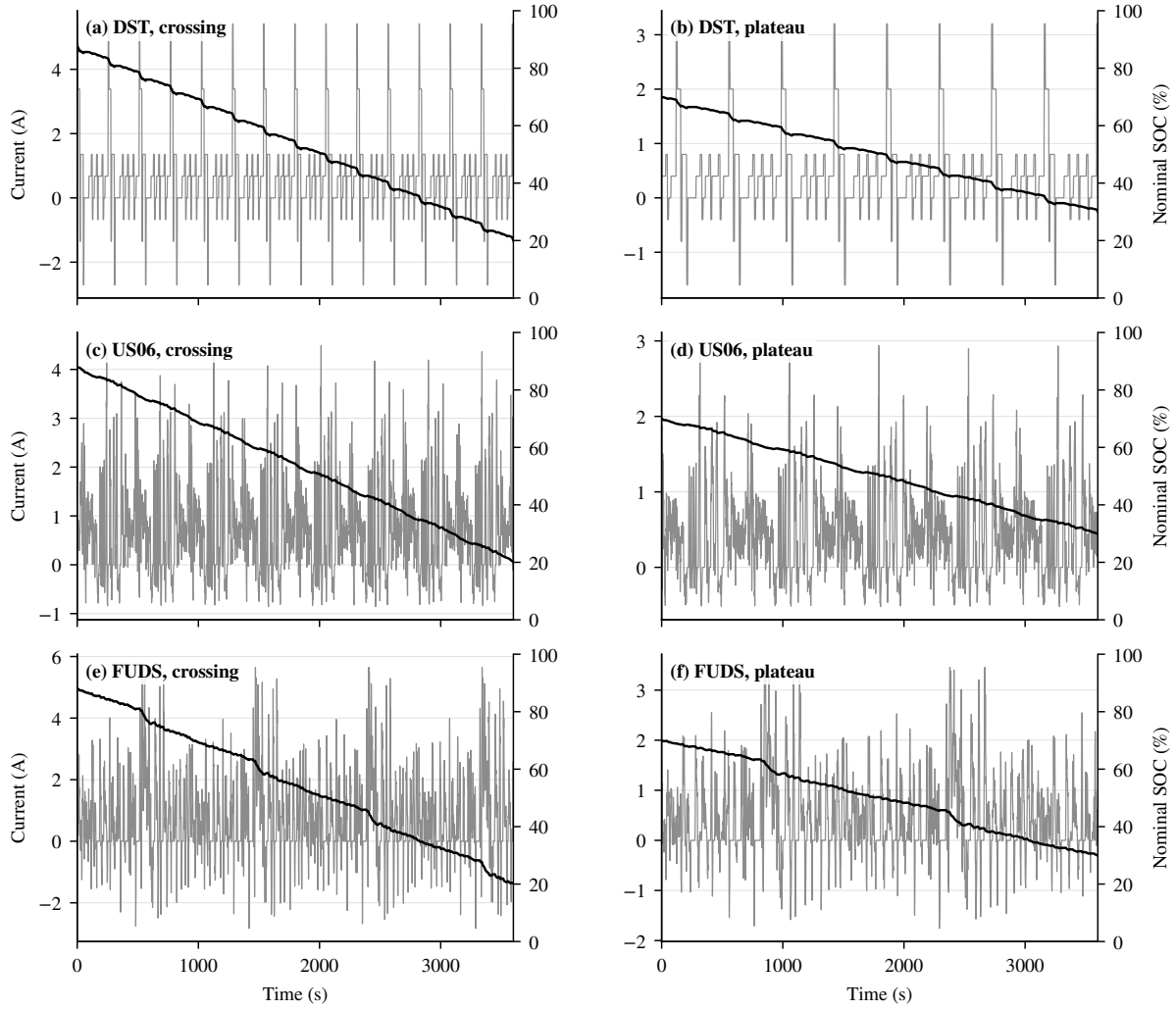

Figure S2: Current of the driving profiles after rescaling to 3600 s (grey, left axis; positive values are discharge) and nominal SOC (black, right axis) for the crossing (left) and plateau (right) segments of (a, b) DST, (c, d) US06, and (e, f) FUDS.

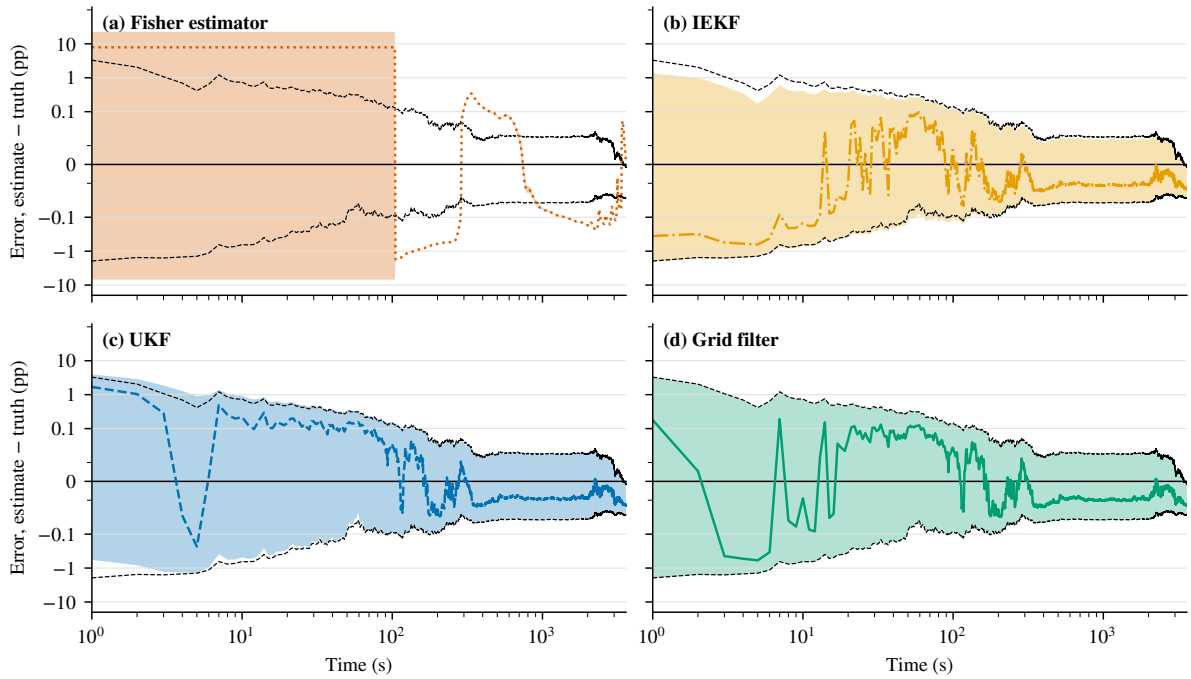

Figure S3: Estimation error (colored line) and 95% credible interval (shaded) of each estimator for one control trajectory of the main study, the realization ranked 25th of 50 by the RMSE of the grid filter in its condition (DST, crossing segment, curve A, S2, prior standard deviation of 10 percentage points, $q_{\mathrm{RC}} = 0$), both relative to the true SOC, so that the interval contains the true SOC whenever the shaded band contains zero. Dashed lines: 95% interval of the exact reference. Time on a logarithmic axis; error on a symmetric logarithmic axis, linear within ±0.01 percentage points. After 100 s, the interval of the Fisher estimator is narrower than 0.01 percentage points and is not visible.

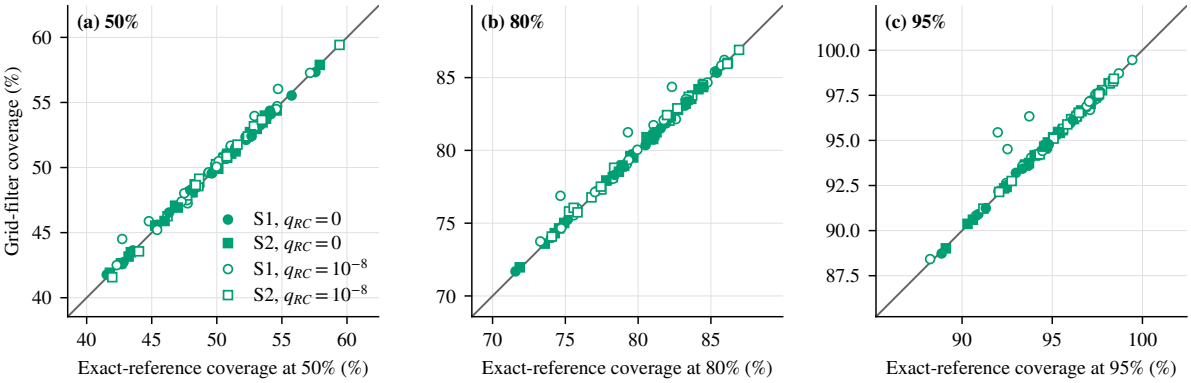

Figure S4: Coverage of the grid filter against that of the exact reference in each of the 96 control conditions at the nominal levels of the credible intervals of (a) 50%, (b) 80% and (c) 95%. Points on the identity line indicate equal coverage; their spread along the line is the sampling variability of the coverage of a condition with 50 realizations, which both methods share because they process the same data. The three points clearly above the line at 95% are the conditions that contain the five trajectories in which the exact reference did not converge; the protocol counts these trajectories as misses of the reference, which lowers its coverage in those conditions.

S1 Numerical validation of the reference and of the grid filter

Numerical rules. After each prediction, the grid filter sets the grid spacing to the predicted standard deviation of the SOC divided by $d = 40$ and interpolates the density and the conditional RC moments linearly to the new nodes. Conditioning the RC moments on the SOC increment Δz from a source node shifts their mean by $\boldsymbol{\beta}\,\Delta z$, with $\boldsymbol{\beta} = \mathbf{Q}_{\mathrm{RC},z}/Q_{zz}$ taken from the process covariance, and reduces their covariance by $\mathbf{Q}_{\mathrm{RC},z}\mathbf{Q}_{\mathrm{RC},z}^{\top}/Q_{zz}$. The update refines the grid, up to six times and with target $R \geq 40$, whenever the resolution ratio R is below 20 or fewer than 20 nodes hold 99% of the mass. The support is trimmed only at tail mass below 10^{-12}, the number of nodes is capped at 8192, and the truncated prior and the OCV domain are enforced exactly. The exact reference uses a grid over the initial SOC that covers the whole prior support and is finer near the posterior mode; its spacing is halved, up to eight times, until the 2.5, 10, 25, 75, 90 and 97.5% quantiles change by less than 10^{-3} posterior standard deviations at every time step. Amendments to the preregistration fixed these rules before the confirmatory run.

All results in this appendix are exploratory and were obtained before the confirmatory run, with seeds different from the confirmatory ones; the 120 trajectories of the simulation-based calibration check reused a fixed set of exploratory seeds.

Grid factor. With the reference available at that time, the grid factor $d = 40$ was selected because it reduced the excess of the posterior standard deviation of the grid filter over that of the reference from 0.9% (curve A) and 2.9% (curve B) at $d = 10$ to 0.02% and 0.12%. Table S4 reports this study. Because that reference was later found to be inaccurate on curve B, the final validation of the grid filter is the confirmatory comparison HA2-ref against the new reference.

Likelihood resolution. A first confirmatory execution stopped on an overflow of the grid filter in a condition with sensor S1 and a prior of 10 percentage points: a single measurement produced a likelihood narrower than the grid spacing, and the posterior collapsed onto one node. The same diagnosis showed that the reference then in use was also under-resolved in every condition with the wider prior. Amendments to the protocol introduced the likelihood-driven refinement of Section 3 of the main text and the full-support reference, and on 96 exploratory control trajectories every step of both met the resolution criterion ($R \geq 10$, at least 20 nodes holding 99% of the mass).

Validation of the reference. The new and the earlier reference disagreed on some of the 120 calibration trajectories. On the ten with the largest disagreement, all on curve B, a uniform grid over the whole prior support (up to 1.7×10^6 nodes, converged to 10^{-4} standard deviations by halving) provided an independent evaluation of the same posterior. The new reference agreed with it within 2.2×10^{-3} standard deviations, while the earlier reference deviated by up to 3.75 standard deviations. The earlier reference was therefore discarded.

Residual deviations of the grid filter. Against the uniform-grid evaluation, the grid filter deviated by up to 4.1 posterior standard deviations in short episodes (1182 samples over the ten trajectories) where the exact posterior was bimodal and strongly skewed. The deviations entered at measurement updates, were not caused by loss of support, and were not removed by forced refinement up to $R = 320$; an exploratory check of the conditional RC moments found no material error before the episodes. Over the 120 trajectories, the difference in temporal coverage between the grid filter and the reference was -0.15 [-0.56, $+0.19$], $+0.15$ [-0.06, $+0.35$] and $+0.08$ [-0.05, $+0.24$] percentage points at 50, 80 and 95%. The grid filter was kept unchanged (Section 5.1 of the main text), and the confirmatory comparison HA2-ref quantified the same effect over the 96 control conditions.

S2 Power analysis

The power analysis used the time coverage of the grid filter on 360 pilot trajectories from two earlier exploratory studies (120 with noiseless current and 240 with current noise). Within

Table S4: Grid factor study against the earlier reference (exploratory, 120 trajectories). SD ratio: mean over time of the posterior standard deviation of the grid filter divided by that of the reference. Last column: largest absolute bound of the 95% interval of the difference in coverage, in percentage points.

d	SD ratio A	SD ratio B	max. bound (pp)
10	1.0086[a]	1.0287[a]	–
20	1.0018	1.0103	1.25
40	1.0002	1.0012	0.17

[a] From a diagnostic subset of 12 trajectories.

each source stratum and nominal level, their time coverage was recentered to the nominal level, $c' = \min\{1, \max\{0, c - \bar{c} + \gamma\}\}$, which keeps the observed spread between realizations (standard deviation 0.26, 0.23 and 0.12 at $\gamma = 0.50$, 0.80 and 0.95). The clipping to $[0, 1]$ was active for 1.7, 13.3 and 26.7% of the trajectories, so the mean of c' was 49.98, 79.59 and 94.25% instead of exactly nominal. The simulated estimator was therefore slightly undercovering at 80 and 95%. With a margin symmetric about the nominal level, an exactly centered estimator would be expected to have a probability of support at least as high, but this was not simulated, and the estimates below are conditional on the simulated distribution of coverage and on the registered decision rule. Each of 1000 simulated studies drew $n = 50$ realizations per condition, applied the registered decision rule with 1000 bootstrap resamples, and was repeated with independent conditions ($\rho = 0$) and with fully correlated conditions within each group ($\rho = 1$). The probability that this simulated estimator supports H1 was at least 0.977, and the probability that an exact reference passes the reference check was 0.978 with $\rho = 1$ and 1.000 with $\rho = 0$. The script and outputs of the power analysis are available from the author on request.

S3 Exploratory sensitivity to the hysteresis rate

This analysis was not part of the preregistered design. The hysteresis rate γ of the main study was set to the nominal value 10, so the 96 control conditions were repeated with $\gamma = 5$, 10 and 20. The random draws were new, disjoint from those of the main study, and shared across the three values of γ, so that the comparisons between them are paired. For each value, the initial hysteresis state of each segment was recomputed by integrating the source cycle from $h = 1$, and the generator and all estimators used the same γ; as in the main study, the estimators received the true hysteresis state. The change of γ is not small: the initial hysteresis state of the crossing segments was between -0.40 and -0.37 with $\gamma = 10$, between 0.03 and 0.06 with $\gamma = 5$, and between -0.73 and -0.62 with $\gamma = 20$. Because of its cost, the analysis used 3 realizations per condition (288 trajectories per value of γ) instead of 50, so its intervals are wider than those of the main study, and its $\gamma = 10$ arm differs from the main results only through the random draws. The aggregation and the block bootstrap with 12 groups and 10 000 resamples were the same as in the main study.

For the three values of γ (Table S5), the nominal level lay within the 95% interval of the coverage of the exact reference at the three levels, all reference trajectories converged, and the difference between the grid filter and the reference stayed within the margin of ± 1 pp at the three levels, with a largest value of +0.15 pp [0.00, 0.31] at 80% with $\gamma = 10$. Relative to $\gamma = 10$, the paired differences in coverage were at most 0.65 pp for the reference and for the grid filter, at most 3.5 pp for the IEKF, with -3.1 pp [-5.3, -0.8] at 95% with $\gamma = 20$, and at most 2.0 pp for the UKF, with +2.0 pp [0.3, 3.8] at 95% with $\gamma = 5$. The ordering of the estimators and their qualitative behavior did not change: the IEKF and the UKF remained below the nominal 95% level, and the Fisher estimator remained near 3%. The mean SOC RMSE of the grid filter and of the UKF stayed between 0.11 and 0.13 percentage points; that of the IEKF was 1.29,

1.26 and 1.90 percentage points for $\gamma = 5$, 10 and 20.

Table S5: Exploratory sensitivity to the hysteresis rate γ over the 96 control conditions, with 3 realizations per condition: aggregate 95% coverage with 95% block-bootstrap intervals (%), grid filter minus exact reference at the three nominal levels (pp), and number of conditions in which the 95% coverage of the IEKF is below 90%.

	$\gamma = 5$	$\gamma = 10$	$\gamma = 20$
Exact reference	92.6 [88.5, 96.2]	93.1 [89.3, 96.4]	93.2 [89.5, 96.4]
Grid filter	92.7 [88.6, 96.1]	93.2 [89.4, 96.4]	93.3 [89.5, 96.5]
IEKF	58.6 [53.1, 64.0]	58.2 [52.5, 63.9]	55.1 [49.3, 60.7]
UKF	88.5 [84.1, 92.3]	86.4 [82.5, 90.0]	86.8 [82.8, 90.3]
Fisher estimator	3.1 [2.9, 3.3]	3.1 [2.9, 3.3]	3.2 [3.0, 3.3]
Grid filter minus reference, 50/80/95%	+0.00/+0.01/+0.02	+0.08/+0.15/+0.05	+0.13/+0.12/+0.06
IEKF conditions below 90%	74 of 96	75 of 96	76 of 96

S4 Computational cost

The cost benchmark of hypothesis H4 measured the grid filter in isolated processes on an Apple M3 Pro laptop: 0.57 ms per update on average and 1.07 ms at the 95th percentile. During the confirmatory run, with eight concurrent processes on the same machine, the mean cost per update was 1.4 ms for the grid filter and 0.12 ms for the UKF; this ratio describes the present Python implementation and need not hold for optimized or embedded code. The cost of the grid filter grows with its number of nodes: over all confirmatory trajectories, the mean number of nodes of a trajectory had a median of 647 and a 95th percentile of 1996, and in the worst trajectory, run with eight concurrent processes, the 95th percentile of the cost per update reached 27 ms.

S5 Aggregation of the RMSE ratio

The registered percentile interval of the mean over conditions of the log RMSE ratio of the grid filter to the IEKF, $[-1.516, -1.270]$, does not contain the point estimate, -1.523. An exploratory check traced this to a finite-sample shift of the bootstrap distribution: because the RMSE of a condition is computed before the logarithm, resamples that omit one of the few IEKF trajectories with very large errors lower its RMSE sharply; the decision on H3 does not depend on this statistic. In an exploratory sensitivity analysis, the median over conditions of the same logarithm was -1.078 $[-1.077, -0.741]$, and the mean over trajectories of the logarithm of the ratio of their RMSE was -0.281 $[-0.339, -0.220]$. All three aggregations favor the grid filter, but the typical ratio for a single trajectory, about 0.75, is much closer to one than the ratio between conditions, which the few IEKF trajectories with very large errors dominate.

S6 Coverage and accuracy of all estimators

Table S6: Coverage (%) and mean trajectory RMSE (percentage points) over the 144 primary conditions, including S3, with 7200 trajectories per estimator.

Estimator	50%	80%	95%	RMSE
Fisher estimator	1.6	2.5	3.1	2.61
IEKF	31.4	49.1	59.4	1.56
UKF	51.6	79.4	92.5	0.130
Grid filter	51.5	80.0	94.0	0.130

Table S7: Robustness conditions with fixed initial errors (216 conditions, 20 realizations each): 95% coverage (%) and mean trajectory RMSE (percentage points, in parentheses) for initial errors of 0, +10 and −10 percentage points.

Estimator	0	+10	−10
Fisher estimator	3.1 (2.44)	3.7 (3.05)	3.1 (3.05)
IEKF	69.4 (0.12)	40.5 (5.10)	42.8 (5.30)
UKF	94.3 (0.08)	81.3 (0.31)	92.3 (0.20)
Grid filter	94.4 (0.08)	94.2 (0.16)	94.2 (0.19)