

Validating predictive uncertainty in cable condition monitoring

A Chaudhary¹, A Keprate¹, Y Parvini¹ and S Riemer-Sørensen²

¹Faculty of Technology, Art and Design, Oslo Metropolitan University, Oslo, Norway

²SINTEF Digital, Norway

E-mail: apcha3338@oslomet.no, arvind.keprate@oslomet.no, yasha.parvini@oslomet.no, signe.riemer-sorensen@sintef.no

Abstract. Probabilistic models are increasingly used in industrial condition monitoring because they report uncertainty with every prediction. It is usually validated by checking coverage, the fraction of observations inside the nominal interval. This paper shows that a model can pass such a check while predicting almost nothing. Using twenty-one months of spatial impedance profiles from a sub-sea inter-array cable at an operational offshore wind farm, measured by frequency-domain reflectometry, we fit a Gaussian process with time as its only input. Coverage is close to nominal, but the model recovers an estimated five per cent of the error reduction available to a time-only predictor, and almost all of its predictive variance is a fitted noise term. A constant profile, with intervals from its own residuals, matches that coverage to within a tenth of a percentage point. From four forecast origins, coverage ranges from 0.88 to 0.93. It rises and falls with the variability of the test period, for the model and the constant profile alike. On a synthetic signal, coverage stays near nominal whether the model has skill or not. Coverage should therefore be reported with interval width against a stated usability criterion and a skill score against a matched-information baseline.

1. Introduction

Subsea cables are a significant cause of lost availability in offshore wind, and the failures and reliability of transmission and array systems have had sustained attention [1, 2]. Condition monitoring aims to catch degradation early enough for maintenance to be scheduled rather than forced. Offshore access is weather-dependent and expensive, so a monitoring system’s value lies less in a point estimate of condition than in how confidently that estimate can be acted upon. A planner does not act on a predicted value; they act on a probability.

This has made probabilistic models attractive. Methods such as Gaussian processes (GP) return a full predictive distribution rather than a single value, giving each prediction an uncertainty interval [3]. Those intervals are often validated by checking coverage, the fraction of observations falling inside the nominal interval [4]. Coverage is easy to compute and to report, and a value close to nominal is routinely offered as evidence that a model can be trusted. It is a necessary condition but a weak one: an interval can reach nominal coverage while being too wide to inform any decision, and without the model having learned anything about the process it monitors [5, 4].

Proper scoring rules are the better instrument, and we report one. The continuous ranked probability score rewards sharpness as well as calibration, so an interval widened for safety is penalised [5]. Coverage is nevertheless what condition-monitoring practice reports, because it

reads directly against a stated confidence level and needs no reference forecast. It is that practice this paper addresses.

The underlying observation is not new. A forecast that always issues the climatological distribution is calibrated but not sharp, and forecast verification has long assessed the two properties together [4]. Established tools include probability integral transform (PIT) histograms [6, 4], reliability diagrams with consistency bars [7], and skill scores built on the continuous ranked probability score (CRPS), whose decomposition separates reliability from resolution [8, 5]. This paper does not propose a new diagnostic. It shows how the known gap appears in condition monitoring, where coverage is often the only check reported. It does so on an operational asset, against a baseline given the same information as the model and an untrained model, by forecast horizon and position along the cable, with uncertainty intervals on each comparison.

Here both failures occur together. Using twenty-one months of spatial impedance profiles from an operational offshore inter-array cable, we fit a GP with time as its only input and test its intervals against what a coverage check omits: width against a stated criterion for usefulness, and skill against a baseline given the same information. Coverage is near nominal at every level. A constant profile is calibrated identically, an untrained model better, and across four forecast origins coverage follows the variability of the test period, for the model and the constant profile alike. Machine learning is widely applied to condition monitoring in offshore wind [9]; several neural architectures were examined on this dataset in earlier work and are not revisited here.

A model with time as its only input is studied because it is what a monitoring system without covariates can deploy: it needs only the indicator’s own history. Covariates such as load current or conductor temperature were not available for this asset. Section 3 shows that on this record the model’s outcome is largely fixed by the data before any fitting. That is part of the point, because a coverage check does not reveal it. To show that the argument does not rest on this one record, Section 5.5 repeats it on a synthetic signal whose predictable fraction is controlled.

2. Asset, measurement and data

The asset is a 36 kV cross-linked polyethylene (XLPE) submarine inter-array cable in an operational offshore wind farm; the site is withheld under a non-disclosure agreement. The measured span is one complete turbine-to-turbine segment of 670 m, itself a single link in a longer inter-array string, so results describe that segment and are not extrapolated to the string. The cable is three-core and this work uses one phase, so the study concerns a single signal path whose geometry is unchanged across the record.¹

Non-destructive methods for subsea cable assessment [10] include dielectric loss measurement [11], sheath current analysis [12] and reflectometry in the time or frequency domain [13, 14]. The measurements used here were acquired by broadband frequency-domain reflectometry, also called line impedance spectroscopy, on the energised cable. A swept signal is applied and the line impedance taken from the ratio of voltage to current at each step, over 750 Hz to 7.5 MHz in ten thousand uniform steps. The impedance phase is then converted into a profile along the cable, using a propagation velocity ratio to express electrical delay as distance. Each acquisition yields one profile of a scalar condition indicator, in decibels at 568 positions between the cable ends; Figure 1 shows examples.

Three elements of this chain are proprietary to the industrial partner and are not disclosed: the measurement hardware, the transform from impedance phase to the reported indicator, and the procedure that fixes the velocity ratio. Reported here are the measurement principle, the frequency range, the velocity ratio and the resulting spatial sampling. [Features along the profile](#)

¹The remaining two phases were acquired over the same period but were not licensed for this study.

mark positions where the impedance of the cable changes. The analysis does not depend on the instrument's design or on how the indicator is defined internally. It assumes that the indicator is measured consistently across the record. Repeatability could not be tested directly: only one phase was licensed, and no repeated acquisitions at a common time are available. Measurement noise and genuine short-term change are therefore not separated here. Section 3 treats both as one fast component and tests one candidate measurement effect.

Bandwidth limits the finest spatial detail the measurement can resolve. Range resolution is the propagation velocity divided by twice the bandwidth, of the order of ten metres for the velocity ratio applied here. The profile is reported on a grid roughly ten times finer, so neighbouring samples are not independent observations and the cable supports on the order of sixty independent resolution cells. This directly influences the analysis of output dimension in Section 5.

Acquisition was scheduled every eighteen hours and ran from April 2023 to December 2024, giving 810 profiles from 852 scheduled slots. The shortfall is instrument downtime rather than irregular sampling, since almost every gap is an integer multiple of the interval. Five interruptions exceeded forty-eight hours, the longest a little under twenty days.

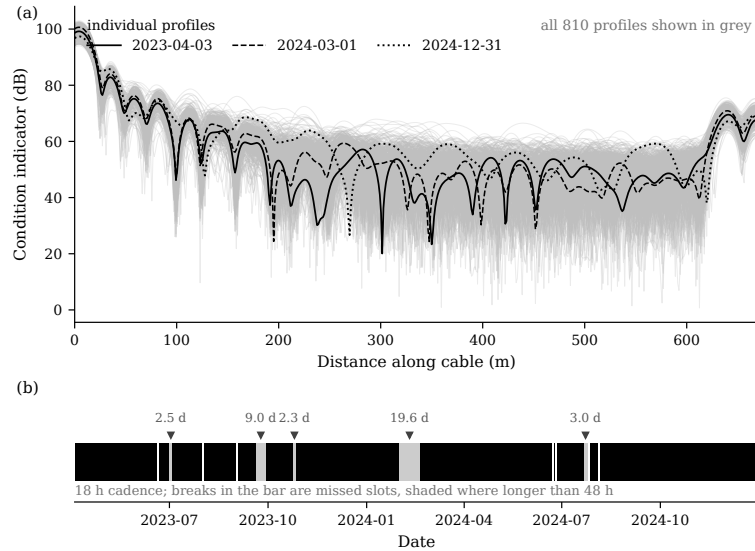


Figure 1. The measurement record. (a) All 810 profiles in grey, with three from the start, middle and end of the record in black. (b) Acquisition history: the bar is continuous where measurements arrived on schedule, and breaks are missed slots, labelled where longer than 48 h.

3. How much of the signal is predictable from time

Write $y(x, t)$ for the condition indicator at position x along the cable at time t , so one acquisition is a profile $y(\cdot, t)$ sampled at 568 positions. The model examined here treats each position independently and takes time as its only input: it fits one function $f_x(t)$ per position and returns a predictive mean and variance for $y(x, t)$ at any future t . It can therefore reproduce only variation that persists from one measurement to the next, and how large that part is, is a property of the data rather than of any model.

Profiles were compared pairwise against their separation in time τ , through the standardised structure function

$$S(\tau) = \frac{\langle [y(x, t + \tau) - y(x, t)]^2 \rangle_{x,t}}{\langle \text{Var}_t [y(x, t)] \rangle_x}, \quad (1)$$

where $\langle \cdot \rangle$ averages over the quantities in its subscript. Two unrelated measurements of the same cable give $S = 2$, the reference drawn in Figure 2. The dissimilarity approaches that value within about a day of separation and stays there. It does not drop again at weekly, monthly or annual separations, so no seasonal return of similarity is detectable in a record of this length.

Splitting the per-point variance into a part that changes faster than the sampling interval and a part that changes more slowly puts 94.4% in the fast part. Only the remaining 5.6% is slow enough for a time-input model to follow. That slow part is small but real: permuting profiles against their timestamps destroys the timing while preserving spatial structure, and the observed slow fraction lies far outside the range returned by 500 such permutations ($z = 6.1$, $p = 0.002$).

The fast part contains genuine short-term change, measurement noise, and apparent change introduced by the measurement. One candidate for the last is a temperature-dependent propagation velocity, which would stretch the profile and move narrow features, such as the notches in Figure 1, between acquisitions. To test this, each profile was stretched by the factor within $\pm 2\%$ that best matched the mean profile, and the variance split repeated. The fitted stretch has a standard deviation of 0.42% (2.7 m at the far end) and is uncorrelated between consecutive acquisitions ($r = 0.04$), whereas a temperature effect would change slowly. Alignment removes 2.4% of the fast variance, and the fast share is 94.8% before and 94.9% after. A uniform velocity change is therefore not a major contributor. Local shifts are not excluded, so 94.4% is an upper bound on the share of genuine condition change that a time-only model cannot follow.

The predictable fraction is also distributed differently along the cable from the variability as a whole. Total variability changes by more than a factor of two between sections, while the slow part stays approximately flat. The spatial pattern therefore sits almost entirely in the part a time-only model cannot forecast; its physical cause is outside the scope of this paper. Section 4 makes the resulting bound explicit and uses it as the reference for the results reported here.

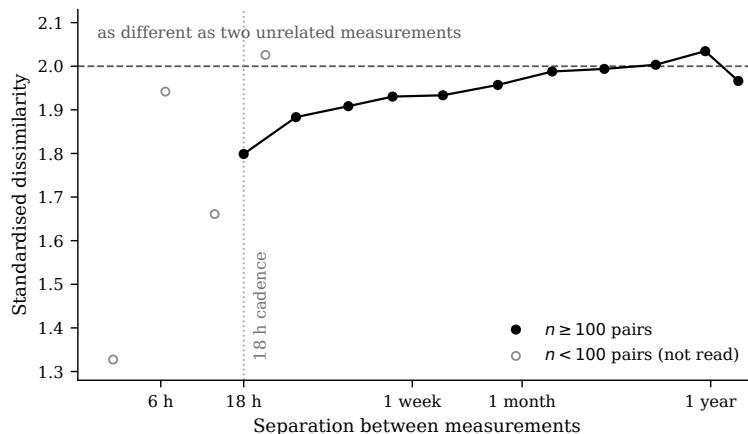


Figure 2. Dissimilarity between pairs of profiles against their separation in time, standardised by equation (1) so that two is the value expected for unrelated measurements. Open circles are bins of fewer than 100 pairs, not used in the analysis.

4. Method

4.1. Split and forecast origin

The record was split in time, not at random, with the cut at a fixed fraction of the elapsed days rather than of the number of measurements. Because the instrument missed slots unevenly, the two differ: the cut falls at 80% of the timeline but leaves just under 79% of the measurements for training. That difference is downtime, not design.

Predictions are issued from a single forecast origin at the end of the training window, and the test window is the remainder of the record. Lead times run from under a day to about four months. All five interruptions longer than 48 hours fall inside the training window, so no lead-time group in [Section 5](#) is thinned by an outage.

With one origin, lead time and calendar period move together: the longest lead times are also the last weeks of the record. To separate them, the model was refitted from three earlier origins, at 50, 60 and 70% of the timeline, each followed by a test window of the same length [15]. The headline origin was refitted alongside as a control, and had to reproduce the recorded predictions exactly. These fits were added in revision and are reported in [Section 5.3](#).

4.2. Ceiling and floor

Two bounds follow from the data and limit any model fitted to it. The ceiling is the variance a time-only predictor cannot reach: the part of the signal that changes faster than the interval between measurements. It is estimated from the mean squared difference between consecutive profiles, and recomputed on the test window rather than carried over from the record, which is less variable. The estimator assumes the fast part is uncorrelated with itself at the sampling interval; [Section 3](#) shows this is very nearly but not exactly true, so the true ceiling is a little higher and every margin here is slightly optimistic. [The ceiling is an estimate under that assumption, not a universal limit.](#) Rapid decorrelation does not by itself rule out every time-only predictor; one that exploited the small residual correlation at the sampling interval, or a process that changed after the test window, could do better. Its sampling uncertainty is also large, and [Section 5](#) carries it into every margin. Every accuracy figure in [Section 5](#) is evaluated against it.

The floor is the error introduced by reducing the profile to a lower output dimension and reconstructing it. No reduction is applied in the main analysis, so the floor is zero by construction.

4.3. Baselines and the two regimes

The model takes time as its only input, so a baseline allowed to see recent measurements is not competing on equal terms. Baselines are therefore reported in two regimes. Write t_0 for the forecast origin and $t_1, \dots, t_m$ for the training times, all at or before t_0 . Regime A gives a baseline what the model is given, everything up to t_0 and nothing after it:

$$\text{persistence : } \hat{y}(x, t) = y(x, t_m), \quad (2)$$

$$\text{mean training profile : } \hat{y}(x, t) = \bar{y}(x) = \frac{1}{m} \sum_{i=1}^m y(x, t_i), \quad (3)$$

$$\text{linear trend : } \hat{y}(x, t) = a_x + b_x t, \quad (4)$$

with a_x and b_x fitted by least squares on the training window. All three are frozen at t_0 . Regime B is rolling one-step persistence, $\hat{y}(x, t_j) = y(x, t_{j-1})$, which carries the most recent measurement forward and so uses test-window observations the model never sees. It shows what the time-only formulation costs and is never treated as a like-for-like comparison.

Table 1 gives both regimes with the ceiling and floor rows. Its error measures average over the test times and all 568 positions,

$$\text{MSE} = \langle [\hat{y} - y]^2 \rangle_{x,t}, \quad \text{RMSE} = \sqrt{\text{MSE}}, \quad \text{MAE} = \langle |\hat{y} - y| \rangle_{x,t}, \quad (5)$$

and skill is the fractional reduction in mean squared error against the mean training profile of equation (3),

$$SS = 1 - \frac{MSE}{MSE_{\bar{y}}}, \quad (6)$$

so that baseline scores zero and a negative score means worse than a constant profile. The gap between that baseline and the ceiling is the entire error reduction available to a time-only predictor: about 7% of the baseline’s own error, placed above zero by a paired bootstrap. [That bootstrap resamples consecutive pairs of whole test profiles. All positions move together, so the correlation along the cable is kept, and pairs are treated as independent in time, which the rapid decorrelation in Section 3 supports.](#)

Table 1. Baselines on the test window, with the ceiling and floor reference rows. Skill is measured against the mean training profile. Regime A uses only information available at the forecast origin; regime B uses test-window observations the model does not have, and is a reference point rather than a comparison.

Regime	Predictor	MSE (dB ²)	RMSE (dB)	MAE (dB)	Skill
—	Ceiling, time-only	27.5356	5.2474	—	—
A	Persistence	65.6508	8.1025	6.1334	−1.2174
A	Mean training profile	29.6069	5.4412	3.9712	0.0000
A	Linear trend	29.4988	5.4313	3.9678	0.0037
B	Rolling persistence	55.0081	7.4167	5.4000	−0.8579
—	Floor, reduction cost	0.0000	0.0000	0.0000	—

4.4. Model

One exact Gaussian process is fitted per cable position, with time as the only input. The 568 processes are fitted as a batch but stay independent, so no correlation between positions is modelled. Each has a constant mean, a Matérn-3/2 covariance function and a Gaussian likelihood. Predictive variances are computed exactly rather than by a low-rank approximation, because the variance is the quantity under examination.

The covariance function was fixed in advance rather than chosen on held-out performance, which would have made the comparison a tuning exercise; [Section 3](#) establishes a rough, fast-decorrelating signal, for which a Matérn-3/2 prior is appropriate. No periodic component is included: a periodicity test on the training window found the best-fitting cycle explained less variance than the same test on date-shuffled data. Scaling is fitted per position on the training window alone. Training runs for a fixed number of iterations rather than until the objective stops falling, because where the signal variance is negligible the timescale is unidentifiable and the likelihood keeps improving without changing anything reported.

[The model is implemented in GPyTorch \[16\] in double precision. The exact marginal likelihood is maximised with Adam, learning rate 0.05, for 1500 iterations, from a timescale of 20 days and the library’s default signal and noise variances \(0.69 in standardised units\). Positivity is enforced by a softplus transform, with a floor of \$10^{-4}\$ on the noise variance; no priors are placed. Fitted timescales have a median of 17.9 days \(10th to 90th percentile 2.0 to 46.8\), and the median signal and noise variances are 0.032 and 0.966. Continuing one fit to 6000 iterations changed the mean 95% width by 0.014% and coverage by 0.0002. The untrained model of Section 5 has the same structure and initial values, without optimisation.](#)

4.5. Metrics and pre-registered thresholds

Coverage is reported at five nominal levels in Table 2 and Figure 3, each beside the mean interval width and a skill score against the matched-information baseline of equation (3). Width is given in dB and as a fraction of the training-window dynamic range at each position, since that is what is available when a forecast is issued. The continuous ranked probability score and the negative log predictive density are also reported. Coverage is evaluated marginally at each position, ignoring correlation between positions.

Every headline figure carries a 95% interval from a spatio-temporal block bootstrap [17]. The resampling unit is a block of 10 adjacent positions (11.8 m, about one resolution cell) crossed with a block of 5 consecutive test profiles (about 3.75 days, several decorrelation times), drawn independently with replacement 2000 times. The ceiling is recomputed in each resample, differences between models are taken on the same resample, and lead-time groups are resampled within themselves. The cable thus contributes about 57 spatial units, not 568.

Two thresholds were fixed in the original analysis plan. Coverage is acceptable within five percentage points of nominal, and an interval is usable if its width is below half the training-window dynamic range at that position. Neither was revised once the results were seen. One is met and one is not, and both are reported below.

The order of events was as follows. (i) The thresholds were set. (ii) An exploratory reproduction of the earlier thesis pipeline was run and its test coverage and error were seen; its split differed, but its final 15% of measurements lies inside the present test window, so these data were not unseen. (iii) On 4 September 2026 the framing was revised in a written plan, keeping the thresholds and pinning the definition of dynamic range. (iv) The confirmatory analysis of Sections 3 to 5 was run; decisions made during it are listed in Section 7. (v) After review, the bootstrap, the additional origins, the synthetic case and the alignment test were added. These are exploratory.

The width criterion is illustrative. No defect-scale change in the indicator is known for this asset, so it could not be tied to a maintenance decision. It is also strict: for 639 Gaussian samples, the training-window size, a calibrated 95% interval spans on average 0.63 of their range, and less than half in 0.4% of draws.

5. Results

5.1. Coverage and skill

Coverage falls inside the pre-registered band at all five nominal levels; Table 2 and Figure 3 give it. Read on its own, that is the outcome a validation exercise sets out to obtain. Read beside the other two columns, it certifies nothing. At 95% the block-bootstrap interval is 0.909 to 0.940; at the other levels the intervals are of similar width, ± 0.01 to ± 0.03 .

The mean training profile, given intervals from the spread of its own training residuals, matches the fitted model's coverage at every level to within a tenth of a percentage point: on this measure a constant profile and a fitted Gaussian process cannot be told apart. The interval for the difference contains zero at every level. The untrained model, which has default noise, a hand-set timescale and no fitting at all, is closer to nominal at the 95% level, by 2.7 percentage points (2.0–3.4), and its error is indistinguishable from the fitted model's (difference in MSE -0.08 dB², -0.46 to 0.21). It is wider at every level and overshoots the tolerance band at 50%. Fitting by maximum likelihood made calibration worse, by tightening intervals that were already a little too narrow.

Accuracy explains why coverage is uninformative here. The fitted model captures about 5% of the margin between the mean baseline and the ceiling, with an interval of 2 to 31% that mostly reflects uncertainty in the ceiling itself. Its skill is 0.0038 (0.0012–0.0079): above

zero, but small. On the CRPS the fitted model is better than the mean baseline by 0.005 dB (0.001–0.011), a CRPS skill of 0.002, and the difference in NLPD is not distinguishable from zero. The likelihood noise term supplies a median 97% of the predictive variance, and a mean of 92%. Interval width scales with the standard deviation, not the variance, so removing the fitted function’s contribution would narrow the 95% interval by a median 1.6% and a mean 4.5%. Near-nominal coverage follows from estimating the noise level correctly, and requires little predictive skill.

The mechanism is not specific to this dataset, but it needs stated conditions. If the residuals of the mean function are Gaussian, with the same variance in training and test, a maximum-likelihood noise variance approaches their mean square, and coverage approaches nominal however little the mean function contributes. Matching the variance alone is not enough; the shape and stability of the residuals also matter, and both hold only approximately here. The standardised test residuals have an excess kurtosis of 4.2, so 3.0% of observations fall outside the 99% interval rather than 1%, which is why coverage at 99% is 0.970. Section 5.2 shows a change in variance over the first 190 m. Where the conditions hold, a coverage check tests the variance estimator, not the predictor.

The second pre-registered threshold is not met. Mean interval width at the 95% level is 0.5013 of the training-window dynamic range against a criterion of below 0.50, and 250 of the 568 positions pass. The threshold is reported as failed, as registered, but the excess is within sampling uncertainty: the block bootstrap gives 0.486 to 0.516. The result is also sensitive to the criterion: 136 positions pass at 0.45 of the range, 396 at 0.55 and 519 at 0.60. It is met at 250 positions at all only because the observed ranges are wider than Gaussian, a median 7.6 training standard deviations against 6.2 expected (Section 4).

Table 2. Coverage and mean interval width (dB) at five nominal levels, with the accuracy block beneath. Skill is against the mean training profile and the ceiling for this test window is 27.5356 dB². The untrained model has default noise and a hand-set timescale, and was not fitted.

Nominal	Fitted GP		Mean baseline		Untrained GP	
	Coverage	Width	Coverage	Width	Coverage	Width
0.50	0.5001	6.334	0.5017	6.355	0.5611	7.296
0.80	0.7808	12.035	0.7818	12.075	0.8385	13.862
0.90	0.8756	15.447	0.8763	15.497	0.9167	17.792
0.95	0.9260	18.407	0.9267	18.466	0.9531	21.201
0.99	0.9701	24.190	0.9705	24.269	0.9798	27.862
MSE (dB ²)	29.4936		29.6069		29.4177	
MAE (dB)	3.9661		3.9712		3.9526	
CRPS (dB)	2.8459		2.8510		2.8515	
NLPD	3.0241		3.0255		3.0149	
Skill	0.0038		0.0000		0.0064	

5.2. Along the cable

Interval width reproduces the training-window standard deviation at each position almost exactly, with a rank correlation of 0.9998; its rank correlation with skill is 0.0065. Width follows the unpredictable part of the variance more closely than the predictable part, 0.79 against 0.32, and is effectively uncorrelated with the fitted timescale, at 0.10. Nothing the model learned

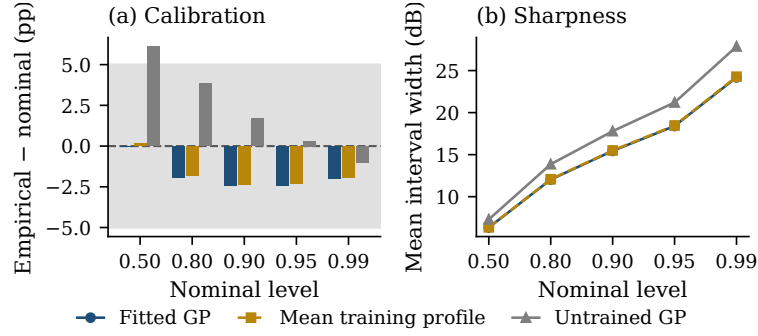


Figure 3. Calibration and sharpness at the five nominal levels of Table 2. (a) Empirical coverage minus nominal, in percentage points, against the pre-registered ± 5 pp band; the fitted process and the mean training profile are indistinguishable. (b) The mean interval width each level costs. The untrained model reaches nominal coverage at 95% by being wider everywhere.

about time influences how wide its intervals are.

Figure 4 shows the consequence. Coverage is not uniform: over roughly the first 190 m it sits below the tolerance band at most positions, and beyond that lies inside it almost everywhere. Pooled coverage passes, but it averages over a cable one part of which does not. Coverage at 95% is 0.891 (0.860–0.919) over the first 190 m and 0.940 (0.927–0.950) beyond. There the test-window variance is 57% above its training value, against 9% elsewhere, and the residual excess kurtosis is 5.6, against 2.0. Under-coverage is consistent with a change in variability that a noise term fitted to training cannot anticipate. Width exceeds the usability criterion at 318 of the 568 positions (bootstrap interval for the share passing, 36 to 52%), concentrated where the unpredictable variance in panel (d) is largest, so the intervals are widest exactly where they are least usable. Skill shows no spatial structure, because there is no skill for a pattern to appear in.

5.3. Forecast horizon

Table 3 groups the test window by lead time. Width is almost constant across the groups, as expected once Section 3 has established that correlation is lost within about a day. Coverage, however, is highest in the longest group, at 91 to 130 days ahead, and that is the only group in which skill is negative. Neither difference is resolved by the block bootstrap: the coverage intervals overlap, and skill in the longest group, -0.0001 (-0.0019 to 0.0008), is indistinguishable from zero. The two shortest groups hold 9 and 10 profiles, so their intervals are indicative only.

The model reverts to the mean baseline as lead time grows. The gap between the two predictions passes below 0.01 of a training standard deviation at roughly 85 days, so across the final third of the test window the fitted process is numerically indistinguishable from a constant profile. But the longest group is also the last calendar period of the record, and a quieter one, at 26.5 dB^2 against about 31 dB^2 in the first three groups. The additional origins separate the two explanations.

Table 4 gives the four origins. The refit at the headline origin reproduced the recorded predictions exactly. At every origin the fitted model’s coverage is within 0.12 percentage points of the mean baseline’s, and its skill is indistinguishable from zero except at the headline origin. Coverage itself is not stable. At 95% it ranges from 0.876 to 0.926, and at the 60% origin it falls outside the pre-registered tolerance. The four origins rank in exactly the order of how much more variable the test window was than the training window, from 1.14 to 1.34 times. The same holds within origins. Across the 22 lead-time groups of the four origins, coverage follows

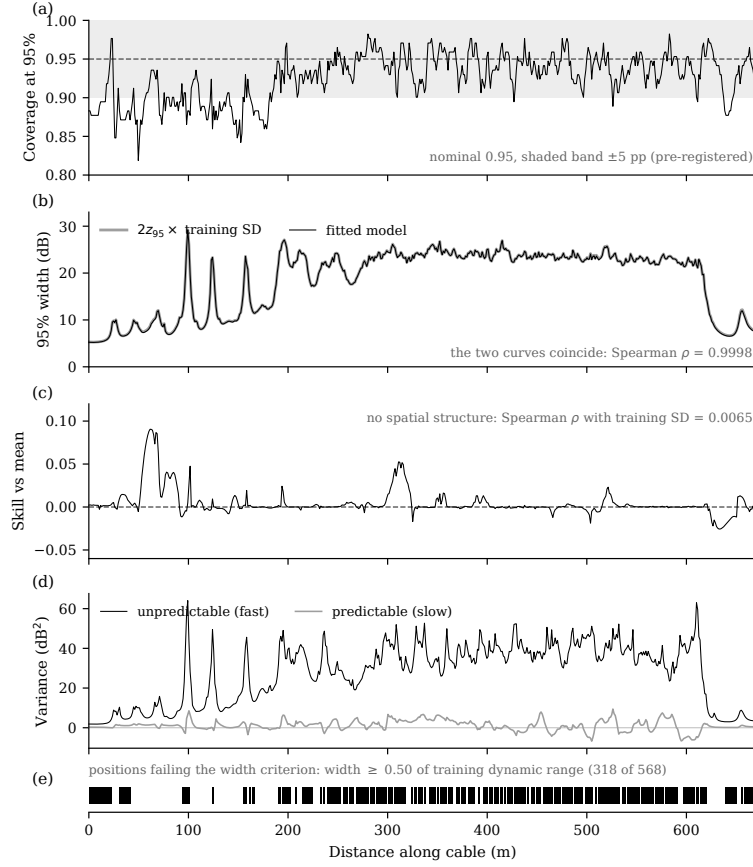


Figure 4. Coverage, sharpness and skill along the cable. (a) Empirical coverage at 95%. (b) Mean 95% interval width, with the width implied by the training-window standard deviation alone in grey; the two coincide. (c) Skill against the mean training profile. (d) Test-window variance split into the fast and slow parts; negative slow values, shown unclipped, mark a predictable part indistinguishable from zero. (e) Positions failing the pre-registered width criterion.

the ratio of the group's test variance to the training variance (Spearman $\rho = -0.88$), and not lead time ($\rho = -0.19$, not significant). The longest group has the highest coverage only at the headline origin; at the other three it ranks fourth. In every group, the constant baseline's coverage is within 0.6 points of the model's.

The high coverage at long lead times in Table 3 is therefore a property of a quiet calendar period, not of the horizon. What remains is a sharper form of the paper's point. Coverage passes or fails with the stability of the noise level between training and use, and a constant profile passes or fails with it. A coverage check measures the test period, not the model.

5.4. Dimensionality

Earlier work on this asset reduced the profile to fewer spatial blocks before modelling and reported a spatial pattern in performance. How much of such a pattern survives the reduction was answered on the training window before any model was fitted: reducing to 71 blocks costs more achievable error than the whole predictable signal is worth, so a perfect model at that dimension cannot beat a constant profile.

A confirming fit bears this out and sharpens it. Skill against the mean baseline is -0.0483 . The excess error over that baseline was predicted in advance as $+0.72 \text{ dB}^2$ and observed as $+1.43 \text{ dB}^2$, about twice as large. The direction was right and the magnitude was not: the

Table 3. Coverage, interval width and skill by forecast lead time, with 95% block-bootstrap intervals and the test-window variance in each group. Skill is measured against the mean training profile. The longest group, in bold, has the highest coverage and the only negative skill; neither difference is resolved by the intervals, and Table 4 shows the high coverage is specific to this origin.

Lead (days)	n	Coverage at 95%	Width (dB)	Skill	Test var. (dB ²)
0–7	10	0.909 (0.830–0.924)	18.314	+0.0114 (−0.0003–0.0313)	30.8
7–14	9	0.881 (0.831–0.908)	18.364	+0.0129 (0.0022–0.0268)	32.4
14–28	19	0.915 (0.882–0.958)	18.389	+0.0074 (−0.0019–0.0155)	31.4
28–56	37	0.929 (0.909–0.951)	18.411	+0.0051 (0.0005–0.0102)	26.2
56–91	46	0.921 (0.878–0.953)	18.421	+0.0010 (−0.0011–0.0031)	29.7
91–130	50	0.944 (0.930–0.962)	18.423	− 0.0001 (−0.0019–0.0008)	26.5

Table 4. The four forecast origins. Each test window is 20% of the timeline. n is the number of training and test profiles. The variance ratio is the mean test-window variance over the mean training-window variance. Baseline is the mean training profile’s coverage at 95%. Intervals are 95% block-bootstrap intervals. Width is the mean 95% interval width as a fraction of the training dynamic range. The 80% origin is the headline analysis.

Origin	n	Var. ratio	Coverage at 95%	Baseline	Skill	Width
50%	390 / 168	1.19	0.908 (0.877–0.934)	0.908	−0.0016 (−0.0051–0.0008)	0.552
60%	472 / 167	1.34	0.876 (0.839–0.904)	0.877	−0.0004 (−0.0026–0.0015)	0.530
70%	558 / 167	1.29	0.900 (0.878–0.920)	0.900	+0.0012 (−0.0045–0.0057)	0.521
80%	639 / 171	1.14	0.926 (0.909–0.940)	0.927	+0.0038 (0.0012–0.0079)	0.501

training-window calculation treated the cost of reduction as additive, whereas averaging also smooths part of the fast component.

The reduced model also gives sharper intervals, 16.9 dB against 18.4 dB, and fails calibration at 0.8794 against a 95% nominal level, outside the pre-registered tolerance. Its predictive variance carries no term for the error the reduction introduces, since every position inherits its block’s variance. This is a second instance of the paper’s central point: sharper intervals a coverage check would have rejected, alongside calibrated intervals it accepts.

Figure 5 places the ladder against the range resolution implied by the acquisition parameters, about 11 m. Every rung, including the one that destroys the model, covers less ground than the measurement can resolve: averaging within a single resolution cell already costs more than the whole predictable signal is worth. Two-thirds of that cost, 4.20 of 6.25 dB² on the training window, changes between acquisitions rather than coming from the fixed shape of the mean profile. This is consistent with narrow features, such as the notches in Figure 1, whose fine structure varies between acquisitions. Section 3 found that a uniform stretch of the profile explains little of the fast variance; local shifts of single features were not tested and remain possible.

5.5. A synthetic signal with a controlled predictable fraction

The analysis was repeated on synthetic profiles at the same acquisition times, with the same split, model and fitting. Each of 128 independent positions is a slow Matérn-3/2 component with a 20-day timescale plus independent fast noise. The slow share of the variance, p , is 0.056 (this cable’s estimate), 0.25, 0.5 or 0.9. The noise is Gaussian, or Student- t with four degrees of freedom for heavy tails; in one further case its variance rises by 10% in the test window only.

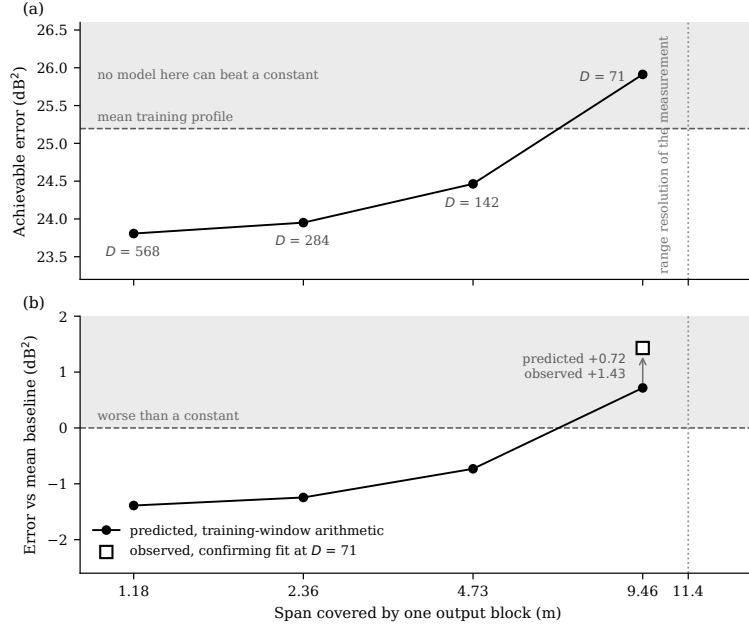


Figure 5. What spatial reduction costs, against the span each output block covers, so that the instrument’s range resolution shares the axis. (a) Achievable error on the 568-point grid, computed on the training window before any model was fitted; the dashed line is the mean training profile, and in the shaded region no model at that dimension can beat a constant. (b) The same as error relative to the mean baseline, with the confirming fit at $D = 71$ marked. Every rung lies inside one resolution cell.

Table 5 gives the result. Coverage at 95% lies between 0.926 and 0.947 in every case, and the mean baseline’s is within 0.7 percentage points of the fitted model’s in each. This holds whether the model has almost no skill, 0.0005 at $p = 0.056$ with heavy tails, or real skill, 0.110 at $p = 0.9$. Coverage therefore does not separate a model that predicts from one that does not, even when the predictable fraction is sixteen times that of this cable. Skill stays modest even at $p = 0.9$, because one origin must forecast up to four months ahead of a process whose timescale is 20 days.

The two departures of Section 5 behave as predicted. With heavy-tailed noise, coverage at 99% falls to 0.976–0.982 while coverage at 50% rises to as much as 0.584: under-coverage in the tails, as on the cable. A 10% rise in noise variance in the test window lowers coverage at 95% from 0.946 to 0.933, for the model and the baseline alike. The width criterion behaves as Section 4 predicts. Width is 0.63 to 0.69 of the training range with Gaussian noise, so every Gaussian case fails it. With heavy-tailed noise at $p \leq 0.5$ it is 0.38 to 0.46, because the tails widen the range, so those cases pass. Whether the criterion is met depends on the shape of the noise, not on skill.

6. Discussion

An uncalibrated interval is dangerous because a planner will act on it as though it were calibrated. A calibrated but uninformative interval is dangerous for a less obvious reason: it passes the check most often run, and so gains a credibility it has not earned. Section 5 is an instance of the second case, and nothing about it is exotic. Wherever most of a predictive variance is a noise term estimated from the data, near-nominal coverage follows almost arithmetically, provided the residuals are close to Gaussian and their spread does not change between training and

Table 5. Synthetic profiles: 128 positions per case at the real acquisition times, with the same split, model and fitting as the cable. p is the true share of variance in the slow component; $\hat{p}$ is its estimate from the training window by the method of Section 3. Skill is against the mean training profile. Width is the mean 95% interval width as a fraction of the training dynamic range.

p	Noise	$\hat{p}$	Coverage, fitted			Baseline		
			50%	95%	99%	95%	Skill	Width
0.056	Gaussian	0.053	0.494	0.946	0.989	0.946	0.0025	0.635
0.25	Gaussian	0.221	0.492	0.945	0.989	0.945	0.0150	0.632
0.5	Gaussian	0.466	0.480	0.938	0.986	0.937	0.0351	0.640
0.9	Gaussian	0.893	0.477	0.937	0.985	0.931	0.1103	0.693
0.056	t_4	0.047	0.584	0.947	0.976	0.947	0.0005	0.379
0.25	t_4	0.223	0.543	0.945	0.977	0.946	0.0069	0.417
0.5	t_4	0.485	0.515	0.946	0.982	0.947	0.0274	0.456
0.9	t_4	0.891	0.478	0.926	0.978	0.928	0.0554	0.644
0.056	Gaussian, test +10%	0.053	0.477	0.933	0.984	0.934	0.0026	0.637

use. When the spread does change, coverage changes with it, and the four origins show that a constant profile then passes or fails alongside the model. A coverage check on one test period says more about that period than about the model.

Coverage should therefore not be reported on its own. Three cheap quantities together prevent the error: the empirical coverage; the interval width against a criterion for what width would be useful for the decision at hand; and a skill score against a baseline given the information the model receives. Omit one and a model with no predictive value can be presented as validated.

The width criterion is the one most often left implicit. It cannot be derived from the model, only from the decision the model supports. Here it was half the training range, fixed in advance. It is illustrative: no defect-scale change in the indicator is known for this asset, and it is stricter than a calibrated Gaussian interval usually meets. A criterion tied to the smallest change a planner must detect before scheduling an intervention would make the check decisive. Until one exists, width should be reported against a stated reference, with its sensitivity. Matched information matters just as much: a baseline given different information tells the reader about the information, not about the model. Rolling persistence sees the previous measurement and does far worse here than a constant profile, so presented as a comparison it would have made the model appear strong.

7. Limitations and disclosure

The findings rest on one asset, one phase and one turbine-to-turbine segment, at one acquisition cadence. Nothing here shows that Gaussian processes are poorly calibrated in general, nor that this signal is unpredictable by any method. What is shown is that on this signal, at this cadence, with time as the only input, coverage is achieved without skill, and that a coverage check alone would not have revealed it. A study that argues for fuller reporting must report fully on itself.

1. The framing followed an exploratory reproduction run whose test metrics had been seen, on a split whose test set lies inside the test window used here (Section 4). The two pre-registered thresholds were not revised afterwards; one is met and one is not.
2. The criterion for choosing the output dimension was revised twice and the convergence criterion once, each time after its output had been seen. Neither is pre-registered, and the full history is retained.

3. An initial comparison of four covariance functions was run in which every candidate failed to converge. Its ranking was discarded and the covariance function was then fixed a priori. A defect in the first stopping rule, which halted training early, was found and corrected before any reported fit.
4. The split is 80% of the timeline but 78.9% of the measurements, because of instrument downtime.
5. The variance estimator's independence assumption is nearly but not exactly met, so all margins are slightly optimistic (Section 4).
6. The confirming run used a different patch release of the interpreter. All package versions were unchanged, and every regenerated artefact matched value for value.
7. The confirming fit at reduced dimension was planned before it ran and did not revise the dimension decision. Its pass condition was first a test on the sign of skill, which it passed while the predicted magnitude was wrong by about a factor of two; the magnitude comparison was added afterwards so that this is reported.
8. Every headline figure is a five-seed mean. The model has no randomly initialised component, so the spread across seeds is exactly zero, measured rather than assumed.
9. The block bootstrap, the additional forecast origins, the synthetic case, the alignment test and the width-criterion sensitivity were added in response to review, after every result in the reviewed manuscript had been seen. They are exploratory. Their design was fixed in code before each was run and was not changed afterwards.

Two planned analyses were not carried out: a study of temporal averaging, and confirming fits at the two intermediate output dimensions. Forecasts from multiple origins, planned and first omitted, were added in revision. A sensitivity check across covariance functions was also not run, because with the function fixed a priori it could only have acted as selection. No covariates, such as load current or temperature, were available, so the time-only model could not be compared with one that uses them.

Two checks are the natural next step. Repeatability is not established here; simultaneous measurements on the other phases, if licensed, or repeated acquisitions minutes apart would bound it and separate measurement noise from change in the asset. If the fast component is largely measurement noise rather than the asset, averaging blocks of N consecutive acquisitions should suppress it in proportion to $1/N$ and raise the predictable fraction, turning the bound reported here into guidance on acquisition cadence.

8. Conclusions

A Gaussian process fitted to spatial impedance profiles from an operational subsea cable produces predictive intervals whose empirical coverage is close to nominal at every level tested, while capturing almost none of the error reduction available to it. A constant profile given intervals from its own training residuals is calibrated to the same degree, and an untrained model slightly better. From four forecast origins, coverage rises and falls with how much more variable the test period is than training, and the constant profile's coverage moves with it. Nearly all of the predictive variance is a noise term, and interval width restates how variable each position was during training. These results carry block-bootstrap intervals, and a synthetic signal with a known predictable fraction reproduces them. Coverage approaches nominal in this way only when residuals are close to Gaussian and stable over time; where they are not, as over the first 190 m of this cable, coverage falls short.

Coverage alone therefore certifies very little. Where an interval is acted upon, it should be reported with its width against a stated criterion and a skill score against a baseline given the same information.

Data availability statement

The data that support the findings of this study were provided by a third party under a non-disclosure agreement. Restrictions apply to the availability of these data, which were used under licence for this study. The data are not publicly available. Derived quantities that do not disclose the underlying measurements, including all summary statistics and model outputs reported here, are available from the corresponding author on reasonable request and with the permission of the third party.

References

- [1] Warnock J, McMillan D, Pilgrim J and Shenton S 2019 Failure rates of offshore wind transmission systems *Energies* **12** 2682
- [2] Galski E, Anders G J, Jongen R A, Parciak J, Siemiński J, Piesowicz E and Paszkiewicz S 2021 Discussion of electrical and thermal aspects of offshore wind farms' power cables reliability *Renew. Sustain. Energy Rev.* **151** 111580
- [3] Rasmussen C E and Williams C K I 2006 *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press)
- [4] Gneiting T, Balabdaoui F and Raftery A E 2007 Probabilistic forecasts, calibration and sharpness *J. R. Stat. Soc. B* **69** 243–268
- [5] Gneiting T and Raftery A E 2007 Strictly proper scoring rules, prediction, and estimation *J. Am. Stat. Assoc.* **102** 359–378
- [6] Dawid A P 1984 Statistical theory: the prequential approach *J. R. Stat. Soc. A* **147** 278–292
- [7] Bröcker J and Smith L A 2007 Increasing the reliability of reliability diagrams *Weather Forecast.* **22** 651–661
- [8] Hersbach H 2000 Decomposition of the continuous ranked probability score for ensemble prediction systems *Weather Forecast.* **15** 559–570
- [9] Keprate A, Sheikhi S, Siddiqui M S and Tanwar M 2023 Comparing deep learning based image processing techniques for unsupervised anomaly detection in offshore wind turbine blades *2023 IEEE Int. Conf. on Industrial Engineering and Engineering Management (IEEM)* pp 274–278
- [10] Nicholls-Lee R, Thies P R, Dulieu-Barton J M, Ólafsson G, Hughes R and Hernandez Arroyo A 2022 Non-destructive examination (NDE) methods for dynamic subsea cables for offshore renewable energy *Prog. Energy* **4** 042011
- [11] Fothergill J C, Dodd S J, Dissado L A, Liu T and Nilsson U H 2011 The measurement of very low conductivity and dielectric loss in XLPE cables: a possible method to detect degradation due to thermal ageing *IEEE Trans. Dielectr. Electr. Insul.* **18** 1544–1553

- [12] Dong X, Yuan Y, Gao Z, Zhou C, Wallace P and Alkali B 2014 Analysis of cable failure modes and cable joint failure detection via sheath circulating current *2014 IEEE Electrical Insulation Conf. (EIC)* pp 294–298
- [13] Sommervogel L 2020 Various models for faults in transmission lines and their detection using time domain reflectometry *Prog. Electromagn. Res. C* **103** 123–135
- [14] Hirai N, Yamada T and Ohki Y 2012 Comparison of broadband impedance spectroscopy and time domain reflectometry for locating cable degradation *2012 IEEE Int. Conf. on Condition Monitoring and Diagnosis* pp 229–232
- [15] Tashman L J 2000 Out-of-sample tests of forecasting accuracy: an analysis and review *Int. J. Forecast.* **16** 437–450
- [16] Gardner J R, Pleiss G, Weinberger K Q, Bindel D and Wilson A G 2018 GPyTorch: black-box matrix-matrix Gaussian process inference with GPU acceleration *Advances in Neural Information Processing Systems* **31** 7576–7586
- [17] Künsch H R 1989 The jackknife and the bootstrap for general stationary observations *Ann. Stat.* **17** 1217–1241