FLOATSENSE: A TIME-SERIES DATASET AND BENCHMARK FOR FATIGUE LOAD RECONSTRUCTION ON FLOATING OFFSHORE WIND TURBINE TOWERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fatigue limits the lifetime of an offshore wind turbine tower and is tracked from the bending moment along it, yet strain gauges are rare and cover one section, while a tower-top accelerometer is standard. Reconstructing the moment at every height from it is a time-series problem, and no public benchmark scores it by fatigue damage. We introduce **FLOATSense**, a dataset and benchmark for fatigue load reconstruction on 22 MW floating offshore wind turbines: 717,948 time series at 10 Hz from the 19,404 OpenFAST simulations of FLOATBench on three tower geometries, with accelerometer and SCADA inputs and the fore-aft moment at eleven heights as target. A reconstruction is scored by the fatigue damage it implies, not by waveform error, and the six realizations of an operating point are kept together, so a model that reads only the condition gets no credit on the within-condition metric. Twenty learned models and an adapted field-validated physics method are evaluated within tower, across towers and under sensor ablation. Sensor-level leaderboards mislead: base log-damage R^2 saturates among the leading models and its leaders collapse at the top, where at the declared budget only three learned models beat a one-gain floor and a condition-only control given the true operating condition exceeds them all in aggregate R^2, so realization tracking must be read within condition; longer training recovers the top for the two TCNs, not for the tested attention and state-space models; five labeled simulations of a new tower recover its base by fine-tuning and its top by one scale per height, which fine-tuning often misses; and a second accelerometer axis improves the top at every budget. Limitations: the campaign is simulated, wind and waves are aligned, and only the fore-aft moment is scored. Resources: data and code at `https://anonymous.4open.science/w/FLOATSense-56B2/review/index.html`.

1 INTRODUCTION

Many scientific and engineering problems require inferring physical quantities that cannot be measured directly from signals that can (Arridge et al., 2019; Deng et al., 2022; Farrar & Worden, 2007; Kapteyn et al., 2021). When both measurements and targets evolve in time, this becomes an inverse time-series problem: reconstructing an unobserved response history from observed signals. Modern machine-learning models offer a scalable way to learn such mappings from data.

The challenge is evaluation. A benchmark must establish whether a model reconstructs the target across target locations and beyond its training conditions: under regimes outside the sampled envelope, on unseen structures and under different sensor configurations. It must also separate within-condition correlation, the ability to explain variation among realizations of one condition, from recognition of the condition. Because a model can learn the conditional mean of the target without tracking the realization, aggregate accuracy can conflate condition-level prediction, which tabular models achieve (Section 2), with within-condition correlation (Gelman & Hill, 2007). To our knowledge, existing time-series (Godahewa et al., 2021; Wang et al., 2026; Ansari et al., 2024), scientific-ML (Takamoto et al., 2022; Ohana et al., 2024) and inverse-problem (Zheng et al., 2025; Deng et al., 2022) benchmarks do not bring these requirements together in one task.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Floating offshore wind turbines (FOWTs) bring these challenges together in a single system. As turbines scale toward 22 MW (Zahle et al., 2024) and deployment moves into water too deep for fixed foundations (International Energy Agency, 2019), their towers respond to wind, waves and platform motion across operating regimes, including extremes measurement campaigns rarely capture. Since wind and waves are stochastic, one condition produces different load histories. Fatigue, a key lifetime constraint, is assessed from bending-moment histories along the tower, not only where a sensor is installed. Strain gauges measure these moments but are rare and limited to one section, whereas tower-top acceleration and Supervisory Control and Data Acquisition (SCADA) signals (rotor speed, blade pitch, wind speed) are readily available (Pimenta et al., 2024). Assessing fatigue at uninstrumented heights therefore requires reconstructing moment histories from the measurements available on a given turbine, generalizing across tower designs, and computing their damage.

We therefore introduce **FLOATSense** (Figure 1), built on the FLOATBench campaign (Anonymous, 2026): from the same simulations it releases the synchronized sensor and load histories, previously unavailable, and poses a distinct inverse task, reconstructing moment histories from measured signals. Reconstructions are scored through their implied fatigue damage and against a physics-based baseline. Our contributions are: **(i) Dataset:** the release, for reconstruction, of the 717,948 channel time series (37 channels, 10 Hz, 1,000 s) of the 19,404 FLOATBench OpenFAST simulations on three 22 MW towers, 1,078 operating points $\times$ six turbulence seeds per tower, fore-aft moments at 11 heights as targets, plus window statistics. **(ii) Protocol:** damage-based scoring at eleven heights, with a within-condition metric separating realization from condition, under regime shift, cross-tower transfer and sensor ablation, against our adaptation of the field-validated method of Pimenta et al. (2024). **(iii) Benchmark:** 20 models, eight families: base log-damage R^2 saturates, the top separates them, part of the top gap is under-training, and aggregate accuracy hides weak realization tracking, lifetime bias and a condition-only predictor beating every benchmark sensor model at the top, showing that aggregate scores can be high for the wrong reason.

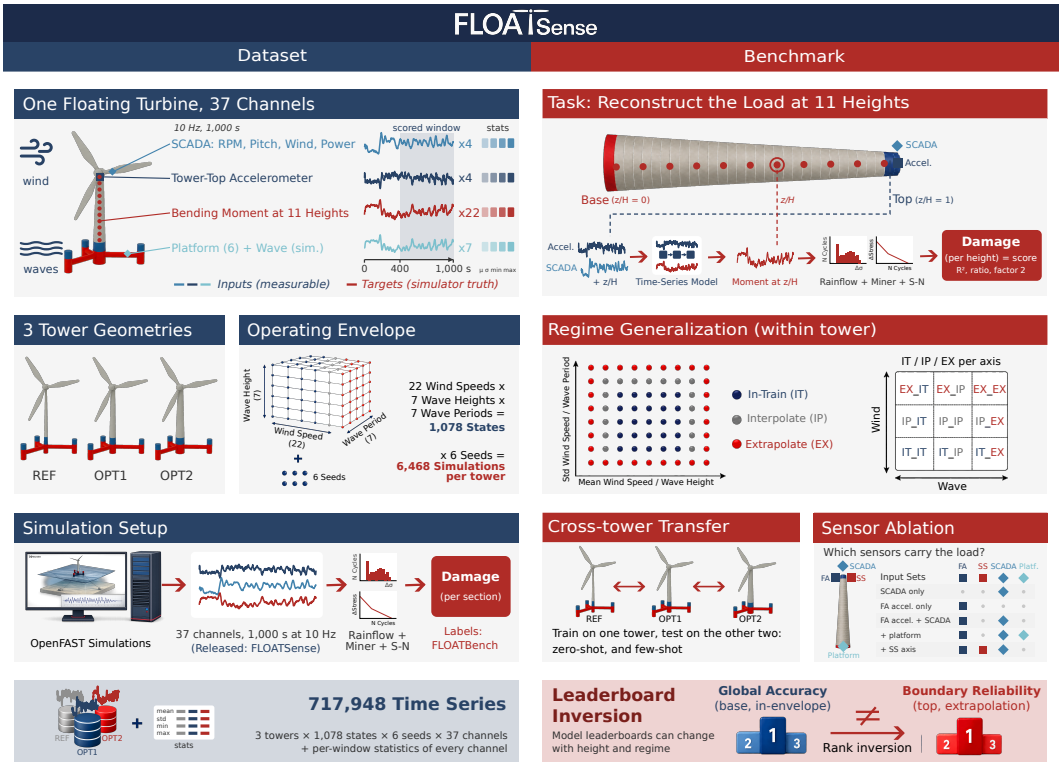


Figure 1: **FLOATSense:** a time-series dataset and benchmark for fatigue load reconstruction on 22 MW FOWT towers. Dataset: 717,948 channel time series (37 channels) from 19,404 simulations on three geometries, moments at 11 heights as targets. Benchmark: damage reconstruction along the tower, regime generalization, cross-tower transfer and sensor ablation, against a physics baseline.

2 RELATED WORK

Time-series, scientific-ML and shift benchmarks. Time-series archives (Godahewa et al., 2021; Wang et al., 2026) and time-series foundation models (Ansari et al., 2024; Goswami et al., 2024) cover forecasting, imputation and anomaly detection. Extrinsic regression (Tan et al., 2021) predicts one number per series. Scientific-ML benchmarks provide simulator reference solutions, mostly for forward problems (Takamoto et al., 2022; Ohana et al., 2024; Leyli-abadi et al., 2022), some with physics baselines (Rasp et al., 2024; Nathaniel et al., 2024), extrapolation splits (Bonnet et al., 2022) or inverse tasks (Zheng et al., 2025). Distribution-shift benchmarks evaluate generalization across domains, environments or time (Koh et al., 2021; Ragab et al., 2023; Gagnon-Audet et al., 2023; Yao et al., 2022). To our knowledge, no benchmark combines inverse time-series reconstruction with held-out operating regimes, repeated realizations, multiple structures and varying input subsets. Repeated realizations are crucial because aggregate performance can otherwise be driven by differences between operating conditions rather than by reconstruction of individual realizations (Gelman & Hill, 2007; Geirhos et al., 2020).

Load reconstruction on wind turbine towers. Modal-based reconstruction (Iliopoulos et al., 2016) and joint input-state estimation (Maes et al., 2016) reconstruct structural response from a few accelerometers, while SCADA–accelerometer fusion (Noppe et al., 2016) combines SCADA-based quasi-static and accelerometer-based dynamic load estimates. Kalman filters have been applied onshore (Branlard et al., 2020) and extended to floating turbines (Branlard et al., 2024), and Pimenta et al. (2024) replaces SCADA-based regression with physics-derived frequency-band gains on the tower-top acceleration, validated on full-scale floating-turbine measurements. Learned approaches include physics-guided losses (de N Santos et al., 2023), Bayesian neural networks (Hlaing et al., 2024), Gaussian-process latent-force models (Bilbao et al., 2022), LSTM virtual sensors (Dimitrov & Göçmen, 2022; Gräfe et al., 2024), PCA strain estimation (Tarpø et al., 2022) and GP fatigue surrogates (Avendaño-Valencia et al., 2021); population-based structural health monitoring (Bull et al., 2021; Gardner et al., 2021) further motivates transfer across structures. These methods are evaluated at only one or a few sections where load references are available, and therefore do not test performance along the tower, nor generalization across regimes, geometries and sensor sets.

Wind turbine datasets and benchmarks. Wind-fatigue studies and databases (Müller & Cheng, 2016; Dimitrov et al., 2018; Slot et al., 2020; Papi et al., 2023; Papi & Bianchini, 2024) label operating points at one or a few tower sections and support surrogate models (Haghi & Crawford, 2024; Singh et al., 2024; 2025; Ren & Xing, 2023). FLOATBench extends this setting to multiple tower

Table 1: Tower load reconstruction (inverse) and wind fatigue datasets (forward) vs. FLOATSense.

						Data and evaluation axes					
Reference	Power	Support	Scale	#sect	#labels	Regime	Struct.	Realiz.	Sensor	Phys.	Public
Load reconstruction (inverse, from a measured series)											
Branlard et al. (2020)	5	Onshore	60 sims	3	–	✗	✗	✓	✗	✓	✗†
Bilbao et al. (2022)	–	Onshore	17×10 min	1	–	✗	✗	✗	✗	✓	✗
Maes et al. (2016)	3	Monopile	8×10 min	2	–	✗	✗	✗	✗	✓	✗
Noppe et al. (2016)	3	Monopile	4 months	1	–	✗	✗	✗	✗	✓	✗
de N Santos et al. (2023)	9.5	Monopile	9 months	1	–	✗	✗	✗	✗	✗	✗
Hlaing et al. (2024)	–	Monopile	2 years	1	–	✗	✗	✗	✓	✗	✗
Branlard et al. (2024)	3.6	TetraSpar	4 days	3	–	✗	✗	✗	✗	✓	✗†
Pimenta et al. (2024)	8.4	Semisub.	6 months	1	–	✗	✗	✗	✗	✓	✗
Fatigue datasets (forward, from the operating condition)											
Dimitrov et al. (2018)	10	Onshore	10 k (9-D)	2	160 k	✗	✗	✓	✗	✗	✗
Slot et al. (2020)	5	Onshore	0.6 k (5-D)	1	0.6 k	✗	✗	✗	✗	✗	✗
IWT 7.5 MW (2023)	7.5	Onshore	15 k (4-D)	1	47 k	✗	✗	✓	✗	✗	✓
Haghi & Crawford (2024)	5	Onshore	33 k (3-D)	2	65 k	✗	✗	✗	✗	✗	✓
Müller & Cheng (2016)	5	Jacket	4.4 k (5-D)	2	9 k	✗	✗	✓	✗	✗	✗
Singh et al. (2024)	10	Onshore+monopile	9 k (3-D/6-D)	2	107 k	✗	✗	✓	✗	✗	✓
FLOATECH (2023)	5/10	Semisub./Spar/Hexa	0.5 k (5-D)	1	1.5 k	✗	✗	✗	✗	✗	✓
Liu et al. (2024a)	5	Semisub. (OC4)	0.2 k (3-D)	1	0.2 k	✗	✗	✗	✗	✗	✗
Papi & Bianchini (2024)	5	Semisub. (OC4)	447 k (hind.)	2	894 k	✗	✗	✗	✗	✗	✓
Singh et al. (2025)	6	Spar (Hywind)	9 k (7-D)	2	22 k	✗	✗	✓	✗	✗	✗
FLOATBench (2026)	22	Semisub. (IEA-22)	6.5 k (3-D)	30*	582 k	✓	✓	✓	✗	✗	✓
Dataset and load-reconstruction benchmark on the FLOATBench simulations (this work)											
FLOATSense (ours)	**22**	**Semisub. (IEA-22)**	**6.5 k (3-D)**	**11**	**718 k**	✓	✓	✓	✓	✓	✓

Power: rated MW; *Scale*: simulations or campaign length; *#sect*: tower sections scored for damage; *#labels*: released labels; *Regime*: out-of-envelope test; *Struct.*: other design; *Realiz.*: repeated turbulence realizations per condition; *Sensor*: varying measurement/input sets; *Phys.*: physics-based method or baseline; †code only; *30 sections interpolated from the 11 heights.

heights, operating regimes and geometries, but predicts fatigue scalars from operating conditions instead of reconstructing load histories from measured signals. To our knowledge, no public resource offers a sensor-to-load time-series benchmark combining regime and structure shifts, repeated realizations, varying sensor sets and a physics baseline (Table 1).

3 DATASET

FLOATBench released the fatigue labels of this campaign; FLOATSense releases its synchronized sensor and load time series.

Operating envelope and simulation setup. The campaign comprises 22 wind speeds and, at each, a 7×7 grid of 49 sea states in significant wave height H_s and peak period T_p, for 1,078 operating points. The IEA 22 MW turbine (Zahle et al., 2024) on a three-column semi-submersible is simulated with OpenFAST (Jonkman, 2013) under IEC 61400-3-2 DLC 1.2 (IEC TS 61400-3-2:2019, 2019), with aligned wind and waves. Each operating point has six realizations per tower, differing in the turbulence seed (Jonkman & Buhl, 2006). A further 22 parked runs per tower (rotor parked, waves only, no wind), one at the median sea state of each wind speed, calibrate the physics baseline (Appendix F). Each run lasts 1,000 s at 10 Hz; the scored window is the post-transient 400–1,000 s. Simulation and environmental specifications are in Anonymous (2026).

Tower geometries. The campaign is repeated on three towers of equal height sharing platform, controller and mooring, for 19,404 operating simulations (plus 66 parked calibration runs): the IEA 22 MW baseline REF and the fatigue-aware redesigns OPT1 and OPT2 (Ribeiro et al., 2026). REF is the flexible outlier (first fore-aft frequency 0.34 Hz versus 0.57 and 0.54 Hz) and carries about 40 times the base lifetime damage of the two redesigns: cross-tower shift at two levels (Appendix E.1).

Channels. Each simulation is an array of 10,001 samples by 37 channels (Table 5, Figure 2): tower-top accelerations, raw and gravity-corrected ($\times 4$; Appendix F), SCADA signals ($\times 4$), platform motions ($\times 6$), wave elevation ($\times 1$), and fore-aft and side-side moments at 11 heights ($\times 22$).

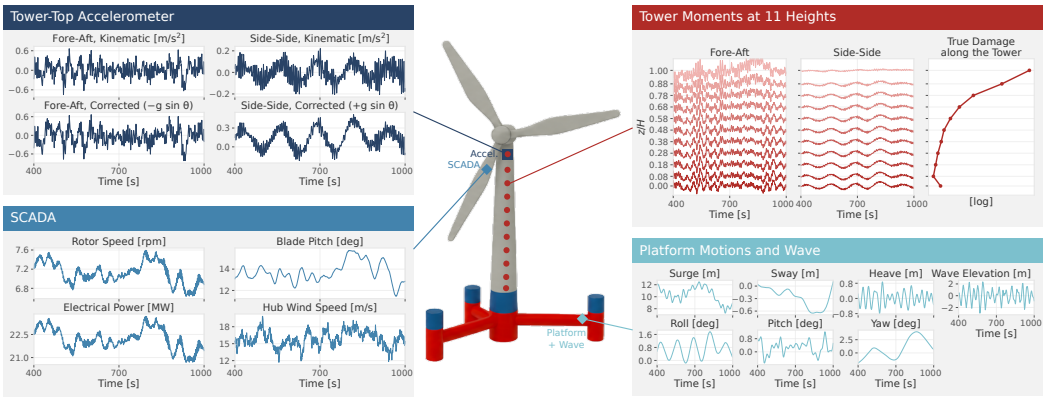


Figure 2: **FLOATSense channels.** The 37 released time series of one simulation (1,000 s at 10 Hz).

Dataset schema. Everything is Parquet, keyed by the FLOATBench simulation identifier: the series, their scored-window statistics, the reference damage at the 11 heights, the tower geometry and a metadata table (operating point, realization, split, regime labels and lifetime weight); the 66 parked runs come in one further file (Appendix E.3).

Hosting and licensing. During review a fixed subset, 47 simulations per tower with the complete tabular and parked-run files and the trained checkpoints, is served from an anonymized repository (Appendix B); the full dataset is released on Hugging Face on publication, under CC-BY-4.0, with a Croissant 1.1 manifest and a Datasheet for Datasets (Gebru et al., 2021). The benchmark code is released under MIT license (Appendix A–Appendix C).

Compute. The simulation campaign required approximately 19,400 CPU-hours across the three towers, about 2.2 core-years on a commercial cloud HPC platform.

4 BENCHMARK PROTOCOL

4.1 TASK

Given the gravity-corrected fore-aft tower-top acceleration, three SCADA channels (rotor speed, blade pitch and hub wind speed; the generator power is released and not used, Appendix M.4) and a requested height z/H, a fraction of the tower height $H = 149.4\,\mathrm{m}$, the task is to reconstruct the fore-aft bending-moment history at that height over the scored window, 400 to 1,000 s. One model serves the full tower with z/H as a constant input channel: training samples one of the 11 target heights per example, testing evaluates all 11 against the OpenFAST reference histories. Only the fore-aft moment is scored, as in the fatigue design of these towers (Ribeiro et al., 2026); the second acceleration axis and field-style SCADA inputs are sensor extensions (Section 5.4).

4.2 EVALUATION PROTOCOLS

Within tower. We use the regime-aware partition of FLOATBench, with the same simulations on the three towers: 1,728 train (18 wind speeds, inner 4×4 sea states) and 4,740 test, with all six seeds of an operating point on the same side. Each test simulation has its FLOATBench regime label on the wind and wave axes, in-train (IT), interpolation (IP) or extrapolation (EX; Appendix E.2).

Cross tower. A model trained on one tower is applied to each of the other two (six source–target pairs); the towers share the wind and wave realizations, so the transfer isolates the change of tower design. Settings: *zero-shot*, the source model unchanged, and *ten-shot*, the same model after 100 adaptation steps on ten target training simulations. Other budgets (5 to 100 simulations, 50 to 1,250 steps) are analysis, not protocol (Appendix L.2).

Sensor ablation. The task uses the field method's sensors (Pimenta et al., 2024), one accelerometer axis and SCADA. The ablation tests what other sensors would add, on the three towers and on the selected multichannel models (Section 4.3): the accelerometer alone, SCADA alone, both accelerometer axes, or the six platform motions added (Figure 2; results in Section 5.4, Appendix M).

4.3 BASELINES AND LEARNED MODELS

Physics baseline. We use an adaptation of the field-validated method of Pimenta et al. (2024), in which the moment spectrum is the acceleration spectrum times a gain with six constants, calibrated on the parked runs and the training split. We extend it to every height with a height factor of our own, tested only on these simulations (Appendix F); across towers it is calibrated on the target tower.

Floor and learned models. The floor (NAIVE) is one scalar gain from the tower-top acceleration to the moment, fitted by least squares on the training split in standardized units, then rescaled by each height's training moment standard deviation (Appendix G.1), with no physics, no conditioning and no other input. Above it, twenty models in eight families (convolutional, recurrent, attention, state-space, spectral, period-folding, pretrained encoders with a trained head and physics-anchored), among them a Prob-TCN with a heteroscedastic head, a condition-dependent spectral gain, and a hybrid that learns a correction of the physics gain, bounded a priori at half the physics reconstruction rather than fitted (Table 9). The pretrained encoders are univariate: they read the accelerometer and ignore SCADA and the height. Architectures, parameter counts and citations are in Appendix G.1.

4.4 TRAINING, SELECTION AND SEEDS

Training. All models follow the same default training recipe, with architecture-specific exceptions stated below, so that a gap reflects the implementation under this recipe, not the tuning, and no choice is made on the test split: 50 epochs of Adam at 10^{-3}, batch 16, mean squared error on the standardized moment, no weight decay, scheduler or early stopping, with the last epoch scored. This is the

fixed-budget track of the leaderboard. Inputs and targets are standardized from the source-tower training split, channel by channel and per height. Adaptation continues from the source checkpoint on ten target simulations, at batch 4 with the same normalization. Mamba uses 3×10^{-4} because it diverges at the default rate, Prob-TCN replaces the mean squared error by a Gaussian likelihood and is scored on one draw of its predictive distribution (Appendix J.3), and the five length-fixed models train on the full window. Cost is in Appendix G.2; Appendix J.1 retrains ten models at three times the budget and four at six, to test the sensitivity to convergence.

Selection and seeds. A held-out 20% of the training operating points serves only for model selection: one representative of each of the eight families by R^2 of $\log_{10}$ damage over the eleven heights, plus the union of the top three at the base, over the heights and at the top, which yields the ten models that receive the additional experiments (Appendix G.5). At the longer budgets, the sensor sets are probed on the TCN, first under the ranking rule and never divergent, and the TCN, Prob-TCN and S4D run three seeds. Seed 0 (initialization, crops, batch order) runs everywhere, with three seeds for thirteen models within tower, ten in adaptation and eight in the sensor ablation (Appendix H.2).

4.5 Metric and statistical significance

The true and the reconstructed moment are low-passed at 3 Hz and pass through the damage model that generated the FLOATBench labels (Anonymous, 2026): rainflow counting (Matsuishi & Endo, 1968; Downing & Socie, 1982), the DNV-RP-C203 S–N curve (DNV, 2024) and Miner's rule. No single scalar ranks a reconstruction, so two axes are scored. Level: R^2 of $\log_{10}$ damage (sorts the leaderboard), the median ratio $\hat{D}/D$ (bias), the mean relative error (MRE, per regime) and the fraction within a factor of two. Realization: over the six realizations of each operating point, the within-condition correlation ρ_{wc} (which one is high) and slope b_{wc} of the reconstructed on the true log-damage deviation from the point mean (how much; 1 is exact), on which a condition-only model scores zero. Waveform fidelity (correlation, NRMSE, variance ratio) is reported as a third axis (Appendix K.5); no axis names a global winner. Each carries a 95% percentile-bootstrap interval ($B = 1{,}000$), resampling operating points, not simulations, since the six realizations are not independent. Pairwise differences are significant when the paired-bootstrap interval excludes zero on all three towers. Definitions: Appendix G.3 and Appendix G.4.

5 Results

Numbers are at 50 epochs unless stated, the median of three seeds or seed 0 (†); triples read REF / OPT1 / OPT2. Rankings are of this fixed-budget track, not of architectures; longer budgets are probes on a subset of models (Appendix J.1; last column of Table 2).

5.1 Along the tower: the base saturates, the top separates the models

At the base the sequence models saturate log-damage R^2, and the leading ones beat the physics. Ten of the 22 entries reach $R^2 \geq 0.95$ at the base and the first seven lie within 0.023 (Table 2; paired tests still separate some, Appendix H.5): Mamba, PatchTST and TCN score 0.99 to 1.00, the physics 0.967 / 0.944 / 0.961 and the floor 0.742 on average. Longer budgets move it by at most 0.04.

The top spreads the entries from -2.3 to 0.84, and the base ranking does not predict it. The base-to-top rank correlation over the 22 entries is -0.37 to 0.22 (Appendix I.1). Under the fixed-budget recipe PatchTST and Mamba, first at the base, fall below zero at the top, while TCN keeps 0.73 to 0.78 and Prob-TCN 0.82 to 0.86 (Table 2). The physics holds to z/H 0.78 and at the top of REF (0.93), but its height factor extrapolates on the redesigns and fails there (Appendix F). Only Prob-TCN, TCN and FITS beat the floor at the top (0.70), TCN by less than its seed spread.

A condition-only control wins aggregate top R^2; only the realization metrics expose it. An XGBoost given the true wind, H_s and T_p, which no sensor model sees, is a diagnostic control, not an entry: it scores 0.95 at the top on every tower, above all 22 entries (Table 2), yet it is constant within an operating point, so $\rho_{\text{wc}} = b_{\text{wc}} = 0$. Top R^2 therefore scores the damage level; the realization is scored by ρ_{wc} (TCN 0.68, Prob-TCN 0.66) and b_{wc} (TCN 0.61 to 0.75, Prob-TCN

Table 2: **Within tower along the height.** R^2 of $\log_{10}$ damage at three heights; at the top also % within a factor of two, ρ_{wc} and lifetime damage ratio; last column: top R^2 at the longest budget (epochs as superscript). Tower mean of three-seed medians; subscript: lowest tower. *Selected on validation. †Seed 0. §Controls, not entries (seed 0, one noise draw; Appendix J.3; condition-only: XGBoost on the true U, H_s, T_p, constant within an operating point, Appendix K). "–": not run (longer budgets only for the ten selected models; physics and floor are calibrated, not trained). All 22 entries in Appendix H.1.

	R^2 of $\log_{10}$ damage			top			top
Model	base	z/H 0.78	top	within 2 (%)	ρ_{wc}	lifetime	longest
Prob-TCN*	0.959	0.96	$0.84_{0.82}$	59	0.66	1.59	0.94^{300}
TCN*	0.993	0.98	$0.76_{0.73}$	59	0.68	0.63	0.93^{300}
FITS*	0.764	0.81	$0.72_{0.66}$	40	0.38	0.50	$0.63^{150\dagger}$
DLinear	0.907	0.88	$0.67_{0.54}$	36	0.54	0.30	–
Chronos*	0.827	0.74	$0.50_{0.31}$	32	0.43	0.23	$0.49^{150\dagger}$
Spectral gain	0.972	0.93	$0.48_{0.38}$	1	0.60	0.21	–
S4D*	0.984	0.92	$0.43_{0.32}$	3	0.56	0.18	0.63^{300}
LSTM*	0.988	0.92	$0.42_{0.33}$	2	0.62	0.21	$0.62^{150\dagger}$
Mamba*	0.995	0.92	$-0.28_{-0.38}$	0	0.55	0.07	$0.17^{150\dagger}$
PatchTST*	0.994	0.78	<-1	0	0.47	0.02	$-0.23^{300\dagger}$
TimesNet*	0.857	0.66	<-1	2	0.25	0.02	$<-1^{150\dagger}$
Hybrid*	0.960	0.82	<-1	0	0.49	0.02	$<-1^{150\dagger}$
Physics†	0.957	0.86	<-1	24	0.48	0.23	–
Floor†	0.742	0.80	$0.70_{0.65}$	32	0.57	0.36	–
TCN + colored noise§	*0.990*	–	$0.90_{0.88}$	*76*	*0.68*	*1.12*	–
Prob-TCN mean head§	*0.984*	–	$0.75_{0.68}$	*26*	*0.73*	*0.47*	–
Condition-only§	*0.93*	–	$0.95_{0.95}$	*89*	*0*	–	–

0.31 to 0.51), where every model attenuates the realization-specific deviation (Appendix K). LSTM and S4D reach 0.3 to 0.55 at the top with at most 4% within a factor of two and ρ_{wc} at most 0.63 (0.95 to 0.97 at the base, Table 23).

Why the top fails: mostly training and variance, and below rated the fore-aft axis. At 300 epochs TCN and Prob-TCN reach 0.92 to 0.95 at the top, level with each other and, on REF, with the physics (Figure 3a, dotted); the TCN seed range shrinks from 0.06–0.15 to 0.004–0.03, and its validation loss is lowest at epochs 230 to 300. S4D stays near 0.6 and PatchTST below zero. Trained for the top alone, TCN and Prob-TCN reach 0.90 to 0.98 while PatchTST, S4D and Mamba still fail; a damage-aware loss rescues S4D (Appendix J.2). The all-height TCN keeps 0.73 to 0.78 of the true variance at the top (Table 35); restoring it with noise lifts the score, not ρ_{wc} (Appendix J.3).

At this budget a top score measures convergence as much as architecture. The collapse of PatchTST, Mamba and S4D and the lifetime bias survive a longer budget; the TCN–Prob-TCN gap does not.

5.2 By regime: at the top, low wind is the hard condition

Waves matter far less than wind and height; at the base wind extrapolation costs the physics and Prob-TCN; at the top low wind costs everyone. Wave extrapolation moves the fraction within a factor of two by -10 to $+9$ points pooled over wind for every model at any height (Appendix I.5). At the base, on extrapolated wind, the physics falls to 73 to 80% on the redesigns (its rotor constants are fitted at low wind) and Prob-TCN to 58 to 60%, while PatchTST and Mamba stay at 100%. At the top the axis that matters is wind speed, not the regime label (Figure 3b): TCN is within a factor of two on 0 to 12% of simulations below 10 m/s, in train or not, and on 92 to 99% above 13.5 m/s. The cause is spectral: below rated wind the 3P rotor harmonic (26% of the top variance) falls in the wave band, where the fore-aft acceleration is weakly coherent with the top moment (0.11 to 0.38) and the side-side axis is not (0.67 to 0.75) (Figure 16, Figure 17). Training longer lifts the in-train low winds far more than the extrapolated ones, so the gap widens; winds below 10 m/s carry only 3 to 5% of the lifetime top damage.

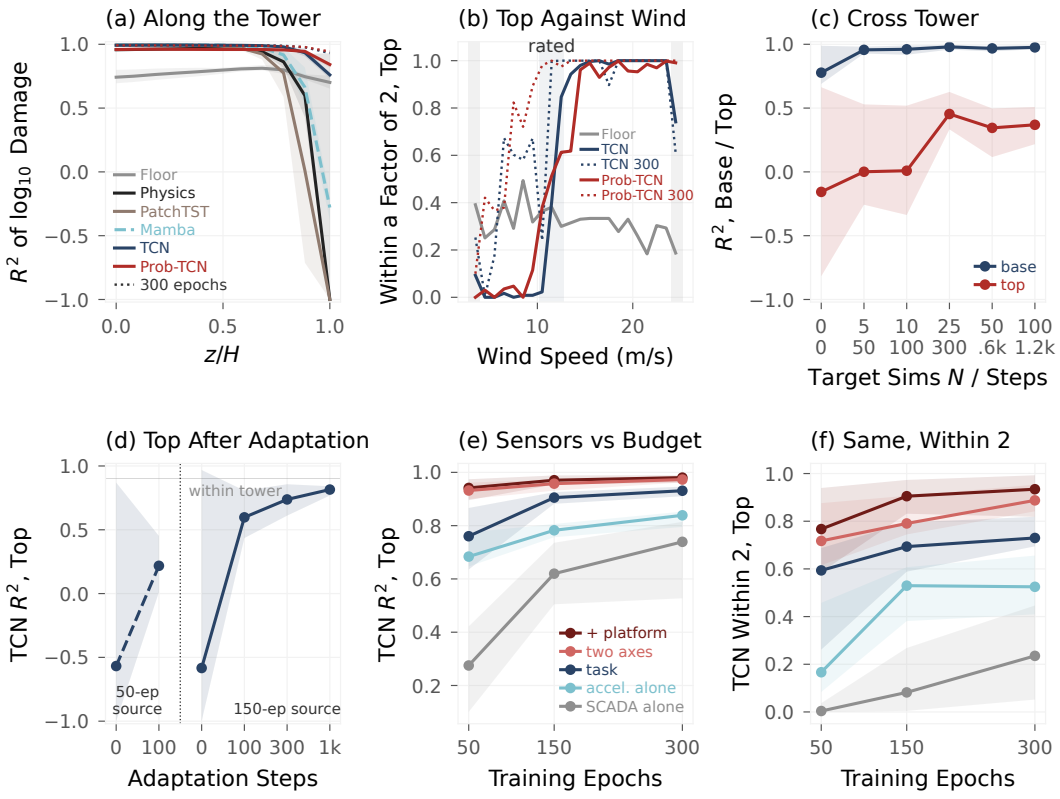


Figure 3: **Where the benchmark separates models, and what survives a longer budget.** (a) R^2 of $\log_{10}$ damage along the tower, 50 epochs (solid), 300 epochs for TCN and Prob-TCN (dotted). (b) Top fraction within a factor of two against wind speed, TCN, Prob-TCN (dotted: 300 epochs) and floor; extrapolated speeds shaded. (c) Cross tower: base and top R^2 against target simulations N (steps under the ticks), median of six models, mean over five pairs (sixth: Table 39); seed 0, draw 0. (d) TCN top against adaptation steps from a 50- and a 150-epoch source, mean over six pairs, seed 0; gray: within tower. (e, f) TCN top R^2 and within 2 under five input sets against epochs; band: nine runs. Unless stated: tower mean of three-seed medians; band: range over towers (a) or pairs (c, d); clipped at -1.

5.3 Cross tower: zero-shot keeps the ranking, not the level

Zero-shot fails on scale, not on ranking. Over the 20 learned models the median zero-shot R^2 at the base is 0.93 and 0.84 between the redesigns, 0.75 and 0.71 into REF, where the models over-predict (median ratio 2.6 and 2.5), and 0.41 and 0.48 out of it, where they under-predict (0.30 and 0.34); ρ_{wc} stays at 0.84 to 0.95, so the realizations are ranked right and placed at the wrong level (Table 37). TCN into REF scores -0.34 / -0.14 (median ratio 10.5 and 8.5).

Five target simulations recover the base by fine-tuning and the top by one scale per height. With five labeled target simulations (50 steps) the sequence models score 0.89 to 1.00 at the base on every pair (Figure 3c), and the physics recalibrated on ten 0.94 to 0.99. At the top, after ten-shot adaptation, Prob-TCN averages 0.71, the floor 0.56 and TCN 0.44 (Table 39). One scale factor per height fitted on the same five simulations (no gradient step, labels at every height) brings the TCN top to 0.75 to 0.94 on all six pairs, above its fine-tuning at every budget (Appendix K.4). The budget curve mixes data and optimization budget (steps grow with N); at fixed N (Figure 3d) the TCN top goes from 0.60 after 100 steps to 0.82 after 1,000 (0.90 within tower).

5.4 SENSORS: THE SECOND ACCELEROMETER AXIS HELPS THE TOP AT EVERY BUDGET

The ordering of the input sets holds at every budget; the R^2 gain of the second axis is mostly a 50-epoch effect, its fatigue gain is not. At the top the TCN order is platform motions added (0.94), both accelerometer axes (0.93), task inputs (0.76), accelerometer alone (0.68) and SCADA alone (0.28), and it holds at 300 epochs (Figure 3e). The R^2 gain of the second axis shrinks from 0.17 to 0.04, but at 300 epochs it still brings 87 to 90% of simulations within a factor of two against 72 to 76% (Figure 3f), and ρ_{wc} 0.94 to 0.95 against 0.82 to 0.88. At the base only the accelerometer alone (0.96) and SCADA alone (0.89) fall behind; generator power and field SCADA stay within seed spread (Appendix M).

The accelerometer is the sensor; a second axis brings the top fatigue metrics close to the base.

5.5 LIFETIME DAMAGE: THE UPPER TOWER GOVERNS, AND RECONSTRUCTION FAILS THERE

Over a 25-year life (DLC-1.2-equivalent fore-aft damage; test split, 72% of the lifetime weight), the top of the redesigns carries at least as much damage as the base (1.5 times on OPT1, 1.07 times on OPT2, where it peaks at z/H 0.88; Figure 4): the upper tower, where every model struggles, governs their fore-aft fatigue life. At the base the sequence models are within about 20% and the physics is conservative. At the top TCN recovers 0.61 to 0.66 of the lifetime damage and Prob-TCN 1.44 to 1.74 (closer to 1 at 300 epochs); the physics recovers almost none on the redesigns (Appendix I.3).

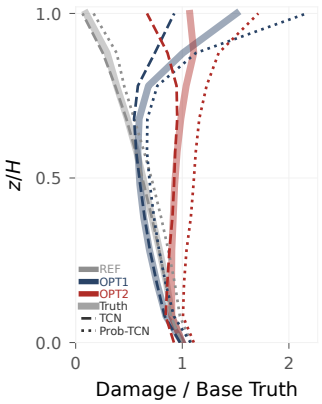


Figure 4: **Lifetime damage along the tower** over the true base damage: truth (thick), TCN (dashed), Prob-TCN (dotted); three-seed median, test split.

6 LIMITATIONS AND FUTURE WORK

Simulation only. No measured load series enters the benchmark; the calibrated physics constants give a plausible platform pitch period (Appendix F), a check, not a validation. Inputs are ideal; accelerometer noise, bias and drift and a noisy hub wind change neither the base nor the top (Appendix K.6), but a winner must still survive sampling jitter and unmodeled dynamics. OpenFAST is mid-fidelity (beam tower, engineering aero- and hydrodynamics); FEA and CFD re-runs will follow.

One turbine family, aligned seas. One 22 MW turbine on one platform under DLC 1.2: cross tower is a change of tower design, not of turbine class or load case. Aligned waves leave side-side without direct wave excitation, so that moment is released but not scored (Section 4.1) and combined-direction fatigue is not assessed. Supervision uses all eleven heights, not the one gauge of a field turbine; a base-gauge-only protocol and modal and Kalman baselines (Iliopoulos et al., 2016; Maes et al., 2016; Branlard et al., 2020) will be added to the leaderboard.

One recipe, short of convergence. All models share one 50-epoch budget, declared in advance and not converged at the top (Section 5.1, Appendix J.1): every top number is a number at this budget, and the longer probes cover four models at 300 epochs. A validation-tuned track (per-model learning rate, validation checkpointing, three seeds) will be added to the public leaderboard.

7 CONCLUSION

FLOATSense releases 717,948 synchronized time series from 19,404 operating OpenFAST simulations (six realizations per operating point, eleven heights) and a benchmark with a fixed protocol, a damage metric, named shifts and a physics method to beat. Its lesson is that aggregate reconstruction scores can be high for the wrong reason: base log-damage R^2 saturates, and at the top a condition-only control given the true condition beats every sensor model in R^2 while tracking no realization, so level and realization are scored apart. Under the fixed-budget track, longer training recovers the TCN top, five labeled target simulations adapt a new tower, and a second accelerometer axis improves the top at every budget.

Ethics statement

The dataset is simulated: it contains no personal data, no measurements of an operating asset and no proprietary design (the turbine and the platform are the public IEA-22 reference). Its intended use is the structural monitoring of floating wind turbines; a reconstruction used for a safety decision must be validated on the target structure, which the cross-tower protocol is designed to measure (Appendix D).

Reproducibility statement

Code, split files and the per-simulation damage files behind every cell are released, with a fixed data subset and the trained checkpoints during review and the full dataset on publication; every number in the paper is reproduced by three scripts with fixed seeds and draws (Appendix A). Appendices F and H give the physics equations, the calibrated constants and the full result matrix.

References

Anonymous. FLOATBench: A tabular dataset and benchmark for fatigue prediction on floating offshore wind turbine towers, 2026. Under review. Anonymized copy: `https://osf.io/s6dfj/?view_only=2bf4e9d06fad45bb86763a6798cea6f2`.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=gerNCVqqtR`.

Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019. doi: 10.1017/S0962492919000059.

Luis David Avendaño-Valencia, Imad Abdallah, and Eleni Chatzi. Virtual fatigue diagnostics of wake-affected wind turbine via Gaussian process regression. *Renewable Energy*, 170:539–561, 2021. doi: 10.1016/j.renene.2021.02.003.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Joaquin Bilbao, Eliz-Mari Lourens, Andreas Schulze, and Lisa Ziegler. Virtual sensing in an onshore wind turbine tower using a Gaussian process latent force model. *Data-Centric Engineering*, 3: e35, 2022. doi: 10.1017/dce.2022.38.

Florent Bonnet, Jocelyn Mazari, Paola Cinnella, and Patrick Gallinari. AirfRANS: High fidelity computational fluid dynamics dataset for approximating Reynolds-averaged Navier–Stokes solutions. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 23463–23478, 2022.

Emmanuel Branlard, Dylan Giardina, and Cameron S. D. Brown. Augmented Kalman filter with a reduced mechanical model to estimate tower loads on a land-based wind turbine: a step towards digital-twin simulations. *Wind Energy Science*, 5(3):1155–1167, 2020. doi: 10.5194/wes-5-1155-2020.

Emmanuel Branlard, Jason Jonkman, Cameron Brown, and Jiatian Zhang. A digital twin solution for floating offshore wind turbines validated using a full-scale prototype. *Wind Energy Science*, 9 (1):1–24, 2024. doi: 10.5194/wes-9-1-2024.

L. A. Bull, P. A. Gardner, J. Gosliga, T. J. Rogers, N. Dervilis, E. J. Cross, E. Papatheou, A. E. Maguire, C. Campos, and K. Worden. Foundations of population-based SHM, Part I: Homogeneous populations and forms. *Mechanical Systems and Signal Processing*, 148:107141, 2021. doi: 10.1016/j.ymssp.2020.107141.

Cristian Challu, Kin G. Olivares, Boris N. Oreshkin, Federico Garza Ramirez, Max Mergenthaler Canseco, and Artur Dubrawski. N-HiTS: Neural hierarchical interpolation for time series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6989–6997, 2023. doi: 10.1609/aaai.v37i6.25854.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 10148–10167, 2024. URL https://proceedings.mlr.press/v235/das24c.html.

Francisco de N Santos, Pietro D'Antuono, Koen Robbelein, Nymfa Noppe, Wout Weijtjens, and Christof Devriendt. Long-term fatigue estimation on offshore wind turbines interface loads through loss function physics-guided learning of neural networks. *Renewable Energy*, 205:461–474, 2023. doi: 10.1016/j.renene.2023.01.093.

Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. OpenFWI: Large-scale multi-structural benchmark datasets for full waveform inversion. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 6007–6020, 2022. doi: 10.52202/068431-0435.

N. Dimitrov, M. C. Kelly, A. Vignaroli, and J. Berg. From wind to loads: wind turbine site-specific load estimation with surrogate models trained on high-fidelity load databases. *Wind Energy Science*, 3(2):767–790, 2018. doi: 10.5194/wes-3-767-2018.

Nikolay Dimitrov and Tuhfe Göçmen. Virtual sensors for wind turbines with machine learning-based time series models. *Wind Energy*, 25(9):1626–1645, 2022. doi: 10.1002/we.2762.

DNV. Fatigue design of offshore steel structures. Recommended Practice DNV-RP-C203, DNV, Høvik, Norway, October 2024. URL https://www.dnv.com/energy/standards-guidelines/dnv-rp-c203-fatigue-design-of-offshore-steel-structures/. Edition October 2024.

S.D. Downing and D.F. Socie. Simple rainflow counting algorithms. *International Journal of Fatigue*, 4(1):31–40, 1982. ISSN 0142-1123. doi: 10.1016/0142-1123(82)90018-4.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. doi: 10.1214/aos/1176344552.

Charles R. Farrar and Keith Worden. An introduction to structural health monitoring. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365 (1851):303–315, 2007. doi: 10.1098/rsta.2006.1928.

Jean-Christophe Gagnon-Audet, Kartik Ahuja, Mohammad-Javad Darvishi-Bayazi, Pooneh Mousavi, Guillaume Dumas, and Irina Rish. WOODS: Benchmarks for out-of-distribution generalization in time series. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=mvftzofTYQ.

P. Gardner, L. A. Bull, J. Gosliga, N. Dervilis, and K. Worden. Foundations of population-based SHM, Part III: Heterogeneous populations – mapping and transfer. *Mechanical Systems and Signal Processing*, 149:107142, 2021. doi: 10.1016/j.ymssp.2020.107142.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021. doi: 10.1145/3458723.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007. doi: 10.1017/CBO9780511790942.

Rakshitha Godahewa, Christoph Bergmeir, Geoffrey I. Webb, Rob J. Hyndman, and Pablo Montero-Manso. Monash time series forecasting archive. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS 2021)*, 2021. URL https://openreview.net/forum?id=wEc1mgAjU-.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 16115–16152, 2024. URL https://proceedings.mlr.press/v235/goswami24a.html.

Moritz Gräfe, Vasilis Pettas, Nikolay Dimitrov, and Po Wen Cheng. Machine-learning-based virtual load sensors for mooring lines using simulated motion and lidar measurements. *Wind Energy Science*, 9(11):2175–2193, 2024. doi: 10.5194/wes-9-2175-2024.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=tEYskw1VY2.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983, 2022.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations (ICLR)*, 2017.

R. Haghi and C. Crawford. Data-driven surrogate model for wind turbine damage equivalent load. *Wind Energy Science*, 9(11):2039–2062, 2024. doi: 10.5194/wes-9-2039-2024.

Nandar Hlaing, Pablo G. Morato, Francisco de Nolasco Santos, Wout Weijtjens, Christof Devriendt, and Philippe Rigo. Farm-wide virtual load monitoring for offshore wind structures via Bayesian neural networks. *Structural Health Monitoring*, 23(3):1641–1663, 2024. doi: 10.1177/14759217231186048.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

IEC TS 61400-3-2:2019. Wind energy generation systems – part 3-2: Design requirements for floating offshore wind turbines. Technical Specification IEC TS 61400-3-2:2019, International Electrotechnical Commission, Geneva, CH, 2019. URL https://webstore.iec.ch/en/publication/29244.

Alexandros Iliopoulos, Rasoul Shirzadeh, Wout Weijtjens, Patrick Guillaume, Danny Van Hemelrijck, and Christof Devriendt. A modal decomposition and expansion approach for prediction of dynamic responses on a monopile offshore wind turbine using a limited number of vibration sensors. *Mechanical Systems and Signal Processing*, 68-69:84–104, 2016. doi: 10.1016/j.ymssp.2015.07.016.

International Energy Agency. Offshore wind outlook 2019. Technical report, IEA, Paris, 2019. URL https://www.iea.org/reports/offshore-wind-outlook-2019.

B. J. Jonkman and M. L. Buhl, Jr. TurbSim user's guide. Technical Report NREL/TP-500-39797, National Renewable Energy Laboratory, Golden, CO, USA, September 2006. URL https://docs.nrel.gov/docs/fy06osti/39797.pdf.

Jason Jonkman. The new modularization framework for the FAST wind turbine CAE tool. In *51st AIAA Aerospace Sciences Meeting including the New Horizons Forum and Aerospace Exposition*, 2013. doi: 10.2514/6.2013-202. Software: OpenFAST v3.5, NREL, 2024, https://github.com/OpenFAST/openfast.

Michael G. Kapteyn, Jacob V. R. Pretorius, and Karen E. Willcox. A probabilistic graphical model foundation for enabling predictive digital twins at scale. *Nature Computational Science*, 1(5): 337–347, 2021. doi: 10.1038/s43588-021-00069-0.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 5637–5664, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Milad Leyli-abadi, Antoine Marot, Jérôme Picault, David Danan, Mouadh Yagoubi, Benjamin Donnot, Seif Attoui, Pavel Dimitrov, Asma Farjallah, and Clement Etienam. LIPS – learning industrial physical simulation benchmark suite. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 28095–28109, 2022. doi: 10.52202/068431-2037.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Ding Peng Liu, Giulio Ferri, Taemin Heo, Enzo Marino, and Lance Manuel. On long-term fatigue damage estimation for a floating offshore wind turbine using a surrogate model. *Renewable Energy*, 225:120238, 2024a. ISSN 0960-1481. doi: 10.1016/j.renene.2024.120238.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024b.

Kristof Maes, Alexandros Iliopoulos, Wout Weijtjens, Christof Devriendt, and Geert Lombaert. Dynamic strain estimation for fatigue assessment of an offshore monopile wind turbine using filtering and modal expansion algorithms. *Mechanical Systems and Signal Processing*, 76-77: 592–611, 2016. doi: 10.1016/j.ymssp.2016.01.004.

M. Matsuishi and T. Endo. Fatigue of metals subjected to varying stress. Presented at the Japan Society of Mechanical Engineers, Fukuoka, Japan, 1968.

Kolja Müller and Po Wen Cheng. Validation of uncertainty in IEC damage calculations based on measurements from Alpha Ventus. *Energy Procedia*, 94:133–145, 2016. doi: 10.1016/j.egypro.2016.09.208. 13th Deep Sea Offshore Wind R&D Conference, EERA DeepWind'2016.

Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 37, pp. 43715–43729, 2024.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

David A. Nix and Andreas S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*, volume 1, pp. 55–60, 1994. doi: 10.1109/ICNN.1994.374138.

N. Noppe, A. Iliopoulos, W. Weijtjens, and C. Devriendt. Full load estimation of an offshore wind turbine based on SCADA and accelerometer data. *Journal of Physics: Conference Series*, 753: 072025, 2016. doi: 10.1088/1742-6596/753/7/072025. The Science of Making Torque from Wind (TORQUE 2016).

Ruben Ohana, Michael McCabe, Lucas Thibaut Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond Buschman Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich Kerswell, Suryanarayana Maddu, Jonah M. Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles Cranmer, and Shirley Ho. The Well: A large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, pp. 44989–45037, 2024. doi: 10.52202/079017-1430. URL https://neurips.cc/virtual/2024/poster/97882.

Francesco Papi and Alessandro Bianchini. 50 years of load simulation of the NREL 5MW OC4 floating wind turbine. Zenodo, 2024. Dataset, v1.0.0.

Francesco Papi, Robert Behrens de Luna, Joseph Saverin, David Marten, Cyril Compbreau, Gerardo Mirra, Giancarlo Troise, and Alessandro Bianchini. Deliverable 2.3: Design load case database for code-to-code comparison. Zenodo, 2023. FLOATECH project deliverable, v5.0.0.

Francisco Pimenta, Daniel Ribeiro, Adela Román, and Filipe Magalhães. Predictive model for fatigue evaluation of floating wind turbines validated with experimental data. *Renewable Energy*, 223:119981, 2024. doi: 10.1016/j.renene.2024.119981.

Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. ADATIME: A benchmarking suite for domain adaptation on time series data. *ACM Transactions on Knowledge Discovery from Data*, 17(8):1–18, 2023. doi: 10.1145/3587937.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Chao Ren and Yihan Xing. AK-MDAmax: Maximum fatigue damage assessment of wind turbine towers considering multi-location with an active learning approach. *Renewable Energy*, 215: 118977, 2023. doi: 10.1016/j.renene.2023.118977.

Niklas Requate and Tobias Meyer. Database of short term damage equivalent loads (DEL) of IWT7.5MW wind turbine depending on wind, TI, yaw and derating. Zenodo, 2023. Dataset.

João Alves Ribeiro, Francisco Pimenta, Bruno Alves Ribeiro, Sérgio M. O. Tavares, and Faez Ahmed. FLOAT: Fatigue-aware design optimization of floating offshore wind turbine towers. *Ocean Engineering*, 362:126285, 2026. doi: 10.1016/j.oceaneng.2026.126285. arXiv:2601.01657.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241, 2015. doi: 10.1007/978-3-319-24574-4_28.

D. Singh, R. Dwight, and A. Viré. Probabilistic surrogate modeling of damage equivalent loads on onshore and offshore wind turbines using mixture density networks. *Wind Energy Science*, 9(10): 1885–1904, 2024. doi: 10.5194/wes-9-1885-2024.

D. Singh, E. Haugen, K. Laugesen, R. P. Dwight, and A. Viré. Data-driven probabilistic surrogate model for floating wind turbine lifetime damage equivalent load prediction. *Wind Energy Science*, 10(12):2865–2888, 2025. doi: 10.5194/wes-10-2865-2025.

René M.M. Slot, John D. Sørensen, Bruno Sudret, Lasse Svenningsen, and Morten L. Thøgersen. Surrogate model uncertainty in wind turbine reliability assessment. *Renewable Energy*, 151: 1150–1162, 2020. ISSN 0960-1481. doi: 10.1016/j.renene.2019.11.101.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 1596–1611, 2022. doi: 10.52202/068431-0117.

Chang Wei Tan, Christoph Bergmeir, François Petitjean, and Geoffrey I. Webb. Time series extrinsic regression: Predicting numeric values from time series data. *Data Mining and Knowledge Discovery*, 35(3):1032–1060, 2021. doi: 10.1007/s10618-021-00745-9.

Marius Tarpø, Sandro Amador, Evangelos Katsanos, Mattias Skog, Johan Gjødvad, and Rune Brincker. Data-driven virtual sensing and dynamic strain estimation for fatigue analysis of offshore wind turbine using principal component analysis. *Wind Energy*, 25(3):505–516, 2022. doi: 10.1002/we.2683.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(10):11642–11662, 2026. doi: 10.1109/TPAMI.2026.3690845.

Jared Willard, Xiaowei Jia, Shaoming Xu, Michael Steinbach, and Vipin Kumar. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys*, 55(4):66:1–66:37, 2022. doi: 10.1145/3514228.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. TimesNet: Temporal 2D-variation modeling for general time series analysis. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhijian Xu, Ailing Zeng, and Qiang Xu. FITS: Modeling time series with $10k$ parameters. In *International Conference on Learning Representations (ICLR)*, 2024.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. WildTime: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 10309–10324, 2022.

Frederik Zahle, Thanasis Barlas, Kenneth Lønbæk, Pietro Bortolotti, Daniel Zalkind, Lu Wang, Casper Labuschagne, Latha Sethuraman, and Garrett Barter. Definition of the IEA Wind 22-megawatt offshore reference wind turbine. DTU Wind Energy Report E-0243, Technical University of Denmark, Lyngby, Denmark, April 2024. Model repository: `https://github.com/IEAWindSystems/IEA-22-280-RWT`.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11121–11128, 2023. doi: 10.1609/aaai.v37i9.26317.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary Ross, Katherine Bouman, and Yisong Yue. InverseBench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=U3PBITXNG6`.

TABLE OF CONTENTS FOR APPENDICES

A Reproducibility statement 17
B URL and links 17
C Author statement and data license confirmation 17
D Broader impact 18
E Dataset details 19
E.1 Tower geometry, damage and natural frequencies (Table 3, Figure 5)
E.2 Regime-aware partition (Table 4, Figure 6)
E.3 Schema reference (Table 5, Table 6)
F Physics baseline in detail 23
Four steps: gravity correction, gain, calibration (Figure 7, Figure 8), fit to the data, application along the tower; differences
G Benchmark details 27
G.1 Learned models (Table 9)
G.2 Training cost
G.3 Metric definitions
G.4 Why a cluster bootstrap? Confidence intervals over operating points
G.5 Model selection on a held-out validation split
H Leaderboards (within tower, cross tower, sensor sets) 34
H.1 Within tower, per tower
H.2 Training-seed spread
H.3 Cross tower, per pair
H.4 Sensor sets
H.5 Paired comparisons, intervals and exclusion audit
I Does the ranking survive along the tower? (within tower) 39
I.1 The base saturates in log-damage R^2; the top separates the models
I.2 A good top R^2 can hide the wrong realization
I.3 Over the DLC 1.2 design life the upper tower governs the fore-aft damage, and it is under-predicted
I.4 At the top, low wind is the hard condition
I.5 Waves matter far less than wind and height
J Is the top limited by information or by training? 46
J.1 Longer training closes most of the gap for the convolutional models
J.2 Training for the top alone, or on a damage loss, does the same
J.3 The failing models keep too little variance; adding it back fixes the level, not the realization
J.4 The missing variance is the 3P band below rated wind
K How should the top metrics and the transfer be read? (controls) 49
K.1 A condition-only predictor wins the top in aggregate R^2 (Table 30)
K.2 At the top every model attenuates the within-condition deviation (Table 31)
K.3 The lifetime biases are not sampling noise (Table 32)
K.4 One scale factor per height from five target simulations recovers the top (Table 33)
K.5 The top damage ranking is not a waveform ranking (Table 35)
K.6 Sensor errors do not change the base or the top (Table 36)
L Does the winner survive a change of tower? (cross tower) 53
L.1 Zero-shot keeps the ranking and misses the level
L.2 Fine-tuning on five target simulations recovers the base; a scale per height does better at the top
M What does each sensor add? (sensor ablation) 57
M.1 The accelerometer carries the realization; SCADA carries the level
M.2 The second axis helps the top; platform motions add little beyond it
M.3 Longer training keeps the order of the input sets
M.4 Generator power adds nothing, and field SCADA keeps the conclusions
N Datasheet for Datasets 60
O AI use statement 65

A REPRODUCIBILITY STATEMENT

We provide a repository to facilitate the reproduction of all experiments in the main paper, along with links to download the dataset. The repository includes code to replicate the benchmark experiments and code to generate the figures presented in the paper. During review, the subset and checkpoints of Appendix B reproduce the per-simulation damage of the learned models on 141 test simulations and the within-condition correlation on five operating points; the full dataset is released on publication. The training and experiments were performed on a machine equipped with an Intel Core i9-14900K processor (24 cores, 32 threads), 128 GB RAM, and two NVIDIA GeForce RTX 4090 GPUs (24 GB VRAM each). The simulation campaign required approximately 19,400 CPU-hours across the three towers, about 2.2 core-years on a commercial cloud HPC platform.

B URL AND LINKS

Project page. An anonymized project page gathers the links below, an interactive dataset explorer, the leaderboards and instructions to run the released checkpoints on the review subset, `https://anonymous.4open.science/w/FLOATSense-56B2/review/index.html`.

FLOATSense dataset. During review a fixed subset is available at an anonymized repository, `https://osf.io/h54t6/?view_only=73f55c8d86214fe1b943ede5260ddf4e`: 47 test simulations per tower, the same on the three towers, all at the middle of the wave grid (H_s and T_p index 4 of 7): the first realization at each wind speed ($\texttt{sim_id} = 25 + 294k$, $k = 0, \ldots, 21$) and all six realizations of the operating points at 5.5, 10.5, 14.5, 21.5 and 24.5 m/s. Series are in the release format, the tabular files are complete, and the seed-0 checkpoints of the 13 learned models trained from scratch and of the one-gain floor are included (fine-tuned foundation models left out for size). The subset supports evaluation, not training; a script recomputes ρ_{wc} on its five six-realization operating points (the benchmark value uses all 790 per tower). The full dataset, with all six realizations of every operating point, will be hosted on Hugging Face under CC-BY-4.0 on publication and is available to reviewers on request through OpenReview, which keeps both sides anonymous.

Tower geometry artifacts. The baseline IEA-22-280-RWT reference turbine is available at `https://github.com/IEAWindSystems/IEA-22-280-RWT`. The FLOAT-derived re-designs OPT1 and OPT2 are released with Ribeiro et al. (2026).

Croissant metadata. A Croissant 1.1 metadata record documenting the dataset will accompany it on Hugging Face.

Benchmark code. During review the source code for model training, evaluation, and figure generation is available from an anonymized repository `https://anonymous.4open.science/r/FLOATSense-FEB2/`; it is released under the MIT license. A script in the code (`scripts/download`) downloads the review subset from the link above and checks its integrity.

C AUTHOR STATEMENT AND DATA LICENSE CONFIRMATION

We, the authors, hereby declare that we bear full responsibility for any violations of rights or other issues arising from the use of the data. We confirm that the FLOATSense dataset is released under the Creative Commons Attribution 4.0 International license (CC-BY-4.0), which allows others to share and adapt the work for any purpose, including commercial use, provided that appropriate credit is given. The benchmark code is released under the MIT license.

Third-party assets and their licenses. The existing assets used in this work are released under permissive open-source licenses, all compatible with the redistribution terms of FLOATSense: IEA-22-280-RWT (Zahle et al., 2024) (Apache-2.0); OpenFAST (Jonkman, 2013) and TurbSim (Jonkman & Buhl, 2006) (Apache-2.0); PyTorch (BSD-3-Clause), NumPy (BSD-3-Clause), SciPy (BSD-3-Clause) and PyArrow (Apache-2.0); the pretrained encoders Chronos (Ansari et al., 2024) (Apache-2.0), MOMENT (Goswami et al., 2024) (MIT) and TimesFM (Das et al., 2024) (Apache-2.0); the reference implementations of S4D (Gu et al., 2022) (Apache-2.0), Mamba (Gu & Dao, 2024) (Apache-2.0), PatchTST (Nie et al., 2023) (Apache-2.0) and the Fourier neural operator (Li et al., 2021) (MIT); the Time Series Library (Wang et al., 2026) (MIT), which provides the TimesNet, iTransformer and DLinear baselines; and the `rainflow` package (MIT), which provides the cycle counting. All licenses have been reviewed and respected.

D BROADER IMPACT

Tower fatigue is a major design and lifetime constraint for 22 MW floating wind turbines operating in deep water, yet the bending moment that drives it is measured, if at all, by one or a few strain gauges, usually at the base, and almost never along the tower. To our knowledge, FLOATSense is the first public FOWT sensor-to-load benchmark combining multi-height fatigue targets, repeated realizations, regime and structural shifts, varying sensor configurations and a mechanistic comparator. Its impact spans several research communities:

1. **Floating wind monitoring and digital twins:** Loads reconstructed from accelerometers and SCADA could extend fatigue tracking to sections without a gauge, the load at every section that a structural digital twin needs, supporting inspection planning and life extension.
2. **ML for inverse physical time series:** Six realizations per operating point test whether a model reconstructs the realization or only recognizes the condition, making FLOATSense a testbed for new inverse time-series methods.
3. **Benchmarking under deployment shifts:** Regime extrapolation, cross-tower transfer and accuracy along the tower expose failures that a single score at the instrumented section hides.
4. **Sensor design:** With accelerations, SCADA, platform motions and wave elevation released, engineers can test sensor combinations in simulation and choose the best before installation.
5. **Environmental impact:** Load monitoring from existing sensors could inform maintenance and life-extension decisions, with potential benefits for floating-wind operating costs and deep-water renewable energy.
6. **Safety and certification support:** FLOATSense tests virtual sensors in simulation before deployment, exposing unsafe errors such as under-predicted fatigue damage and testing robustness across regimes, towers and heights. Once validated in the field, such virtual sensors could track consumed fatigue life from measured operation, supporting inspection and life-extension decisions beyond design assumptions.
7. **Cross-disciplinary insights:** FLOATSense sits at the intersection of structural health monitoring, machine learning and floating wind operation: learned reconstructions are scored with the fatigue damage engineers use and against a physics baseline, which encourages cross-disciplinary research and collaboration.
8. **Foundation for follow-up ML research:** FLOATSense releases 717,948 channel time series from the 19,404 OpenFAST runs of FLOATBench, to be extended with FEA and CFD simulations (Section 6). Beyond load reconstruction, they support probabilistic and multi-output models, neural operators, multi-fidelity learning, simulation-to-field adaptation, and hybrid models that pair the time series with the tabular data of FLOATBench.

Beyond floating wind, FLOATSense is a template for inverse monitoring of any partially observed structure: repeated realizations separate condition recognition from reconstruction, spatial targets expose location-dependent failures, and regime and structure shifts test whether performance survives deployment. We expect this template to inform virtual sensing and structural health monitoring in adjacent engineering domains, such as bridges, offshore platforms and aircraft, where accuracy away from the sensors and under realistic deployment shifts is what matters operationally.

E DATASET DETAILS

E.1 TOWER GEOMETRY, DAMAGE AND NATURAL FREQUENCIES

Tower geometry. The three towers are those of FLOATBench (Anonymous, 2026), designed in Ribeiro et al. (2026). They share the height (149.4 m), the platform, the controller and the mooring and differ in the diameter and wall-thickness profiles (Table 3 and Figure 5): REF is the IEA 22 MW baseline designed for a fixed-bottom foundation, OPT1 and OPT2 are the fatigue-aware redesigns of Ribeiro et al. (2026). Both redesigns thicken toward the base and taper toward the top; OPT2 distributes material slightly more evenly than OPT1 (thinner base, wider and thicker top).

Damage profile. The lifetime damage of REF (Table 3) falls by an order of magnitude from the base (33) to the top (3.2). The redesigns cut the base damage about 41-fold and flatten the profile, so that on both the top exceeds the base (1.21 against 0.80 on OPT1, 0.88 against 0.82 on OPT2).

Natural frequencies. On the floating platform the first fore-aft mode of REF lies at 0.34 Hz (Table 3), close to the 3P excitation, the blade-passing frequency of the three-bladed rotor, at rated rotor speed, which can lead to resonance and amplified fatigue loading; the redesigns move it to 0.57 and 0.54 Hz, above the 3P band.

Table 3: Tower geometry, mass, natural frequency and damage of the three towers, from FLOATBench, with the change relative to REF. Frequency: first fore-aft mode on the floating platform. Damage: lifetime-weighted Miner sum over all simulations (1 is the end of life), from the FLOATSense moments at the base and top gauges, with the mean outer radius and wall thickness of the nearest FLOATBench section.

Parameter	REF	OPT1	OPT2
Geometry			
Diameter, bottom [m]	10.000	12.000 (+20.0 %)	12.000 (+20.0 %)
Diameter, top [m]	6.000	6.424 (+7.1 %)	6.741 (+12.4 %)
Thickness, bottom [mm]	66	124 (+87.6 %)	118 (+78.2 %)
Thickness, top [mm]	38	43 (+13.1 %)	45 (+15.8 %)
Mass, dynamics and damage			
Total mass [t]	1,574	2,899 (+84.2 %)	2,656 (+68.8 %)
First fore-aft frequency, floating [Hz]	0.336	0.573 (+70.5 %)	0.537 (+59.8 %)
Lifetime damage at the base	33.3	0.80 (1/41)	0.82 (1/41)
Lifetime damage at the top	3.23	1.21 (1/2.7)	0.88 (1/3.7)

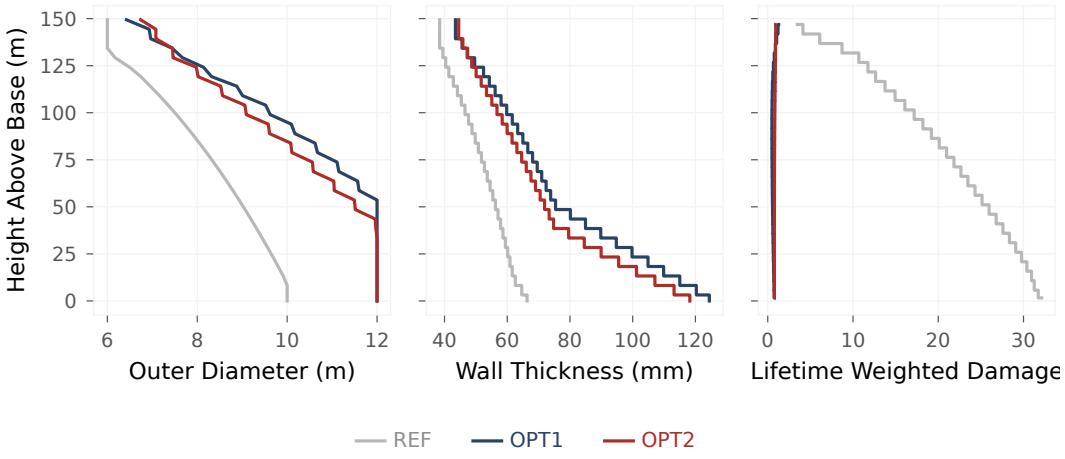


Figure 5: Tower geometry and damage along the height of the three towers, from FLOATBench. Diameter and thickness: outer diameter and wall thickness. Damage: lifetime-weighted Miner sum over all simulations.

E.2 REGIME-AWARE PARTITION

FLOATSense follows the regime-aware partition of FLOATBench (Anonymous, 2026), identical on the three towers (Section 4.2). Training keeps 18 of the 22 wind speeds and the inner 4×4 of the 7×7 sea states; every other operating point is test. Each test simulation is labeled on the wind and on the wave axis as in-train, interpolation or extrapolation, giving nine regime cells; we refer the reader to FLOATBench for the labeling algorithm and its thresholds.

Figure 6 shows the training domain and these labels, and Table 4 gives how the 4,740 test simulations fall into the nine cells. They are uneven by construction: the partition holds out the extreme and middle indices of both axes, so 2,118 test simulations extrapolate on waves against 539 on wind, and the worst case, EX_EX, holds 254.

Table 4: Test simulations per regime cell, identical on the three towers, from FLOATBench. Rows: wind label. Columns: wave label.

Wind \ Wave	IT	IP	EX	Total
IT	752	1,245	1,714	3,711
IP	164	176	150	490
EX	92	193	254	539
Total	1,008	1,614	2,118	4,740

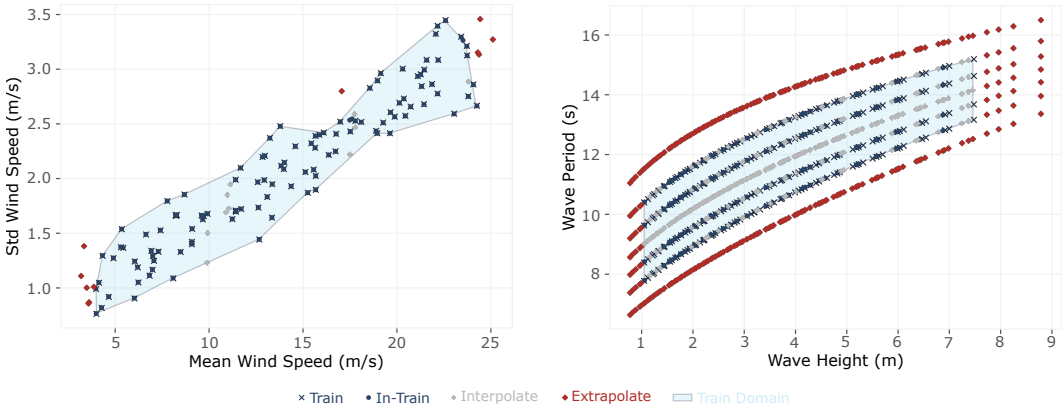


Figure 6: Regime-aware partition from FLOATBench, in the wind (left) and wave (right) subspaces. The training domain is the shaded alpha-shape hull; every test simulation is labeled in-train, interpolation or extrapolation on each axis, giving nine regime cells.

E.3 SCHEMA REFERENCE

Channels. Every simulation carries 37 channels sampled at 10 Hz over 1,000 s (Table 5): the measurable signals (tower-top accelerations, SCADA, platform motions and wave elevation) and the simulator reference bending moments (base, nine intermediate sections and top).

Table 5: The 37 channels of every simulation, with the number of series per tower and in total. Task inputs and target: Section 4.1.

			Series	
Group	Channels	Unit	per tower	total
Tower-top accel.	fore-aft, side-side; raw and gravity-corrected (4)	m/s^2	26 k	78 k
SCADA	rotor speed, blade pitch, power, hub wind speed (4)	rpm, deg, kW, m/s	26 k	78 k
Platform motion	surge, sway, heave, roll, pitch, yaw (6)	m, deg	39 k	116 k
Wave elevation	at the platform (1)	m	6 k	19 k
Fore-aft moment	base, nine intermediate sections and top (11)	kN m	71 k	213 k
Side-side moment	base, nine intermediate sections and top (11)	kN m	71 k	213 k
Total	37 channels × 6,468 simulations per tower		239 k	718 k

Files. Each tower is one directory with the five files of Table 6, and one `parked.parquet` is shared by the three. Every file is keyed by `sim_id`, the FLOATBench simulation identifier, the same on the three towers, so a run can be matched to its FLOATBench row.

- *Time series.* Parquet format, compressed and columnar, read directly by pandas, PyArrow and Hugging Face `datasets`, with one row per time step: `sim_id`; `time_s`, the time in seconds; and the 37 channels. A tower has 6,468 × 10,001 = 64,686,468 rows, about 8.0 GB (24.0 GB for the three), split into 17 shards of about 0.50 GB so that they can be downloaded in parts and read in parallel: 400 simulations per shard, the last holding 68, each simulation one row group so that it is read without scanning the shard. `series_stats.parquet` gives the mean, standard deviation, minimum and maximum of every channel over the 400–1,000 s scored window.
- *Labels.* `damage.parquet` gives, one row per simulation and section as in FLOATBench, the reference fore-aft damage at the 11 target heights, computed as in the evaluation, and `sections.parquet`, for each height, the nearest of the 30 FLOATBench sections with its FLOATBench height, mean outer radius and wall thickness, used for the stress, and the height of the gauge, so a predicted moment can be scored with the released code.
- *Operating point, split and weight.* `metadata.parquet` carries, per simulation, the three grid indices that identify its operating point, which of the six realizations it is, the physical wind and wave values, the train/test split, the wind and wave regime labels, and the lifetime weight: the number of ten-minute periods its operating point is assigned over a 25-year life (the 6,468 weights sum to 1,314,000).
- *Parked runs.* The 22 parked simulations of each tower, under waves only, used to calibrate the physics baseline (Appendix F).

The channel names are the column names of the Parquet files: `tower_top_afa` and `tower_top_ass` for the raw accelerations, the same with `_mod` for their gravity-corrected counterparts (the correction of Appendix F), and `<section>_mfa` or `<section>_mss` for the moments, the section running `tower_bottom`, `tower_1` to `tower_9`, `tower_top` from the base upward (FLOATBench sections 1, 3, 6, . . . , 27, 30); the dataset card pairs each name with its OpenFAST column and documents axis directions, positive signs and the mapping to OpenFAST coordinates. Channels and `time_s` are stored as `float32` in the units of Table 5, and `sim_id` as `int32`.

Table 6: Files released per tower and their columns. One folder per tower, so no tower column is needed; the grid indices identify the operating point and `wind_seed_id` the realization, as in FLOATBench, whose column names and order the files follow.

File	Column	Type	Description
Time series			
`series-<k>-of-00017.parquet`	`sim_id`	int32	FLOATBench simulation identifier, the same on the three towers
	`time_s`	float32	time (s), 0 to 1,000 at 10 Hz
	`<channel>`	float32	the 37 channels of Table 5; 400 simulations per shard, one row group each
`series_stats.parquet`	`sim_id`	int	simulation identifier
	`<channel>_{mean,std, min,max}`	float	statistics over the 400–1,000 s scored window
Labels			
`damage.parquet`	`sim_id`	int	simulation identifier
	`section_id`	int	FLOATBench section, 11 of the 30: 1, 3, 6, . . . , 27, 30
	`damage`	float	reference fore-aft damage at that section, with its mean outer radius and wall thickness; 11 rows per simulation
Operating point, split and weight			
`metadata.parquet`	`sim_id`	int	simulation identifier
	`wind_speed_id`, `wave_hs_id`, `wave_tp_id`	int	grid indices on the $22 \times 7 \times 7$ envelope; the three together are the operating point
	`wind_seed_id`	int	which of the six realizations of that point, 1 to 6
	`wind_speed`, `wave_hs`, `wave_tp`	float	nominal hub-height wind speed (m/s), significant wave height (m), peak period (s)
	`mean_wind_speed`, `std_wind_speed`	float	realized mean and standard deviation of the hub-height wind (m/s), as in FLOATBench
	`split`	str	`train` (1,728) or `test` (4,740)
	`wind_group`, `wave_group`	str	regime label per axis: In-train, Interpolate, Extrapolate
	`damage_weight`	float	ten-minute periods of the operating point over a 25-year life
Tower geometry			
`sections.parquet`	`section_id`	int	FLOATBench section, 1, 3, 6, . . . , 27, 30
	`section_height_m`	float	FLOATBench section midpoint above the tower base (m)
	`section_radius_m`, `section_thickness_m`	float	mean outer radius and wall thickness of that section, used for the stress (m); these four columns are those of FLOATBench
	`channel`	str	prefix of its moment channels, `tower_bottom` to `tower_top`
	`gauge_height_m`, `z_over_h`	float	height of the gauge (m) and the same as a fraction of the tower height, the model input
Parked runs, one file for the three towers			
`parked.parquet`	`tower`, `run`	str	tower and run name; parked settings in the file metadata
	`time_s`	float32	time (s), 0 to 1,000 at 10 Hz
	`<channel>`	float32	28 of the 37 channels of Table 5

F PHYSICS BASELINE IN DETAIL

The physics baseline reconstructs the bending moment from the tower-top acceleration with the frequency-band method of Pimenta et al. (2024). The method was developed for one section of smaller floating turbines (the NREL 5 MW model on its reference floating platform, and the 8.4 MW turbines of the WindFloat Atlantic floating farm); we adapt it to this 22 MW turbine and extend it to every height of the tower. The baseline evaluated here is therefore an adaptation of a field-validated method, not the validated method itself: the original was validated at one section, while the height factor and the changes listed at the end are ours and are evaluated only against the simulations of this paper. This appendix describes the method with these changes, which are listed at the end; the derivation is in the original paper.

The tower is a tall beam clamped at its base, with the rotor and nacelle (1,218 t) on top. When waves move the platform, the top accelerates and the inertia of that mass bends the tower, so the base moment is, to first order, mass times lever arm times acceleration. Wind adds loads that do not follow this ratio, but at frequencies that can be predicted. The method therefore multiplies the spectrum of the tower-top acceleration by a gain whose form depends on the frequency. FA (fore-aft) is the direction along the wind and SS (side-side) the direction across it; both are treated alike.

The method runs in four steps:

1. Correct the acceleration for gravity.
2. Define the gain $G(f)$ that turns its spectrum into the moment.
3. Calibrate G: C_1 on parked runs (no wind, rotor stopped), the rest on training runs.
4. Apply G at the base and scale the result to each height of the tower.

After step 3 we check the calibrated gain against the data; the last paragraph lists where we depart from the published method.

Step 1: acceleration with the gravity projection. A real accelerometer at the tower top tilts with the platform and the bending tower, and so senses part of gravity; the method assumes this (Eq. 10 of Pimenta et al. (2024)). OpenFAST outputs only the kinematic acceleration, so we add the projection:

$$a_{\mathrm{FA}} = \ddot{x} - g\sin(\theta_{\mathrm{pitch}} + \theta_{\mathrm{FA}}), \qquad a_{\mathrm{SS}} = \ddot{y} + g\sin(\theta_{\mathrm{roll}} + \theta_{\mathrm{SS}}), \tag{1}$$

with $\ddot{x}$, $\ddot{y}$ the kinematic accelerations, θ_{pitch}, θ_{roll} the platform rotations and θ_{FA}, θ_{SS} the bending rotations of the tower top (signs follow the OpenFAST axes). This correction is not specific to the baseline: the corrected channels are released with the dataset (`_mod` channels, Table 5), and the corrected fore-aft acceleration is the input of the physics baseline and of every learned model.

Step 2: the gain. The reconstructed base moment is $\tilde{M}(f) = G(f)\,\tilde{a}(f)$, with $\tilde{a}$ the Fourier transform of the corrected acceleration, $\omega = 2\pi f$ and

$$G(f) = \begin{cases} C_1 & \text{background (most frequencies),} \\ C_1\,(C_4 + C_5\,\omega^2) & \text{slow wind response, } 0 < f \le f_{\mathrm{lim}}, \\ C_1/(C_2 + C_3\,\omega^2 - i\,C_d) & \text{rotor loads, in the 3P, 6P and 9P windows.} \end{cases} \tag{2}$$

where f_{lim} is the frequency at which the slow branch meets C_1 ($C_4 + C_5\omega^2 = 1$); it follows from the calibrated C_4, C_5 and falls at the platform pitch frequency, about 0.04 Hz (Eq. 18 of Pimenta et al. (2024)). Each range has one dominant load:

- *Background.* Waves move the platform and the moment is purely inertial; C_1 (MN s^2) is the mass times lever arm of the ratio above (Eq. 21a of Pimenta et al. (2024)).
- *Slow wind response.* Slowly varying wind thrust pitches the platform and bends the tower while the top barely accelerates, so the gain is above C_1 ($C_4 > 1$).
- *Rotor loads.* Each of the three blades sweeps the same wind shear, turbulence and tower wake once per revolution, so the rotor loads the tower at three times its rotation frequency (3P) and at 6P and 9P. The moment follows the tower's response to a force at its top (Eq. 24 of Pimenta et al. (2024)), plus our damping term C_d. The windows are $n\,[\Omega_{\min}, \Omega_{\max}]/60$, $n \in \{3, 6, 9\}$, with Ω the rotor speed (rpm) while the turbine generates power, so they move with the wind (Figure 7).

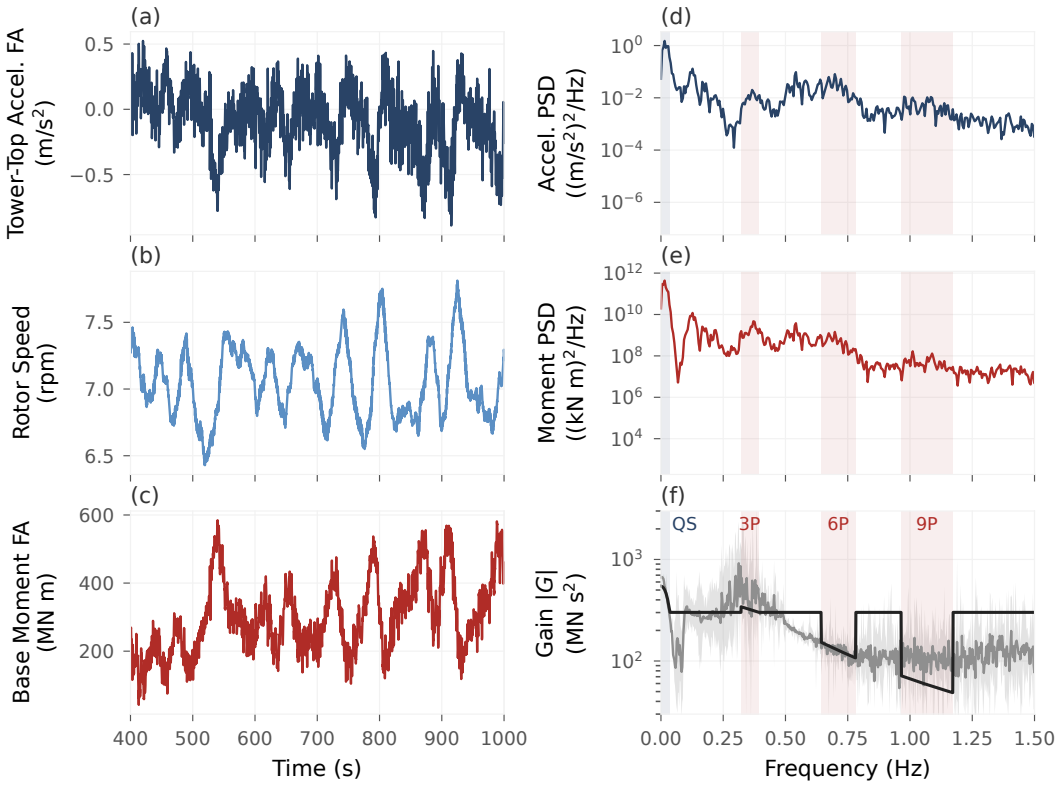


Figure 7: One simulation of OPT2 at 12.5 m/s and H_s 1.8 m, fore-aft, over the scored window. (a) Corrected tower-top acceleration and (b) rotor speed, the inputs; (c) base moment, the target. (d), (e) Power spectral densities (PSD) of acceleration, S_{aa}, and moment, S_{MM}. (f) Calibrated gain $|G(f)|$ (black) against the measured gain $\sqrt{S_{MM}/S_{aa}}$ (grey: median and 10 to 90% band over the 96 training simulations at 12.5 m/s). Shaded: the slow wind band (QS) and the 3P, 6P and 9P windows.

Step 3: calibration. C_1 comes from parked runs (waves only, no wind), the other five constants from the training split; no test simulation is used.

- C_1, *calibrated on 22 parked runs per tower.* C_1 belongs to the structure, so it is measured where the background load acts alone: no wind, rotor stopped, blades turned edge-on, waves only, one run at the median sea state of each wind speed. The learned models do not see these runs. C_1 is the least-squares slope of the standard deviation of the moment against that of the acceleration, with a free intercept (Figure 8a). The fit is tight: R^2 of 0.992 to 1.000, intercept below 4% of the largest moment standard deviation. The same slope measured at height z is written $C_1(z)$, so $C_1 = C_1(0)$.
- C_2 *to* C_5 *and* C_d*, calibrated on the 1,728 simulations of the training split of each tower* (18 wind speeds × 16 sea states × 6 seeds, Appendix E.2), with the turbine operating in wind and waves; the learned models train on the same split. Each simulation gives an estimate from the power spectral densities S_{aa} and S_{MM} of its acceleration and moment, and the median is kept. C_4, C_5 fit $(|M|/|a|)/C_1 = C_4 + C_5\omega^2$ between 0.01 and 0.05 Hz, on every simulation. C_2, C_3, C_d are solved from $C_1^2 S_{aa}/S_{MM} = (C_2 + C_3\omega^2)^2 + C_d^2$ at the peak of each of the three windows. These three constants use only the simulations where this fit has a solution ($n_{2,3}$ for C_2, C_3 and n_d for C_d in Table 7). Most of these simulations have wind below 9 m/s: such runs are 28% of the 1,728 but give most of the successful fits (47 to 96% for C_2, C_3 and 69 to 100% for C_d). So the rotor-load branch of the baseline is calibrated mostly on low-wind runs, but it is used at every wind speed, including 24.5 m/s, which is not in the training split.

Table 7 lists the six calibrated constants of each tower and direction.

Table 7: Calibrated constants of the physics baseline. C_1 comes from the parked runs; the others are medians over the training simulations of each tower; $n_{2,3}$ and n_d are the numbers of those simulations (of 1,728) whose fit gives C_2, C_3 and C_d. f_{lim} is the platform pitch frequency (periods of 25.0 s on REF, 28.3 to 29.1 s on the redesigns, fore-aft). The medians have the expected signs, $C_2 < 0$ and $C_3 > 0$.

Tower	Dir.	C_1 [MN s^2]	C_2 [–]	C_3 [s^2]	C_4 [–]	C_5 [s^2]	C_d [–]	f_{lim} [Hz]	$n_{2,3}$	n_d
REF	FA	229.5	−0.130	0.125	2.226	−19.48	0.922	0.0399	417	314
REF	SS	214.6	−0.645	0.364	2.328	−20.04	0.496	0.0410	501	308
OPT1	FA	312.8	−0.034	0.108	1.743	−15.94	0.719	0.0344	641	325
OPT1	SS	314.8	−0.740	0.156	1.673	−14.93	0.727	0.0338	991	440
OPT2	FA	301.2	−0.044	0.114	1.811	−16.45	0.773	0.0353	621	318
OPT2	SS	311.0	−1.108	0.217	1.694	−15.08	0.737	0.0341	710	474

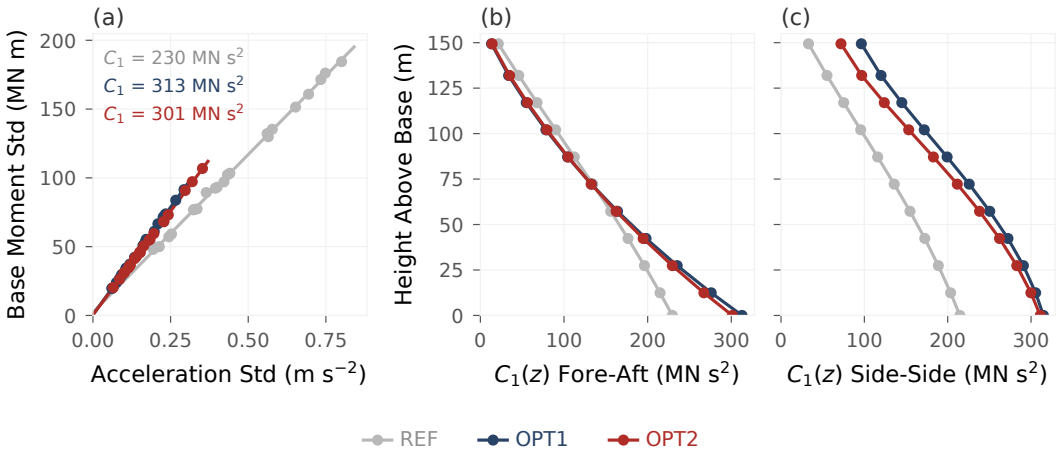


Figure 8: Parked calibration, 22 parked runs per tower. (a) Standard deviation of the base moment against that of the tower-top acceleration, fore-aft; the slope of each line is C_1. (b), (c) The same slope at each of the eleven heights, $C_1(z)$, fore-aft and side-side.

Does the gain match the data? *Closely in the slow band only; elsewhere within a factor of 3.5.* Over the 18 training wind speeds of the three towers (in-sample; Figure 7f shows one), the measured gain $\sqrt{S_{MM}/S_{aa}}$ is within 9% of the calibrated one in the slow band, 0.66 to 2.37 times it in the 3P, 6P and 9P windows, and mostly below C_1 outside the three ranges (0.29 to 1.29 times). We keep the three ranges of the published method; the spectral gain and the hybrid of Appendix G address the gaps, the hybrid only within its $\pm 50\%$ bound.

Step 4: application at the base and along the tower. At the base, the calibrated gain is applied by fast Fourier transform (FFT) to the record from 350 s, with its last 50 s mirrored, so the scored window (400–1,000 s) has a 50 s margin on each side against edge effects. The fore-aft sign follows Eq. (21a) of Pimenta et al. (2024). As for every model, both series are low-passed at 3 Hz before scoring.

A height factor with one fitted parameter. The published method covers one section. Higher up, less mass lies above the section and its lever arm is shorter, so the same acceleration produces a smaller moment. We therefore multiply the gain by a height factor $f(z)$, the moment of a rigid tower rotating about its base at height z relative to that at the base:

$$f(z) = \frac{Q(z)}{Q(0)}, \qquad Q(z) = M_{\mathrm{rna}}\,(H+h)\,(H+h-z) + \int_z^H m(s)\,s\,(s-z)\,\mathrm{d}s, \tag{3}$$

with z the height above the base, H the tower height, $m(s)$ the tower mass per unit length, M_{rna} the rotor and nacelle mass and h the height of their center of mass above the tower top. We fit h so that the height factor matches the parked slope at the top, $C_1(H)/C_1$. The fitted h (19.7 m on REF, about 10 m on the redesigns, against 4.6 m in the design) is an effective value that absorbs what the

rigid model leaves out. The resulting curve passes within 0.07 of the parked ratios $C_1(z)/C_1$ at all eleven heights.

Alternatives. Table 8 compares three ways to extend the gain, calibrated on the training simulations and scored on the test ones by how high the reconstruction stays acceptable:

1. *Height factor $f(z)$ above:* the parked C_1 at the base and the top and the tower mass.
2. *Parked $C_1(z)$ at every height:* the parked C_1 at all eleven heights, with the other constants of the base; the direct use of the published recipe.
3. *All constants calibrated at every height:* the parked C_1 and the training moments at every height.

Table 8: Ways to extend the gain along the tower. Highest z/H up to which at least 80% of the 4,740 test simulations have a damage within a factor of two of the true one.

Extension	REF	OPT1	OPT2	Needs
1. Height factor $f(z)$ (used)	0.78	0.68	0.68	$C_1(z)$ base and top, $m(s)$; training moments at the base
2. Parked $C_1(z)$ at every height	0.28	0.48	0.58	$C_1(z)$ at every height; training moments at the base
3. All constants calibrated at every height	0.58	0.88	0.78	$C_1(z)$ and training moments at every height
Best constant factor per height	0.88	0.78	0.78	training moments at every height

Option 1 is the best on REF (z/H 0.78) and option 3 on OPT1 and OPT2 (0.88 and 0.78), but option 3 fails at mid tower on REF, so option 1 has the best worst case (0.68). Option 2 does worse despite nine more measurements, likely because the parked runs contain only the wave load. We use option 1 because it needs the least instrumentation, a base strain gauge and a parked simulation, as on turbines instrumented with a single strain gauge at the tower base, whereas option 3 needs a load reference at every height. Even option 3, with load labels at all eleven heights, reaches at the top log-damage R^2 0.91 / 0.52 / 0.33 with 66 / 17 / 13% within a factor of two (TCN: 0.78 / 0.77 / 0.73, 60 / 59 / 60%); it fixes the level (lifetime ratio 0.64 to 1.12) but not the realization (ρ_{wc} 0.37 to 0.58). A linear filter calibrated on the training split, $H(f) = S_{Ma}/S_{aa}$, reaches the physics at the base (log-damage R^2 0.95 to 0.995, ρ_{wc} 0.97 to 0.99) but keeps 11 to 16% of the top damage (R^2 0.16 to 0.48), because it shrinks each band by its coherence. No option holds at the top, where loads absent from the parked runs dominate: rotor thrust fluctuations, the gravity moment of the tilted rotor and nacelle, and first-mode bending. Even the best constant factor per height (last row: the median ratio, on training, of true to reconstructed moment standard deviation) stops at z/H 0.78 to 0.88, while the learned models, reading the same acceleration, do better at the top: the limit is the single-gain model, not the sensor.

Differences from the published method. The three-range gain, the estimation of the constants from power spectra, the placement of the windows on the generating rotor speed and the fatigue model (rainflow, DNV-RP-C203) follow Pimenta et al. (2024). We depart from it in six points:

(i) C_1 is the slope of standard deviations over dedicated parked simulations (no wind, waves only), not the median spectral ratio over parked periods of an operating turbine.
(ii) The gravity correction adds the bending rotation of the tower top to the platform rotation.
(iii) The rotor branch has a damping term C_d, identified with a third window, 9P: the rotor of a 22 MW turbine turns slowly (3P at 0.09 to 0.42 Hz), and the undamped form fitted on 3P and 6P crosses zero inside the 3P window of the redesigns. The fit uses magnitudes, so the sign of C_d (the phase lag) is our choice.
(iv) The rotor branch is fitted and applied in both directions; the paper fits it fore-aft only.
(v) C_4, C_5 are fitted to the amplitude ratio $|M|/|a|$, so their signs are opposite to those of the paper's signed constants.
(vi) The median over training simulations, the 0.01–0.05 Hz fit band, the 50 s margins and the extension along the tower are ours.

Pimenta et al. (2024) work on measurements from a real turbine, which is sometimes parked, starting up or shutting down, or operating normally, and they calibrate each state separately. In every simulation of our benchmark the turbine operates normally, so this step does not apply.

G BENCHMARK DETAILS

G.1 LEARNED MODELS

Table 9 lists the 20 learned models and the floor. All read the same five channels, the accelerometer, the three SCADA series and the height z/H, and write the standardized moment; the pretrained encoders read the accelerometer alone, and the gain models and the two hybrids also take the window means as a condition. Five of them (spectral gain, hybrid, Hybrid-TCN, FITS, iTransformer) are length-fixed and train on the full window, the rest on 4,096-sample crops. Nothing was tuned on this dataset, including the hybrid bound, which is a design constraint fixed a priori and never swept (Section 4.3). Each heading cites the published model it implements; the four without a citation are ours, the floor, the spectral gain and the two hybrids; Prob-TCN adds a Gaussian head to the TCN.

Table 9: The 20 learned models and the floor. Parameters in millions (trainable / total).

Model	Family	Params	Notes
NAIVE	floor	$<$0.001	one scalar gain, no conditioning
DLinear	convolutional	0.003	trend/remainder split as in Zeng et al. (2023), with channel-mixing convolutions in place of the per-channel linear maps (our variant)
TCN	convolutional	0.33	dilated residual blocks (Bai et al., 2018)
U-Net	convolutional	4.4	four-level multi-rate encoder-decoder (Ronneberger et al., 2015)
LSTM	recurrent	0.30	bidirectional (Hochreiter & Schmidhuber, 1997)
PatchTST	attention	0.94	transformer over strided patches (Nie et al., 2023)
iTransformer	attention	5.4	attention across channels (Liu et al., 2024b)
S4D	state-space	0.05	diagonal SSM, FFT convolution (Gu et al., 2022)
Mamba	state-space	0.18	selective SSM, chunked scan in plain PyTorch (Gu & Dao, 2024)
Spectral gain	spectral	0.40	condition-dependent complex gain per bin
FITS	spectral	5.9	complex linear map between low-frequency bins (Xu et al., 2024)
FNO	spectral	2.4	Fourier neural operator (Li et al., 2021)
TimesNet	period-folding	0.16	2D convolutions over folded periods (Wu et al., 2023)
Chronos	pretrained	0.15 / 21	frozen T5 encoder, trained head (Ansari et al., 2024)
MOMENT	pretrained	0.004 / 38	frozen patch encoder, trained head (Goswami et al., 2024)
MOMENT-ft	pretrained	38 / 38	the same, fine-tuned end to end
TimesFM	pretrained	0.25 / 232	frozen decoder-only backbone (2.5 checkpoint), trained patch head (Das et al., 2024)
TimesFM-ft	pretrained	232 / 232	the same, fine-tuned end to end
Hybrid	physics-anchored	0.40	$G_{\mathrm{phys}}(1+c)$, c tanh-bounded at ± 0.5
Hybrid-TCN	physics-anchored	0.33	$M_{\mathrm{phys}} + 0.5\,\sigma_{\mathrm{phys}}\tanh(\mathrm{TCN})$
Prob-TCN	convolutional	0.33	TCN with heteroscedastic head, Gaussian NLL (Nix & Weigend, 1994)

NAIVE. One scalar gain from acceleration to moment, $\hat{M}(t) = g\,a(t)$, with no conditioning and no height input. The gain is learned in standardized units with the recipe of the other models, so in physical units the reconstruction at a requested height z is

$$\hat{M}_z(t) = g\,\frac{\sigma_{M,z}}{\sigma_a}\,\big(a(t) - \mu_a\big),$$

with one learned g shared by all heights, μ_a and σ_a the mean and standard deviation of the gravity-corrected fore-aft acceleration, and $\sigma_{M,z}$ the standard deviation of the fore-aft moment at height z, all computed once over the scored window (400 to 1,000 s) of an evenly spaced sample of about 200 training simulations of the source tower (the target is centered per window, which leaves the damage unchanged). The floor is thus one learned gain times eleven fixed training ratios $\sigma_{M,z}/\sigma_a$. Every learned model is de-standardized the same way, so the height-blind pretrained encoders also receive $\sigma_{M,z}$ at the requested height.

DLinear (Zeng et al., 2023). The linear sanity check of the forecasting literature. Each input channel is split into a moving-average trend and its remainder and each part is mapped by its own convolution over all channels and a fixed context (the original uses linear maps over a fixed window; the convolution makes the model length-agnostic). Against the floor it is the same linear hypothesis with memory: a kernel over a context window rather than one scalar.

TCN (Bai et al., 2018). Dilated residual convolutional blocks with symmetric (non-causal) padding, the standard sequence-to-sequence convolutional network. Small (0.33M) and length-agnostic; the model that transfers best across towers.

U-Net (Ronneberger et al., 2015). A four-level multi-rate convolutional encoder-decoder with skip connections: the signal is downsampled four times and reconstructed, so each stage works at one time scale (the wave band, the tower modes, the rotor harmonics). The convolutional counterpart of the multi-rate stacks of N-HiTS (Challu et al., 2023).

LSTM (Hochreiter & Schmidhuber, 1997). Bidirectional LSTM over the five input channels with a linear output per sample.

PatchTST (Nie et al., 2023). A transformer encoder over non-overlapping strided patches of the input; a learned positional embedding is added and a linear head maps each encoded token back to its patch of moment samples.

iTransformer (Liu et al., 2024b). The inverted transformer: every input channel becomes one token holding its whole series, so attention models the relations between acceleration, wind, rotor speed and pitch rather than between time steps; a learned query token is decoded into the moment series. It diverges under the common recipe; at 3×10^{-4} it reaches R^2 0.26, -0.18 and -0.15 at the base of the three towers, no better, so the tables keep the common recipe.

S4D (Gu et al., 2022). A stack of diagonal state-space blocks with zero-order-hold discretization, computed as an FFT convolution. Each state is a damped oscillator, the hypothesis class closest to a linear structural response.

Mamba (Gu & Dao, 2024). The selective state-space model (S6), with the recurrence solved by a chunked scan in plain PyTorch instead of the fused CUDA kernel. It needs a learning rate of 3×10^{-4}; at 10^{-3} it diverges.

FITS (Xu et al., 2024). A single complex matrix between the low-frequency rFFT bins of every input channel and the output bins, everything above the cutoff dropped. Unlike the spectral gain the map is not diagonal and not conditioned; it isolates how much of the task is one fixed linear filter. The published model interpolates the spectrum for forecasting; here the same map is used for regression. Fixed length.

FNO (Li et al., 2021). The Fourier neural operator: layers that mix a truncated set of Fourier modes of the input channels, the standard operator-learning baseline and resolution independent.

TimesNet (Wu et al., 2023). Folds the series at its dominant periods into a 2D image whose rows are successive cycles and convolves it. The rotor harmonics and the wave period make the signal quasi-periodic, so this is the strongest published inductive bias for periodic series and the natural learned competitor of the band-based physics baseline.

Chronos (Ansari et al., 2024). Chronos-T5 (small) used as a frozen encoder in fixed context chunks; only a small convolutional head, two pointwise layers around a width-5 convolution, maps the token embeddings to the moment series. Chronos rescales each chunk internally, so the head also receives the chunk scale. Its embedding interface exposes no gradients, so a fine-tuned variant is not possible with the released model.

MOMENT and MOMENT-ft (Goswami et al., 2024). MOMENT is pretrained for masked reconstruction and maps a window to a window of the same length. MOMENT keeps the patch encoder frozen and trains the reconstruction head; MOMENT-ft fine-tunes all 38M parameters end to end.

TimesFM and TimesFM-ft (Das et al., 2024). TimesFM 2.5 (200M, decoder-only) as an encoder: each context window is normalized and cut into 32-sample patches, the 20-layer stack produces one embedding per patch and a small convolutional head writes the moment samples of that patch back, with the window scale as an extra channel. TimesFM trains the head only; TimesFM-ft fine-tunes the backbone.

Spectral gain. The learned analog of the physics baseline: a diagonal complex filter on the rFFT of the acceleration, one gain per frequency bin, set by a hypernetwork (Ha et al., 2017) from the operating state and the requested height. Fixed length.

Hybrid. Physics-anchored spectral gain, residual modeling in the sense of Willard et al. (2022): the reconstruction applies $G_{\mathrm{phys}}(f)\,(1 + c(f, \mathrm{condition}))$ to the rFFT of the acceleration, where G_{phys} is the calibrated band gain of the physics baseline for that simulation, scaled by the height factor of the requested section, and c a complex learned correction initialized at zero, so training starts exactly at the physics reconstruction. The condition-dependent part of c is bounded by a tanh at ± 0.5, so that the reconstruction stays within a factor 1 ± 0.5 of the physics gain by construction: the anchor is what the model is for, and an unbounded correction would be free to discard it. The bound was not swept, and the unbounded variant is not reported.

Hybrid-TCN. The same bound in the time domain: $M = M_{\mathrm{phys}} + 0.5\,\sigma(M_{\mathrm{phys}})\,\tanh(\mathrm{TCN}(\mathrm{inputs}))$, so the correction never exceeds half the amplitude of the physics reconstruction.

Prob-TCN. The TCN with a heteroscedastic head (Nix & Weigend, 1994) that predicts mean and log-variance per sample, trained with the Gaussian negative log-likelihood; the predicted variance reports how much of the target the model believes it can explain.

G.2 Training cost

One run, training and scoring together, takes 7 to 9 minutes on one RTX 4090 for the small models, 10 to 60 for the pretrained encoders and 142 for Mamba; the median adaptation takes 5. Times are wall clock over the 66 within-tower runs of seed 0, one per GPU (Appendix A).

G.3 Metric definitions

To evaluate reconstructions against the simulated moments, we compare the fatigue damage $\hat{D}_i$ of each reconstruction with the true damage D_i of the same simulation, over the N simulations of the test split. The damage function and the metrics are defined as follows.

Why damage and not the waveform? In the lower panel of Figure 9 the two series overlay to the eye and would score well under any waveform error, yet the damage ratio is 4.2: a 30% overshoot of the first tower mode is raised to the power of three to five by the S–N curve.

Fatigue damage. The true and the reconstructed moment $M(t)$ over the scored window (400 to 1,000 s, 6,000 samples at 10 Hz) are low-passed at 3 Hz by zeroing the rFFT bins above the cutoff, and rainflow-counted (Matsuishi & Endo, 1968; Downing & Socie, 1982) into cycles with moment ranges ΔM_k and counts n_k (half cycles count 0.5). At the scored section, with mean radius r and wall thickness t, the stress range is

$$\Delta\sigma_k = \frac{\Delta M_k}{W}\left(\frac{\max(t, t_{\mathrm{ref}})}{t_{\mathrm{ref}}}\right)^{0.2}, \qquad W = \frac{\pi}{4}\,\frac{r^4 - (r-t)^4}{r},$$

with $t_{\mathrm{ref}} = 25\,\mathrm{mm}$ (DNV, 2024); r and t are the mean outer radius and the wall thickness of the FLOATBench section (of 30) nearest to the gauge, for the true and the reconstructed moment alike and in every table and figure. At the top of the redesigns this radius is 2.5 to 4% larger than the radius at the gauge height, which lowers the absolute damage there by 18 to 26%; it scales the true and the reconstructed damage alike, so the TCN top R^2 changes by at most 0.005; the cycles to failure follow the bilinear S–N curve (detail category E in air)

$$\log_{10} N_k = \begin{cases} 12.010 - 3\,\log_{10}\Delta\sigma_k & N_k \le 10^7, \\ 15.350 - 5\,\log_{10}\Delta\sigma_k & N_k > 10^7, \end{cases}$$

and the damage is Miner's sum $D = \sum_k n_k / N_k$. This is the damage model of FLOAT (Ribeiro et al., 2026), which produced the FLOATBench labels (Anonymous, 2026), and it scores every model and the physics baseline. The 3 Hz low-pass, added here and above the highest rotor harmonic, keeps the high-frequency noise of a sampled reconstruction out of the cycle count. The

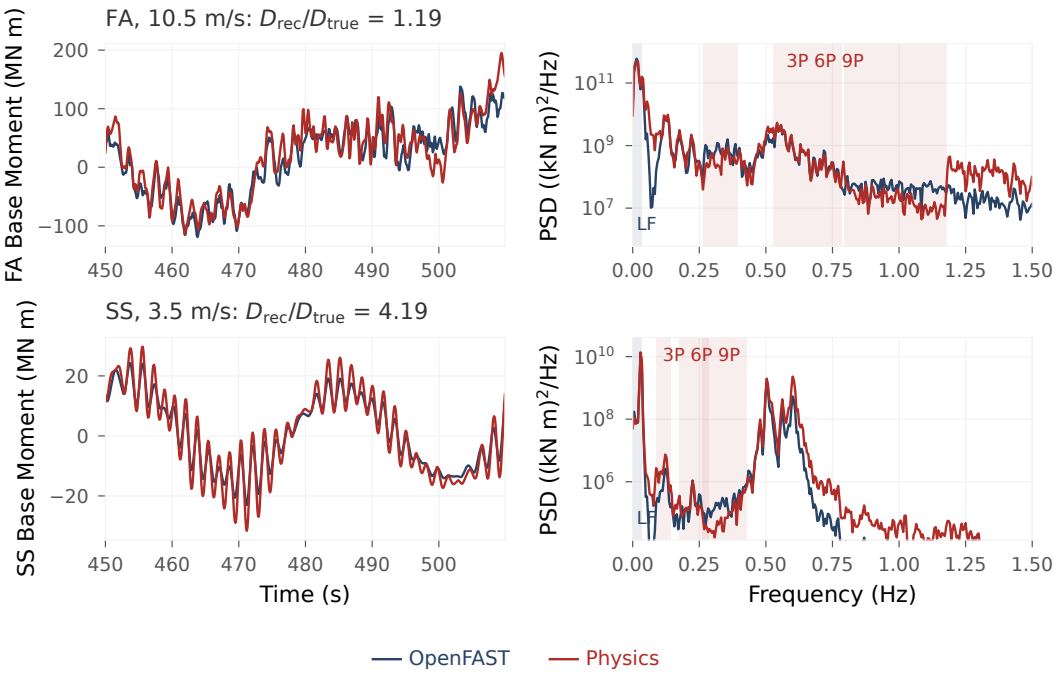


Figure 9: **Damage is not waveform error.** Simulated (blue) and physics-reconstructed (red) base moment on OPT2: 60 s of the scored window (left) and Welch spectra (right); shaded: quasi-static band (blue) and rotor-harmonic windows (red) of the physics gain. Top: fore-aft at 10.5 m/s, damage ratio 1.19. Bottom: side-side at 3.5 m/s, damage ratio 4.2; the series overlay but the reconstruction overshoots the first tower mode near 0.55 Hz.

scored window, 400 to 1,000 s at 10 Hz, holds 6,001 samples; the learned models use the first 6,000 (400.0 to 999.9 s), an even length for the spectral models, and the physics baseline uses all 6,001. The true damage of the two differs by a median of 0.02%.

R^2 of $\log_{10}$ damage. The proportion of the variance of the log damage explained by the reconstruction,

$$R^2 = 1 - \frac{\sum_i (\log_{10} D_i - \log_{10} \hat{D}_i)^2}{\sum_i (\log_{10} D_i - \overline{\log_{10} D})^2}.$$

Damage spans six orders of magnitude, so the logarithm puts every operating point on the same footing. Sorts the leaderboard; it scores the damage level and is read with ρ_{wc} and b_{wc}, which score the realization.

Median damage ratio. The median of the ratio of reconstructed to true damage,

$$\mathrm{median}_i \, (\hat{D}_i / D_i).$$

An unbiased model sits at 1; the ratio exposes a systematic over- or underestimate that R^2 does not see.

Mean relative error (MRE). The point-wise mean of relative deviations,

$$\mathrm{MRE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\hat{D}_i - D_i|}{D_i}.$$

Each simulation contributes equally and no variance enters the denominator, so the metric stays comparable between regime cells of different width. Used as the per-regime metric.

Fraction within a factor of two. The share of simulations whose reconstructed damage is within a factor of two of the true one,

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{0.5 \le \hat{D}_i/D_i \le 2\}.$$

The tolerance used for fatigue in practice; it separates models whose R^2 look alike.

Within-condition correlation. The Pearson correlation of the deviations of true and reconstructed log damage from their means over the six seeds of each operating point,

$$\rho_{\mathrm{wc}} = \mathrm{corr}(\delta, \hat{\delta}), \qquad \delta_i = \log_{10} D_i - \overline{\log_{10} D}_{c_i}, \quad \hat{\delta}_i = \log_{10} \hat{D}_i - \overline{\log_{10} \hat{D}}_{c_i},$$

with c_i the operating point of simulation i. A predictor that maps the operating point to a damage, such as the FLOATBench surrogates, has $\hat{\delta}_i = 0$ everywhere, for which the Pearson correlation is undefined (zero denominator); we set $\rho_{\mathrm{wc}} = 0$ by convention whenever the predicted deviations have zero variance, so such a predictor scores zero (no benchmark entry reaches this case); a positive value shows that the model reads the realization in the signal. The correlation is scale-invariant, so it is read with the median ratio and the fraction within a factor of two, which score the level.

These metrics together capture explained variance, bias, relative error, engineering tolerance and within-condition tracking. Higher is better for R^2, the fraction within a factor of two and ρ_{wc}, lower for MRE, and 1 is best for the median ratio.

G.4 Why a cluster bootstrap? Confidence intervals over operating points

Every interval is a 95% percentile bootstrap (Efron, 1979) that resamples operating conditions, not simulations. The test split of a tower holds K operating points (wind speed, H_s, T_p), each with six turbulence seeds; the seeds of one point share its loads and are not independent draws, so resampling simulations one by one understates the uncertainty. The point is therefore the sampling unit.

Method. Let $\mathcal{S}_k$ be the set of simulations of operating point k, $k = 1, \dots, K$, and $\theta = \mathcal{M}(\{(D_i, \hat{D}_i)\})$ one of the metrics of Appendix G.3. Replicate b draws K points with replacement, $k_1^{(b)}, \dots, k_K^{(b)} \sim \mathrm{Uniform}\{1, \dots, K\}$, keeps every simulation of each drawn point,

$$I^{(b)} = \bigcup_{j=1}^{K} \mathcal{S}_{k_j^{(b)}} \quad \text{(as a multiset)}, \qquad \theta^{(b)} = \mathcal{M}\big(\{(D_i, \hat{D}_i) : i \in I^{(b)}\}\big),$$

and the interval is $\mathrm{CI}_{95\%} = [\theta^*_{0.025}, \theta^*_{0.975}]$, the percentiles of $\{\theta^{(b)}\}_{b=1}^{B}$. The point estimate reported next to the interval is the metric on the full set, not the bootstrap mean.

Implementation. $B = 1{,}000$ replicates, generator seed 0, the three metrics computed in one pass (Algorithm 1). Per regime cell the same procedure runs on the points of the cell, so a cell with few points has a wide interval by construction. Simulations where either damage is not positive are dropped before the metric, and a replicate whose metric is undefined (a cell with a single point has no variance) is ignored in the percentiles. The intervals of the seed repeats in Appendix H are a different quantity: the range over training seeds at a fixed test set.

Paired comparison. Two models scored on the same split are compared on the same resampled points: replicate b draws the K points once, and records $\Delta^{(b)} = \theta_A^{(b)} - \theta_B^{(b)}$ on the simulations that both models score. The difference is significant when the 95% percentile interval of $\{\Delta^{(b)}\}$ excludes zero; the paper requires this on each of the three towers before calling two models different. Pairing removes the variation the two models share, the choice of test points, so the paired interval is narrower than the two marginal intervals suggest and the disjoint-interval rule is not used. Appendix H.5 lists the paired differences against the TCN.

Algorithm 1 CLUSTERBOOTSTRAP: percentile confidence intervals resampling operating points.

Require: True and reconstructed damage $\{(D_i, \hat{D}_i)\}_{i=1}^{N}$, operating point c_i of each simulation, metrics $\mathcal{M} = \{f_m\}_{m=1}^{3}$, replicates B (default 1,000), RNG seed s (default 0)
Ensure: For each metric m: $[\theta^*_{m,0.025}, \theta^*_{m,0.975}]$
Group simulations by point: $\mathcal{S}_k \leftarrow \{i : c_i = k\}$, $k = 1, \ldots, K$; seed RNG with s
for $b = 1, \ldots, B$ **do**
  Draw K points with replacement: $k_1, \ldots, k_K \sim \mathrm{Uniform}\{1, \ldots, K\}$
  $I^{(b)} \leftarrow \mathcal{S}_{k_1} \uplus \cdots \uplus \mathcal{S}_{k_K}$ ▷ every seed of a drawn point is kept
  for $m = 1, \ldots, 3$ **do**
    $\Theta_m[b] \leftarrow f_m\big(\{(D_i, \hat{D}_i) : i \in I^{(b)}\}\big)$
  end for
end for
for $m = 1, \ldots, 3$ **do**
  $\theta^*_{m,0.025}, \theta^*_{m,0.975} \leftarrow \mathrm{percentile}(\Theta_m;\, 2.5,\, 97.5)$ ▷ undefined replicates ignored
end for
return $\{[\theta^*_{m,0.025},\, \theta^*_{m,0.975}]\}_{m=1}^{3}$

Interpretation. A paired interval that excludes zero means the ranking survives a change of test points. The intervals do not cover training-seed variation, which the seed repeats report separately (Appendix H.2); a difference that is significant under the paired bootstrap but smaller than the spread over training seeds is a property of one training run, not of the architecture. Tables that average the three towers give the mean per-tower half-width as $\pm$, not an interval on the mean.

G.5 MODEL SELECTION ON A HELD-OUT VALIDATION SPLIT

Table 10: **The top three by criterion.** R^2 of $\log_{10}$ damage, mean over the three towers. Validation rows: models trained on 80% of the training operating points, scored on the held-out 20%, three seeds (Appendix G.5); their union (TCN, Prob-TCN, FITS, PatchTST, Mamba) adds Prob-TCN and Mamba to the eight family representatives. Test rows: the twenty learned models, seed 0.

Criterion	First	Second	Third
Validation split, base	Mamba (0.994)	PatchTST (0.994)	TCN (0.990)
Validation split, eleven heights	TCN (0.950)	Prob-TCN (0.929)	FITS (0.906)
Validation split, top	Prob-TCN (0.874)	FITS (0.781)	TCN (0.655)
Test, all regimes, eleven heights	TCN (0.953)	Prob-TCN (0.947)	DLinear (0.900)
Test, wind-EX/wave-EX cell, eleven heights	Spectral gain (0.973)	S4D (0.971)	TCN (0.971)
Test, all regimes, base	Mamba (0.997)	PatchTST (0.995)	TCN (0.988)
Test, all regimes, top	Prob-TCN (0.861)	TCN (0.725)	FITS (0.721)
Test, EX/EX cell, top	DLinear (0.912)	Spectral gain (0.860)	S4D (0.860)

The models that receive the extra experiments are selected without the test split. The validation split holds 58 of the 288 training operating points (20%, drawn uniformly with a fixed seed), all six turbulence seeds of a drawn point on the validation side as in the test partition, the same identifiers on the three towers: 1,380 training and 348 validation simulations. Every model is retrained on the 1,380 with the recipe of Section 4.4 and three training seeds on each tower (180 runs), the hybrid models with the physics recalibrated on the same 1,380, and scored on the 348 by three quantities, each averaged over the three towers and the three seeds: S_{base}, the R^2 of $\log_{10}$ damage at the base; S_{global}, its mean over the eleven heights; S_{top}, its value at the top. The family representative is the model with the highest S_{global} in its family (Table 11); the union of the top three under each score (the three validation rows of Table 10) is the set the main text analyzes. Both sets are fixed by the validation scores alone, and the leaderboard covers all twenty models either way, so the selection decides which models are probed further and not which are compared; the final models are those of the matrix, trained on the full training split.

The validation selection agrees with the test-set ranking on six of the eight families and changes two: FITS replaces the spectral gain (0.906 against 0.890, a margin within the seed range but held on six of the nine tower-seed cells) and Chronos replaces TimesFM (0.812 against 0.647, on every tower and seed; TimesFM ranges 0.38 to 0.80 over seeds). Table 12 gives the winners per tower and per seed: seven families have the same winner on every seed and five on every tower; on REF the

iTransformer and Hybrid-TCN win their families, and the iTransformer wins one seed in three, the model that diverges under the common recipe. The union is stable in the same sense: Mamba and PatchTST enter the base top three in eight of the nine cells, FITS the top top three in nine of nine.

Two readings of Table 11 matter for the rest of the paper. The base-to-top inversion is already visible on the validation split: Mamba and PatchTST, first and second at the base, are sixteenth and seventeenth at the top (Spearman correlation between S_{base} and S_{top} over the twenty models -0.07). And the validation split, drawn from inside the training envelope, cannot see the extrapolation regimes: FITS is third on validation and nineteenth on the wind-EX/wave-EX cell of the test split, which is the case for the regime-aware partition.

Table 11: **Validation selection.** R^2 of $\log_{10}$ damage on the validation split, mean over eleven heights, three towers and three seeds ($\pm$ half the seed range); * marks the family representative, the model with the highest score in its family, which with the union of the top three gives the ten of Section 4. No test number enters this table.

Model	Family	Validation R^2	Rank	Rank in family
TCN*	convolutional	0.950±0.007	1	1
Prob-TCN	convolutional	0.929±0.018	2	2
FITS*	spectral	0.906±0.010	3	1
Spectral gain	spectral	0.890±0.008	4	2
S4D*	state-space	0.864±0.025	5	1
DLinear	convolutional	0.862±0.050	6	3
LSTM*	recurrent	0.855±0.031	7	1
Chronos*	pretrained	0.812±0.017	8	1
U-Net	convolutional	0.809±0.063	9	4
Mamba	state-space	0.786±0.035	10	2
MOMENT	pretrained	0.700±0.088	11	2
PatchTST*	attention	0.681±0.074	12	1
TimesFM-ft	pretrained	0.664±0.113	13	3
TimesFM	pretrained	0.647±0.211	14	4
FNO	spectral	0.638±0.027	15	3
Hybrid*	hybrid	0.573±0.029	16	1
Hybrid-TCN	hybrid	0.383±0.004	17	2
TimesNet*	period-folding	0.368±0.163	18	1
MOMENT-ft	pretrained	0.292±0.118	19	5
iTransformer	attention	0.090±0.686	20	2

Table 12: **Stability of the family winner** on validation, per tower (mean over seeds) and per seed (mean over towers); the margin to the second of the family in parentheses.

	per tower (3 seeds)			per seed (3 towers)		
Family	REF	OPT1	OPT2	seed 0	seed 1	seed 2
convolutional	TCN (+0.01)	TCN (+0.01)	TCN (+0.04)	TCN (+0.01)	TCN (+0.01)	TCN (+0.04)
recurrent	LSTM	LSTM	LSTM	LSTM	LSTM	LSTM
attention	iTransformer (+0.05)	PatchTST (+1.53)	PatchTST (+0.29)	PatchTST (+0.74)	PatchTST (+1.24)	iTransformer (+0.20)
state-space	S4D (+0.04)	S4D (+0.06)	S4D (+0.14)	S4D (+0.10)	S4D (+0.08)	S4D (+0.06)
spectral	FITS (+0.04)	FITS (+0.02)	Spectral gain (+0.01)	FITS (+0.01)	FITS (+0.02)	FITS (+0.02)
period-folding	TimesNet	TimesNet	TimesNet	TimesNet	TimesNet	TimesNet
pretrained	Chronos (+0.07)	Chronos (+0.12)	Chronos (+0.10)	Chronos (+0.06)	Chronos (+0.01)	Chronos (+0.09)
hybrid	Hybrid-TCN (+0.09)	Hybrid (+0.30)	Hybrid (+0.36)	Hybrid (+0.17)	Hybrid (+0.23)	Hybrid (+0.18)

H LEADERBOARDS (WITHIN TOWER, CROSS TOWER, SENSOR SETS)

The tables of this appendix are the full results behind Section 5, one per tower, pair or input set; the sections that follow answer one question each. Section 5 reports the median of three training seeds where three exist and seed 0 otherwise; the tables here report seed 0, the run every model has, unless the caption says otherwise. All numbers are at the declared 50 epochs unless epochs are stated.

H.1 WITHIN TOWER, PER TOWER

Tables 13 to 15 give the R^2 of $\log_{10}$ damage of every model at the eleven heights of each tower, seed 0; values below -1 are shown as < -1.

Table 13: **Within tower on REF.** R^2 of $\log_{10}$ damage by height z/H, seed 0.

Model / z/H	0.00	0.08	0.18	0.28	0.38	0.48	0.58	0.68	0.78	0.88	1.00
Physics	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.94	0.93	0.93
Floor	0.80	0.80	0.81	0.81	0.82	0.81	0.81	0.79	0.77	0.72	0.75
TCN	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97	0.83
U-Net	0.99	0.99	0.99	0.99	0.98	0.98	0.99	0.98	0.96	0.86	0.13
DLinear	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.97	0.95	0.93
LSTM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.92	0.33
PatchTST	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.96	0.64	-0.94
iTransformer	0.21	0.29	0.26	0.21	0.17	0.13	0.10	0.06	0.01	-0.06	0.02
S4D	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.96	0.85	0.32
Mamba	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.98	0.82	-0.43
Spectral gain	0.99	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.98	0.88	0.38
FITS	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86	0.86	0.84
FNO	0.89	0.88	0.86	0.84	0.81	0.76	0.68	0.54	0.32	-0.02	0.05
TimesNet	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.96	0.93	0.34
Chronos	0.94	0.94	0.94	0.95	0.95	0.95	0.94	0.94	0.93	0.90	0.86
MOMENT	0.90	0.91	0.91	0.92	0.92	0.92	0.92	0.91	0.89	0.86	0.85
MOMENT (ft)	0.90	0.90	0.91	0.91	0.92	0.92	0.91	0.90	0.89	0.86	0.81
TimesFM	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.89	0.88	0.86	0.80
TimesFM (ft)	0.86	0.86	0.87	0.87	0.88	0.88	0.88	0.87	0.86	0.84	0.79
Hybrid	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.98	0.96	0.79	-0.60
Hybrid-TCN	0.93	0.93	0.93	0.93	0.93	0.94	0.94	0.94	0.95	0.95	0.73
Prob-TCN	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.94	0.86

Table 14: **Within tower on OPT1.** R^2 of $\log_{10}$ damage by height z/H, seed 0.

Model / z/H	0.00	0.08	0.18	0.28	0.38	0.48	0.58	0.68	0.78	0.88	1.00
Physics	0.94	0.95	0.95	0.96	0.96	0.97	0.96	0.93	0.82	0.43	< -1
Floor	0.69	0.70	0.72	0.73	0.75	0.77	0.79	0.80	0.81	0.76	0.70
TCN	0.98	0.98	0.98	0.98	0.98	0.97	0.96	0.95	0.91	0.84	0.64
U-Net	0.97	0.97	0.97	0.96	0.95	0.93	0.89	0.81	0.62	0.21	-0.43
DLinear	0.90	0.91	0.92	0.93	0.93	0.94	0.93	0.91	0.86	0.74	0.58
LSTM	0.96	0.96	0.96	0.97	0.97	0.96	0.95	0.93	0.88	0.77	0.38
PatchTST	0.99	0.99	0.99	0.99	0.99	0.99	0.96	0.89	0.59	-0.71	< -1
iTransformer	-0.24	-0.41	-0.22	-0.16	-0.09	0.00	0.13	-0.05	-0.24	-0.43	-0.34
S4D	0.97	0.97	0.98	0.98	0.98	0.97	0.97	0.94	0.89	0.72	0.40
Mamba	1.00	1.00	1.00	1.00	0.99	0.99	0.98	0.95	0.86	0.53	-0.27
Spectral gain	0.96	0.96	0.95	0.95	0.94	0.94	0.92	0.91	0.85	0.68	0.48
FITS	0.70	0.71	0.71	0.71	0.70	0.71	0.73	0.75	0.78	0.76	0.66
FNO	0.95	0.94	0.93	0.91	0.89	0.85	0.78	0.65	0.41	0.01	-0.24
TimesNet	0.84	0.83	0.83	0.82	0.80	0.78	0.76	0.73	0.64	0.27	< -1
Chronos	0.68	0.68	0.69	0.70	0.71	0.71	0.69	0.64	0.54	0.38	0.24
MOMENT	0.70	0.71	0.72	0.73	0.74	0.74	0.72	0.67	0.58	0.42	0.27
MOMENT (ft)	0.41	0.42	0.44	0.46	0.47	0.48	0.45	0.40	0.30	0.13	0.03
TimesFM	0.83	0.83	0.83	0.84	0.84	0.84	0.82	0.78	0.70	0.55	0.38
TimesFM (ft)	0.66	0.66	0.67	0.69	0.70	0.70	0.68	0.64	0.55	0.40	0.26
Hybrid	0.95	0.96	0.96	0.96	0.96	0.95	0.92	0.87	0.71	0.17	< -1
Hybrid-TCN	0.77	0.77	0.76	0.74	0.71	0.66	0.57	0.44	0.21	-0.33	< -1
Prob-TCN	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.96	0.96	0.87

Table 15: **Within tower on OPT2.** R^2 of $\log_{10}$ damage by height z/H, seed 0.

Model / z/H	0.00	0.08	0.18	0.28	0.38	0.48	0.58	0.68	0.78	0.88	1.00
Physics	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.94	0.83	0.44	< -1
Floor	0.73	0.74	0.76	0.78	0.79	0.81	0.83	0.84	0.83	0.75	0.65
TCN	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.93	0.71
U-Net	0.96	0.96	0.95	0.94	0.94	0.93	0.92	0.90	0.87	0.78	0.54
DLinear	0.90	0.91	0.92	0.93	0.94	0.95	0.94	0.92	0.86	0.72	0.54
LSTM	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.94	0.45	0.55
PatchTST	0.99	0.99	0.99	0.99	0.99	0.98	0.95	0.89	0.69	-0.03	< -1
iTransformer	< -1	< -1	< -1	< -1	< -1	< -1	< -1	< -1	< -1	< -1	< -1
S4D	0.98	0.98	0.98	0.98	0.98	0.98	0.97	0.95	0.90	0.74	0.44
Mamba	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.97	0.91	0.64	-0.19
Spectral gain	0.96	0.96	0.96	0.96	0.95	0.95	0.94	0.93	0.90	0.77	0.57
FITS	0.70	0.71	0.70	0.69	0.68	0.69	0.71	0.74	0.78	0.77	0.66
FNO	0.95	0.94	0.93	0.92	0.90	0.86	0.78	0.65	0.41	0.03	-0.12
TimesNet	0.77	0.77	0.76	0.74	0.73	0.71	0.65	0.56	0.41	-0.13	< -1
Chronos	0.72	0.72	0.74	0.76	0.78	0.78	0.76	0.71	0.61	0.42	0.26
MOMENT	0.75	0.76	0.77	0.79	0.80	0.80	0.78	0.73	0.63	0.45	0.28
MOMENT (ft)	0.28	0.30	0.33	0.36	0.39	0.39	0.36	0.30	0.18	-0.02	-0.12
TimesFM	0.84	0.84	0.85	0.86	0.87	0.87	0.85	0.82	0.73	0.56	0.38
TimesFM (ft)	0.68	0.69	0.71	0.73	0.74	0.75	0.73	0.68	0.59	0.42	0.26
Hybrid	0.95	0.96	0.96	0.96	0.96	0.95	0.94	0.91	0.79	0.32	< -1
Hybrid-TCN	0.75	0.75	0.75	0.74	0.72	0.68	0.60	0.47	0.25	-0.29	< -1
Prob-TCN	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.94	0.85

H.2 TRAINING-SEED SPREAD

Table 16 repeats thirteen models, the ten selected plus DLinear, the spectral gain and TimesFM, with three training seeds (0, 1, 2) on every tower and reports the mean and half the range over the seeds at three heights. The bootstrap intervals of the main tables do not cover this spread; a difference between two models is a finding only when it exceeds both. At the top most of this spread is under-training: the per-tower seed range of the TCN at the top falls from 0.06–0.15 at 50 epochs to 0.004–0.03 at 300 (Appendix J.1).

Table 16: **Training-seed spread.** R^2 of $\log_{10}$ damage within tower, mean $\pm$ half range over three seeds, at the base, z/H 0.78 and the top of each tower.

	REF			OPT1			OPT2		
Model	base	z/H 0.78	top	base	z/H 0.78	top	base	z/H 0.78	top
TCN	0.997±0.001	0.990±0.003	0.789±0.031	0.988±0.007	0.953±0.032	0.732±0.076	0.990±0.003	0.978±0.008	0.770±0.077
LSTM	0.989±0.003	0.923±0.065	0.369±0.104	0.981±0.017	0.877±0.042	0.352±0.208	0.989±0.008	0.928±0.014	0.587±0.101
PatchTST	0.999±0.000	0.946±0.017	−0.956±0.033	0.994±0.002	0.185±0.631	−1.457±0.245	0.987±0.005	0.761±0.058	−1.075±0.419
S4D	0.995±0.003	0.959±0.001	0.360±0.112	0.978±0.004	0.886±0.013	0.450±0.046	0.975±0.002	0.901±0.003	0.519±0.073
Mamba	0.997±0.001	0.976±0.010	−0.385±0.039	0.982±0.019	0.820±0.115	−0.254±0.163	0.995±0.003	0.907±0.016	−0.168±0.191
Spectral gain	0.996±0.002	0.989±0.003	0.385±0.070	0.965±0.018	0.871±0.014	0.461±0.041	0.965±0.018	0.912±0.011	0.564±0.036
DLinear	0.933±0.037	0.935±0.035	0.884±0.046	0.878±0.024	0.830±0.036	0.544±0.045	0.885±0.024	0.834±0.030	0.520±0.040
FITS	0.851±0.018	0.848±0.011	0.829±0.012	0.721±0.019	0.791±0.013	0.661±0.002	0.713±0.008	0.785±0.007	0.657±0.008
TimesNet	0.891±0.115	0.818±0.204	−0.589±0.802	0.702±0.294	0.430±0.352	−2.753±1.124	0.816±0.066	0.350±0.313	−2.850±1.549
Chronos	0.933±0.012	0.922±0.012	0.854±0.011	0.773±0.099	0.642±0.100	0.330±0.087	0.799±0.094	0.689±0.095	0.337±0.092
TimesFM	0.884±0.073	0.883±0.072	0.808±0.089	0.629±0.209	0.510±0.205	0.218±0.176	0.679±0.170	0.581±0.161	0.240±0.144
Hybrid	0.982±0.002	0.959±0.008	−0.528±0.064	0.950±0.002	0.713±0.007	−3.439±0.087	0.950±0.003	0.790±0.005	−2.815±0.086
Prob-TCN	0.959±0.006	0.948±0.010	0.823±0.032	0.959±0.012	0.962±0.009	0.858±0.013	0.963±0.008	0.958±0.003	0.856±0.012

H.3 CROSS TOWER, PER PAIR

Table 17 gives every model zero-shot and after the protocol adaptation (ten target simulations, 100 steps, draw 0, seed 0) at the base on the six ordered pairs. The physics row is calibrated on the training split of the target, so it is a reference, not a zero-shot competitor.

Table 18 repeats the protocol adaptation (ten target simulations, 100 steps, draw 0) over the three training seeds for the ten selected models. The median seed range over models and pairs is 0.04 at the base and 0.34 at the top (TCN 0.03 and 0.45): the base holds across seeds, the top depends on the seed. At $N = 10$ adaptation lowers the Prob-TCN top below its zero-shot value on ten of the twelve redesign pair-seed cells, so its recovery of the top is not seed-robust at this budget; seed 0 is a pessimistic seed for the TCN top out of REF.

Table 17: **Cross tower, all models.** R^2 of $\log_{10}$ damage at the base, zero-shot / after ten target simulations, seed 0. †Physics: calibrated on the training split of the target tower, a within-tower reference, not a transfer result.

Model	REF→OPT1	REF→OPT2	OPT1→REF	OPT1→OPT2	OPT2→REF	OPT2→OPT1
Physics†	0.94	0.96	0.97	0.96	0.97	0.94
Floor	−0.79 / 0.01	−0.74 / 0.07	0.76 / 0.87	0.74 / 0.78	0.76 / 0.87	0.69 / 0.74
TCN	0.67 / 0.96	0.72 / 0.95	−0.34 / 0.99	0.99 / 0.98	−0.14 / 0.99	0.97 / 0.95
U-Net	0.36 / 0.97	0.43 / 0.92	0.89 / 0.96	0.98 / 0.95	−0.01 / 0.98	0.93 / 0.98
DLinear	−0.01 / 0.53	0.06 / 0.67	< −1 / 0.99	0.94 / 0.89	< −1 / 0.97	0.78 / 0.79
LSTM	0.47 / 0.90	0.54 / 0.88	0.78 / 0.98	0.98 / 0.94	0.68 / 0.99	0.98 / 0.97
PatchTST	0.54 / 0.93	0.63 / 0.88	0.86 / 0.97	1.00 / 0.96	0.90 / 1.00	0.98 / 0.87
iTransformer	−0.63 / < −1	−0.17 / −0.82	< −1 / −0.04	−0.16 / −0.41	< −1 / −0.69	< −1 / −0.42
S4D	0.71 / 0.96	0.76 / 0.96	0.69 / 0.97	0.99 / 0.97	0.79 / 0.99	0.96 / 0.98
Mamba	0.81 / 0.99	0.86 / 0.99	0.79 / 0.99	1.00 / 0.99	0.84 / 1.00	0.99 / 0.99
Spectral gain	0.54 / 0.97	0.60 / 0.97	0.85 / 0.86	0.98 / 0.98	0.69 / 0.86	0.93 / 0.97
FITS	0.72 / 0.72	0.71 / 0.71	0.02 / 0.06	0.62 / 0.65	0.15 / 0.17	0.74 / 0.76
FNO	0.69 / 0.94	0.71 / 0.97	0.98 / 0.99	0.96 / 0.93	0.97 / 0.99	0.94 / 0.91
TimesNet	0.63 / 0.34	0.69 / 0.54	0.81 / 0.93	0.82 / 0.71	0.91 / 0.96	0.73 / 0.87
Chronos	0.02 / 0.54	0.19 / 0.63	0.83 / 0.97	0.80 / 0.81	0.80 / 0.97	0.56 / 0.76
MOMENT	−0.28 / 0.51	−0.25 / 0.56	0.21 / 0.98	0.85 / 0.88	0.05 / 0.98	0.56 / 0.85
MOMENT (ft)	0.33 / 0.46	0.32 / 0.37	0.97 / 0.98	0.43 / 0.80	0.94 / 0.98	0.18 / 0.32
TimesFM	−0.63 / < −1	−0.56 / −0.82	0.23 / 0.82	0.92 / 0.12	0.25 / 0.78	0.61 / −0.58
TimesFM (ft)	< −1 / −0.97	< −1 / −1.00	0.78 / 0.74	0.60 / 0.16	0.74 / 0.61	0.26 / 0.12
Hybrid	0.27 / 0.82	0.35 / 0.84	0.03 / 0.79	0.98 / 0.96	−0.30 / 0.85	0.91 / 0.97
Hybrid-TCN	−0.05 / 0.76	0.03 / 0.75	0.75 / 0.01	0.82 / 0.75	0.86 / 0.01	0.70 / 0.76
Prob-TCN	0.78 / 0.96	0.83 / 0.87	−0.29 / 0.95	0.93 / 0.92	0.22 / 0.93	0.95 / 0.95

Table 18: **Cross tower over training seeds.** Log-damage R^2 at the base (upper block) and the top (lower block) after the protocol adaptation, median over three training seeds (range, max minus min, in parentheses). Values below -1 shown as <-1.

Model	REF→OPT1	REF→OPT2	OPT1→REF	OPT1→OPT2	OPT2→REF	OPT2→OPT1
Base						
TCN	0.96 (0.19)	0.95 (0.03)	0.99 (0.01)	0.98 (0.06)	0.99 (0.01)	0.97 (0.03)
Prob-TCN	0.94 (0.04)	0.92 (0.06)	0.94 (0.03)	0.92 (0.05)	0.91 (0.03)	0.95 (0.01)
S4D	0.96 (0.03)	0.96 (0.03)	0.99 (0.02)	0.97 (0.01)	0.99 (0.00)	0.98 (0.01)
LSTM	0.79 (0.15)	0.85 (0.08)	0.98 (0.03)	0.94 (0.07)	0.99 (0.01)	0.96 (0.11)
PatchTST	0.96 (0.06)	0.98 (0.11)	0.97 (0.07)	0.97 (0.03)	1.00 (0.08)	0.96 (0.10)
Mamba	0.97 (0.03)	0.98 (0.02)	0.99 (0.00)	0.97 (0.02)	1.00 (0.00)	0.99 (0.03)
TimesNet	0.71 (0.60)	0.69 (0.36)	0.96 (0.04)	0.71 (0.78)	0.96 (0.03)	0.86 (0.14)
Hybrid	0.82 (0.01)	0.84 (0.01)	0.80 (0.03)	0.96 (0.00)	0.86 (0.02)	0.97 (0.00)
Chronos	0.45 (1.26)	0.58 (1.20)	0.96 (0.17)	0.73 (0.83)	0.95 (0.19)	0.54 (0.87)
FITS	0.76 (0.05)	0.74 (0.05)	0.11 (0.05)	0.70 (0.08)	0.22 (0.05)	0.79 (0.05)
Top						
TCN	0.01 (0.56)	0.08 (0.53)	0.56 (0.30)	0.75 (0.38)	0.48 (0.60)	0.74 (0.36)
Prob-TCN	0.82 (0.26)	0.83 (0.06)	0.52 (0.22)	0.78 (0.20)	0.50 (0.29)	0.82 (0.07)
S4D	−0.24 (0.96)	−0.19 (0.78)	0.37 (0.32)	0.47 (0.28)	0.43 (0.17)	0.52 (0.19)
LSTM	−0.43 (1.00)	−0.25 (0.94)	−0.38 (0.98)	0.54 (0.20)	−0.10 (0.87)	0.62 (0.22)
PatchTST	< −1 (0.78)	< −1 (1.32)	< −1 (2.04)	−0.47 (0.89)	−0.69 (1.69)	−0.05 (1.09)
Mamba	−0.62 (0.24)	−0.54 (0.32)	−0.49 (1.26)	−0.35 (0.21)	−0.20 (0.65)	0.07 (1.04)
TimesNet	< −1 (3.31)	< −1 (1.83)	< −1 (2.02)	< −1 (1.01)	< −1 (0.85)	< −1 (3.05)
Hybrid	< −1 (0.31)	< −1 (0.32)	0.60 (0.11)	< −1 (0.09)	0.67 (0.11)	< −1 (0.18)
Chronos	0.02 (0.91)	0.02 (0.93)	0.92 (0.21)	0.27 (0.64)	0.92 (0.19)	0.28 (0.60)
FITS	0.68 (0.07)	0.65 (0.08)	0.36 (0.02)	0.69 (0.06)	0.38 (0.02)	0.70 (0.07)

H.4 Sensor sets

Table 19 scores the ten selected models under the five input sets of Section 4.2, mean over the three towers and three seeds.

H.5 Paired comparisons, intervals and exclusion audit

Table 20 gives the paired cluster bootstrap of Appendix G.4 against the TCN, seed 0. At the base Prob-TCN is 0.04 below the TCN and at the top 0.14 above it; PatchTST and Mamba are 0.01 above at the base and more than 1 below at the top; the two-axis TCN is 0.22 above at the top. Each of these excludes zero on all three towers. The bootstrap resamples test points, not training seeds; at the top the seed spread of Appendix H.2 is often larger than these intervals, so a difference at the

Table 19: **Sensor ablation.** R^2 of $\log_{10}$ damage at the base and at the top, mean over the three towers and the three training seeds, per input set; Section 5.4 quotes medians (Table 40), within 0.02 of these means. n/a: Chronos and the hybrid read the fore-aft acceleration alone by construction, so their task run is already an accelerometer-only run and the other input sets are not defined for them.

	base					top				
Model	task	accel.	SCADA	2-axis	+platform	task	accel.	SCADA	2-axis	+platform
TCN	0.99	0.96	0.88	0.99	0.99	0.76	0.70	0.28	0.94	0.94
Prob-TCN	0.96	0.97	0.94	0.97	0.98	0.85	0.83	0.86	0.92	0.90
LSTM	0.99	0.92	0.24	0.98	0.97	0.44	0.29	< -1	0.75	0.68
PatchTST	0.99	0.91	-0.31	0.99	0.98	< -1	< -1	< -1	-0.31	0.27
Mamba	0.99	0.86	0.40	1.00	0.99	-0.27	< -1	< -1	0.04	0.34
S4D	0.98	0.97	0.29	0.99	0.98	0.44	-0.09	< -1	0.59	0.52
FITS	0.76	0.97	0.83	0.58	0.30	0.72	0.78	0.61	0.71	0.68
TimesNet	0.80	0.85	< -1	0.89	-0.56	< -1	< -1	< -1	-0.11	-0.88
Chronos	0.83	n/a	n/a	n/a	n/a	0.51	n/a	n/a	n/a	n/a
Hybrid	0.96	n/a	n/a	n/a	n/a	< -1	n/a	n/a	n/a	n/a

top counts only when it exceeds both. Table 21 gives the intervals of the top three models within tower and per pair after adaptation. The exclusion audit (`exclusion_audit.csv`) covers the 16 models and variants of the two tables at the 11 heights on the three towers, 528 cells: no simulation is dropped, 4,740 of 4,740 in every cell.

Table 20: Paired cluster bootstrap of the R^2 difference model $-$ TCN over 1,000 resamples of the operating points, within tower, seed 0: mean difference over the three towers and, in parentheses, the number of towers on which the 95% interval excludes zero.

Model $-$ TCN	base	z/H 0.48	z/H 0.78	top
Physics	-0.03 (3/3)	-0.02 (3/3)	-0.10 (3/3)	< -1 (3/3)
Floor	-0.25 (3/3)	-0.19 (3/3)	-0.16 (3/3)	-0.02 (3/3)
Prob-TCN	-0.04 (3/3)	-0.03 (3/3)	0.00 (3/3)	+0.14 (3/3)
LSTM	-0.01 (2/3)	0.00 (2/3)	-0.03 (3/3)	-0.31 (3/3)
PatchTST	+0.01 (3/3)	0.00 (3/3)	-0.21 (3/3)	< -1 (3/3)
Mamba	+0.01 (3/3)	+0.01 (3/3)	-0.04 (3/3)	< -1 (3/3)
S4D	-0.01 (3/3)	-0.01 (3/3)	-0.04 (3/3)	-0.34 (3/3)
FITS	-0.23 (3/3)	-0.23 (3/3)	-0.15 (3/3)	0.00 (1/3)
TimesNet	-0.13 (3/3)	-0.17 (3/3)	-0.29 (3/3)	< -1 (3/3)
Chronos	-0.21 (3/3)	-0.18 (3/3)	-0.27 (3/3)	-0.27 (3/3)
Hybrid	-0.03 (3/3)	-0.03 (3/3)	-0.14 (3/3)	< -1 (3/3)
Prob-TCN, SCADA only	-0.07 (3/3)	-0.07 (3/3)	-0.03 (3/3)	+0.13 (3/3)
TCN, two axes	0.00 (3/3)	+0.01 (3/3)	+0.03 (3/3)	+0.22 (3/3)

Table 21: **Bootstrap intervals of the top three.** R^2 of $\log_{10}$ damage with its 95% percentile cluster-bootstrap interval $[\theta^*_{0.025}, \theta^*_{0.975}]$ ($B = 1{,}000$, resampling operating points), top three models per tower at the base and the top, and per pair after ten target simulations (draw 0); seed 0. Points: operating points in the test split. The intervals exclude the seed spread (Table 16).

Rank	Model	R^2 of $\log_{10}$ damage $[\theta^*_{0.025}, \theta^*_{0.975}]$	points
Within tower, base			
REF			
1	PatchTST	0.998 [0.998, 0.999]	790
2	Mamba	0.998 [0.998, 0.998]	790
3	TCN	0.996 [0.996, 0.996]	790
OPT1			
1	Mamba	0.995 [0.994, 0.996]	790
2	PatchTST	0.993 [0.992, 0.994]	790
3	TCN	0.979 [0.977, 0.981]	790
OPT2			
1	Mamba	0.997 [0.997, 0.997]	790
2	PatchTST	0.992 [0.992, 0.993]	790
3	LSTM	0.990 [0.990, 0.991]	790
Within tower, top			
REF			
1	Physics	0.929 [0.922, 0.935]	790
2	DLinear	0.925 [0.918, 0.932]	790
3	Chronos	0.861 [0.846, 0.875]	790
OPT1			
1	Prob-TCN	0.872 [0.866, 0.879]	790
2	Floor	0.698 [0.659, 0.731]	790
3	FITS	0.659 [0.640, 0.675]	790
OPT2			
1	Prob-TCN	0.852 [0.844, 0.859]	790
2	TCN	0.713 [0.692, 0.732]	790
3	FITS	0.660 [0.641, 0.677]	790
Cross tower after ten target simulations, base			
REF → OPT1			
1	Mamba	0.985 [0.983, 0.987]	790
2	U-Net	0.969 [0.965, 0.973]	790
3	Spectral gain	0.969 [0.964, 0.973]	790
REF → OPT2			
1	Mamba	0.990 [0.989, 0.992]	790
2	Spectral gain	0.968 [0.962, 0.972]	790
3	FNO	0.967 [0.961, 0.972]	790
OPT1 → REF			
1	TCN	0.993 [0.992, 0.993]	790
2	Mamba	0.991 [0.990, 0.991]	790
3	FNO	0.990 [0.989, 0.991]	790
OPT1 → OPT2			
1	Mamba	0.985 [0.983, 0.987]	790
2	TCN	0.983 [0.981, 0.985]	790
3	Spectral gain	0.981 [0.978, 0.984]	790
OPT2 → REF			
1	Mamba	0.997 [0.997, 0.997]	790
2	PatchTST	0.997 [0.996, 0.997]	790
3	TCN	0.993 [0.992, 0.994]	790
OPT2 → OPT1			
1	Mamba	0.993 [0.992, 0.994]	790
2	U-Net	0.977 [0.975, 0.979]	790
3	S4D	0.975 [0.973, 0.977]	790

I DOES THE RANKING SURVIVE ALONG THE TOWER? (WITHIN TOWER)

I.1 THE BASE SATURATES IN LOG-DAMAGE R^2; THE TOP SEPARATES THE MODELS

Base-to-top ranking under other metrics. Table 22 repeats the base-to-top rank correlation of the 22 entries (seed 0 where only one seed exists) for the fraction within a factor of two and the lifetime-damage error $|\log_{10}(\sum w\hat{D}/\sum wD)|$, with a cluster bootstrap over the 790 operating points ($B = 1{,}000$; operating-point sampling only, not seed spread). Under no metric does the base ranking predict the top: the correlation is negative on every tower for the fraction within a factor of two and at most 0.41 (OPT2) for the lifetime error. With a practical-equivalence margin of $\Delta R^2 = 0.01$, four to five models are indistinguishable at the base on every tower (TCN, PatchTST, Mamba and S4D or LSTM, plus the spectral gain on REF) and one at the top; the base-equivalent models rank from 2nd (TCN on OPT1/OPT2) to 22nd at the top. At the base the sequence models lie within a few

Table 22: **Base-to-top rank correlation under three metrics.** Spearman ρ between the base and the top ranking of the 22 entries, within tower, three-seed median, [95% cluster-bootstrap interval].

	log-damage R^2	within a factor of two	lifetime-damage error
REF	-0.37 [-0.39, -0.32]	-0.32 [-0.33, -0.28]	-0.02 [-0.03, 0.03]
OPT1	0.04 [-0.01, 0.07]	-0.48 [-0.52, -0.45]	0.13 [0.10, 0.16]
OPT2	0.22 [0.18, 0.31]	-0.34 [-0.39, -0.28]	0.41 [0.38, 0.44]

hundredths of each other in log-damage R^2, although paired differences exclude zero on all three towers for several of them (Table 20) and the fraction within a factor of two still separates them (Table 26); at the top they spread over three units of R^2 (Figure 10). Figure 11 draws every model from its base score to its top score; lines that cross are the rank inversion. In Figure 11, PatchTST

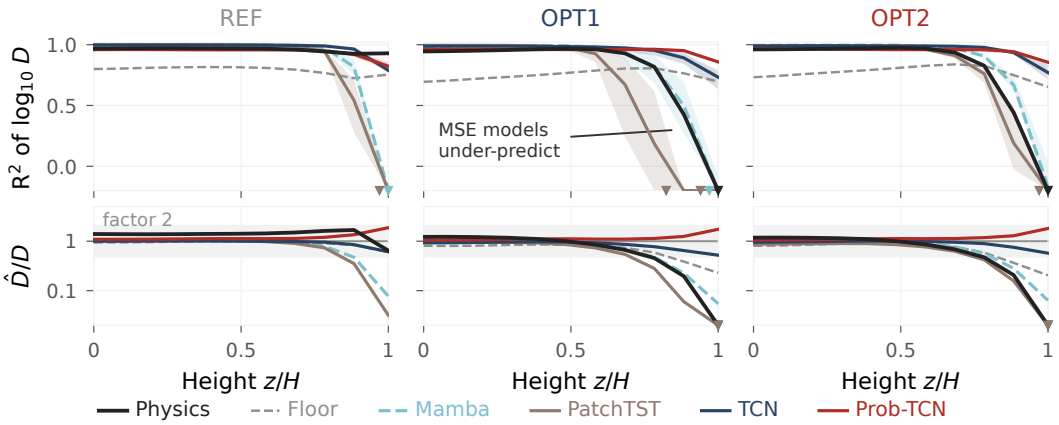


Figure 10: **Along the tower, within tower.** R^2 of $\log_{10}$ damage (top) and median damage ratio $\hat{D}/D$ (bottom, log scale, factor-two band shaded) against height, per tower, for the two best models at the base (Mamba, PatchTST), TCN, Prob-TCN, physics and floor; seed range as a band, triangles mark values below the axis. All models in Tables 13 to 15.

and Mamba, among the best at the base, end in the bottom third on every tower, while TCN stays second at the top on OPT1 and OPT2.

I.2 A GOOD TOP R^2 CAN HIDE THE WRONG REALIZATION

The within-condition correlation of Appendix G.3 scores the realization inside an operating point; Table 23 gives it next to the global R^2. At the base its order follows the leaderboard (Mamba 0.99, PatchTST 0.98, TCN 0.95, physics 0.93, floor 0.79). At the top, at 50 epochs, every single-axis model falls to 0.17–0.7, while Prob-TCN still reads 0.86 in global R^2; its SCADA-only variant keeps that score with a correlation of 0.51, so the sampled variance restores the level, not the realization. Only the two-axis TCN comes close to its base value at the top (0.89 against 0.97).

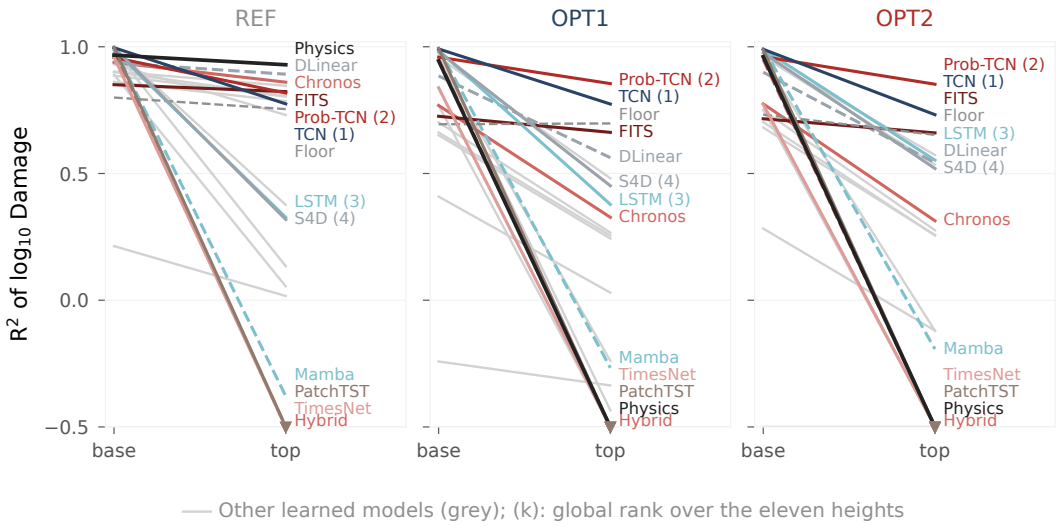


Figure 11: **From the base to the top.** R^2 of $\log_{10}$ damage at the base (left of each panel) and the top (right), one line per model, one panel per tower, three-seed median (seed 0 where only one seed exists). Colored and labeled: the eight family representatives, physics, floor, Prob-TCN, Mamba and DLinear, with the global rank over the eleven heights in parentheses for the top five; gray: other learned models; triangles: top values below -0.5. Crossing lines are the rank inversion.

Table 23: Global R^2 of $\log_{10}$ damage and within-condition correlation ρ_{wc} (Pearson correlation of predicted and true deviations from the six-seed mean of each operating point), within tower, seed 0, mean over the three towers. The correlation is reported without a bootstrap interval.

	R^2 of $\log_{10}$ damage				within-condition correlation			
Model	base	z/H 0.48	z/H 0.78	top	base	z/H 0.48	z/H 0.78	top
Physics	0.96	0.97	0.86	< -1	0.93	0.93	0.87	0.48
Floor	0.74	0.80	0.80	0.70	0.79	0.80	0.77	0.57
TCN	0.99	0.99	0.96	0.73	0.95	0.96	0.94	0.69
Prob-TCN	0.95	0.96	0.96	0.86	0.93	0.93	0.90	0.66
LSTM	0.98	0.98	0.93	0.42	0.96	0.96	0.94	0.62
PatchTST	0.99	0.99	0.75	< -1	0.98	0.97	0.89	0.48
Mamba	1.00	0.99	0.92	-0.30	0.99	0.97	0.92	0.51
S4D	0.98	0.98	0.92	0.39	0.95	0.94	0.90	0.51
FITS	0.76	0.75	0.81	0.72	0.71	0.71	0.68	0.38
TimesNet	0.86	0.82	0.67	< -1	0.57	0.53	0.44	0.17
Chronos	0.78	0.81	0.69	0.45	0.86	0.85	0.77	0.41
Hybrid	0.96	0.96	0.82	< -1	0.97	0.97	0.92	0.49
Prob-TCN, SCADA only	0.91	0.91	0.93	0.86	0.69	0.71	0.71	0.51
TCN, SCADA only	0.89	0.91	0.85	0.26	0.90	0.89	0.88	0.53
TCN, two axes	0.99	0.99	0.99	0.94	0.97	0.97	0.96	0.89
Prob-TCN, two axes	0.96	0.96	0.96	0.91	0.94	0.93	0.91	0.76

I.3 Over the DLC 1.2 design life the upper tower governs the fore-aft damage, and it is under-predicted

Table 24 weights the test split by the FLOATBench lifetime weights (Anonymous, 2026) (72% of the lifetime weight of each tower) and gives the ratio of reconstructed to true lifetime damage; Figure 12 draws the profiles. The true sums agree with the FLOATBench labels of the nearest section to within 7.1% at every gauge. The weights sum to every ten-minute period of 25 years but only DLC 1.2 is simulated, so these sums are a DLC-1.2-equivalent lifetime damage, not a full design-life fatigue assessment (no parked, idling, transient, fault or misaligned cases), and they use the fore-aft moment alone. On the redesigns the upper tower governs: on OPT1 the top's true lifetime damage (0.87) is 1.5 times that at the base (0.58), on OPT2 the two are about equal (0.63 against 0.59) and the largest of the five heights is z/H 0.88 (0.66), and on REF the base governs (24 against 2.3). At the base most of the ten selected models are within 20% on the redesigns, mostly on the unconservative side

(seed 0: TCN 0.84 and 0.92, PatchTST 0.91 and 0.98); FITS over-predicts by 33 to 40%, Chronos keeps less than half, and the physics is 10 to 11% conservative on the redesigns and 44% on REF. At the top (median of three seeds; the table gives seed 0) PatchTST and Mamba keep 1 to 8% of the lifetime damage, TCN 61 to 66%, and Prob-TCN over-predicts by 44 to 74%. On the redesigns the

Table 24: **Lifetime-weighted damage ratio.** Reconstructed over true lifetime damage at five heights, within tower, seed 0, test split weighted by the FLOATBench lifetime weights (72% of the lifetime weight per tower), for the physics, the floor and the ten selected models. Last row: true lifetime-weighted damage over the test split.

	REF					OPT1					OPT2				
z/H	0	0.48	0.78	0.88	top	0	0.48	0.78	0.88	top	0	0.48	0.78	0.88	top
Physics	1.44	1.45	1.69	1.83	0.68	1.11	0.96	0.47	0.20	< 0.01	1.10	0.98	0.49	0.21	< 0.01
Floor	0.97	1.01	0.95	0.91	0.70	0.80	0.82	0.60	0.41	0.21	0.77	0.87	0.60	0.38	0.18
TCN	0.95	0.98	0.96	0.91	0.72	0.84	0.85	0.64	0.51	0.43	0.92	0.99	0.88	0.75	0.59
Prob-TCN	1.08	1.12	1.18	1.30	1.49	1.01	1.06	0.99	1.01	1.31	1.01	1.05	1.02	1.10	1.60
FITS	1.03	1.07	1.02	0.98	0.79	1.33	1.46	1.11	0.76	0.40	1.40	1.57	1.13	0.75	0.39
LSTM	0.82	0.91	0.79	0.61	0.17	0.81	0.82	0.55	0.38	0.21	0.92	1.01	0.74	0.49	0.35
S4D	1.00	1.05	0.87	0.56	0.14	0.84	0.88	0.61	0.36	0.16	0.86	0.93	0.65	0.40	0.17
Chronos	0.66	0.68	0.64	0.60	0.44	0.44	0.47	0.35	0.24	0.12	0.48	0.53	0.37	0.24	0.11
TimesNet	0.87	0.93	0.86	0.75	0.19	1.00	0.84	0.49	0.21	0.01	0.72	0.71	0.40	0.18	0.01
Hybrid	0.87	0.87	0.73	0.40	0.04	0.79	0.76	0.41	0.17	< 0.01	0.80	0.78	0.47	0.19	< 0.01
PatchTST	0.97	1.03	0.81	0.46	0.04	0.91	0.91	0.33	0.07	0.02	0.98	0.95	0.46	0.17	0.02
Mamba	1.06	1.03	0.82	0.53	0.11	0.94	0.89	0.56	0.27	0.08	0.97	0.96	0.62	0.31	0.07
True damage	24	16	9.3	7.1	2.3	0.58	0.35	0.39	0.58	0.87	0.59	0.55	0.61	0.66	0.63

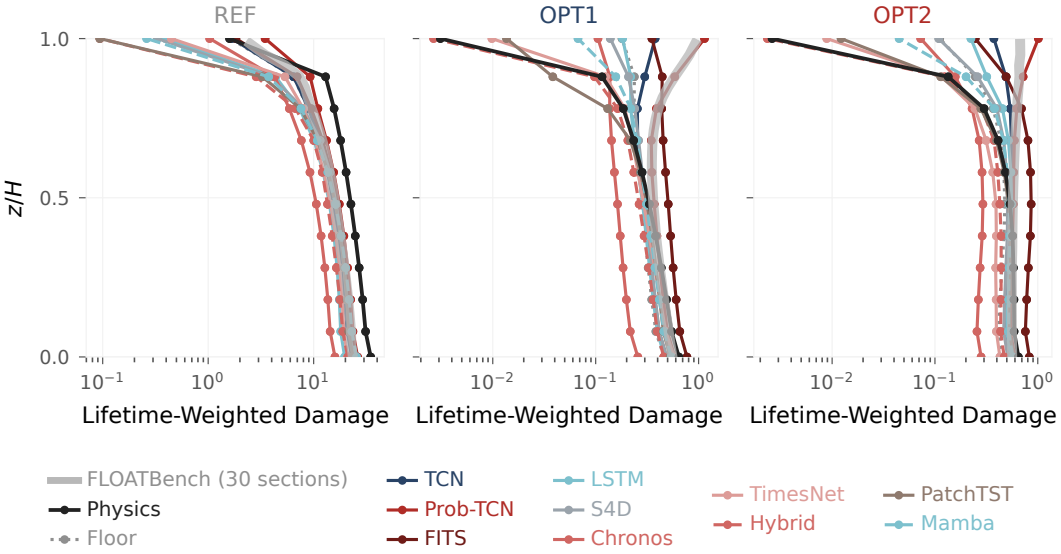


Figure 12: **Lifetime-weighted damage along the tower.** FLOATBench label profile at 30 sections (gray) and reconstructed lifetime-weighted damage at the eleven gauge heights (markers), per tower, log axis, test split, seed 0, for the ten selected models (the eight family representatives and the top-three union), the physics and the floor.

true lifetime damage is at least as large at the top as at the base (Figure 12); PatchTST and Mamba fall one to two decades short of it at the top, TCN less than a factor of 2.5, the physics baseline more than two decades on the redesigns, and Prob-TCN overshoots.

I.4 At the top, low wind is the hard condition

At the top the wind speed matters, not the regime label (Table 25). At 50 epochs TCN is within a factor of two on 0–1% of simulations at 4.5 to 9.5 m/s, in train, and on 92–99% above 13.5 m/s.

Training longer lifts the in-train low winds (TCN 40–51% at 300 epochs) far more than the extrapolated 3.5 m/s (17–33%). Figure 13 gives the median damage ratio per wind speed. In Figure 13,

Table 25: **Top of the tower by wind speed.** Percentage of test simulations within a factor of two of the true damage at the top, by wind band; 3.5 m/s is extrapolated, 10.5 to 12.5 m/s is around rated wind. Range over the three towers of the three-seed median (the floor has one run).

Model	Epochs	Wind speed [m/s]			
		3.5	4.5–9.5	10.5–12.5	13.5–24.5
TCN	50	6–12	0–1	35–41	92–99
TCN	300	17–33	40–51	64–74	90–98
Prob-TCN	50	0–0	2–8	36–49	93–96
Prob-TCN	300	1–15	54–71	96–99	100–100
S4D	50	22–39	2–4	0–0	0–0
S4D	300	27–32	2–3	0–2	0–4
Floor	50	25–47	6–51	12–73	1–85

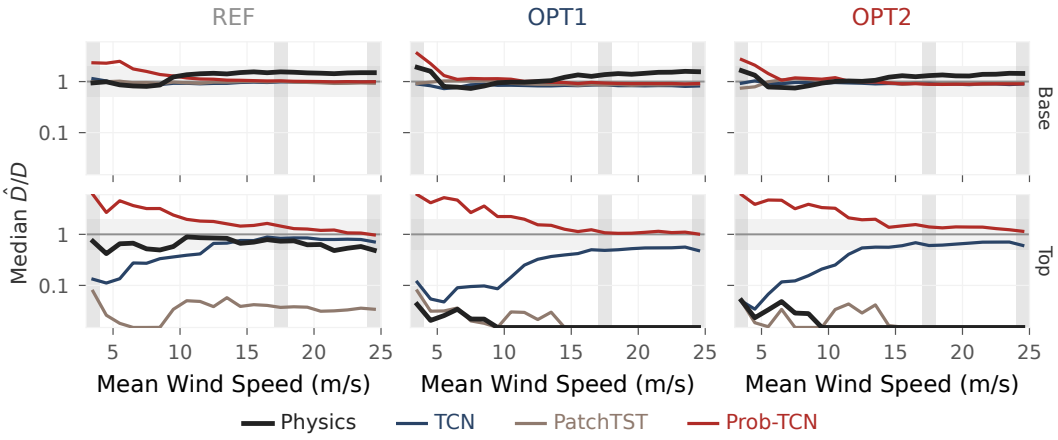


Figure 13: **Median damage ratio per mean wind speed**, within tower, seed 0, at the base (top row) and the top of the tower (bottom row); gray band: factor-two range; gray columns: the four held-out wind speeds.

at the base every model stays in the band at every wind except Prob-TCN, which over-predicts by 2 to 3.6 at 3.5 to 5.5 m/s. At the top the error depends on wind speed: TCN under-predicts by a factor of 3 to 30 below rated wind and recovers to 0.3 to 0.9 above it, Prob-TCN over-predicts by 2 to 9 below rated and approaches 1 at the highest winds, PatchTST stays below 0.1 at every wind, and the physics baseline is below 0.05 on the two redesigns (below 0.02 from 10.5 m/s up; the extrapolated height factor) while it stays between 0.4 and 0.9 on REF.

I.5 Waves matter far less than wind and height

Holding the other axis fixed (27 contrasts per model and height, cluster-bootstrap intervals), the wind effect on the fraction within a factor of two is 3 to 20 times the wave effect at the base for every model and at the top for the floor, the physics and Prob-TCN; for TCN at the top the two are of the same size (mean 6 against 9 points) and almost no interval excludes zero, so there wind speed, not the regime label, sets the error. At fixed wind the wave contrasts span -16 to $+20$ points (pooled over wind: -10 to $+9$). At the base the wave axis stays flat and the wind axis sets the error (Table 26): the physics is within a factor of two on 96 to 98% of simulations in every wave group and on 84% on extrapolated wind. Figure 14 shows the nine cells per tower at the top and at the base for the physics and the three models that beat the floor at the top: the wave blocks repeat at every height, and the top is where every model drops. At the top Figure 15 shows the same: across a row the wave regime changes neither the series nor the ratios; down a column the wind does. In Figure 15, reading across a row, the wave regime changes neither the series nor the ratios; reading down a column, the wind does: at 5.5 m/s the TCN misses the 3P oscillation that the two-axis TCN

Table 26: **Within tower by regime cell, base.** Fraction of test simulations within a factor of two (%), pooled over the wave groups for the wind axis and over the wind groups for the wave axis, and in the cell that extrapolates on both (254 simulations, 62 operating points). Mean over the three towers, seed 0.

	wind			waves			
Model	IT	IP	EX	IT	IP	EX	EX_EX
Physics	99	100	84	96	97	98	90
Floor	69	91	52	68	75	65	47
TCN	99	100	95	99	99	99	94
LSTM	99	100	94	99	99	98	93
PatchTST	100	100	100	100	100	100	100
S4D	99	100	96	99	99	98	96
FITS	87	75	21	78	79	77	19
TimesNet	83	87	73	83	82	81	66
Chronos	47	73	51	52	50	50	53
Hybrid	98	100	94	98	97	97	96

tracks (0.12 to 0.25 against 1.06 to 1.11), at 10.5 m/s both recover part of it, at 24.5 m/s both reach 0.5 to 0.6 and the second axis no longer helps. The physics baseline stays within a factor of two at the base in every cell and at 0.00 to 0.17 at the top.

Summary. The ranking does not survive along the tower: the sequence models lie within a few hundredths of each other at the base and spread over three units of R^2 at the top, where a good R^2 can hide the wrong realization. On the redesigns the upper tower governs the fore-aft lifetime damage and is under-predicted. At the top wind speed, not the regime label, sets the error, low wind being the hard condition; waves matter far less than wind and height.

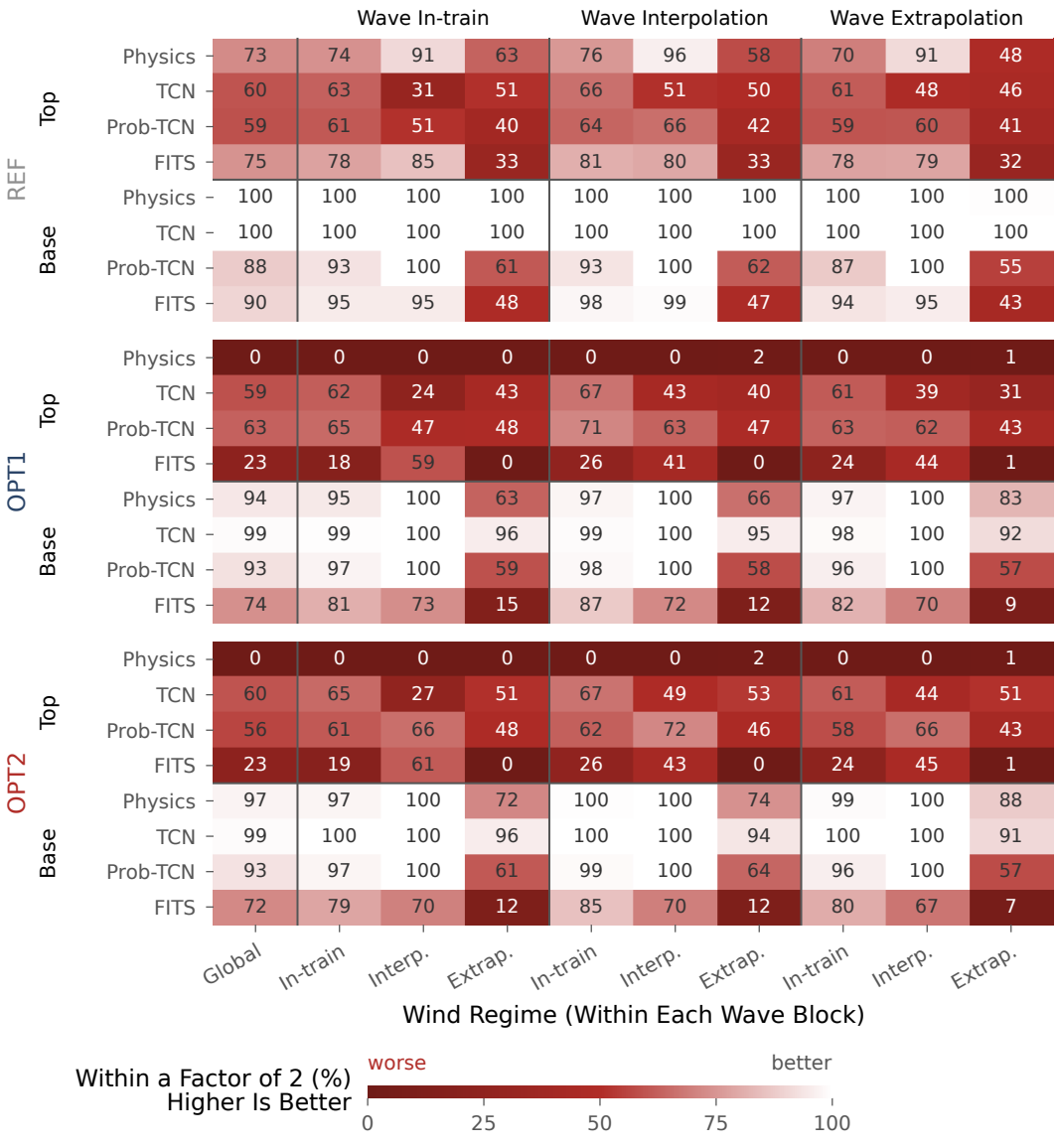


Figure 14: **Regime cells at the top and at the base.** Fraction of test simulations within a factor of two (%), globally and in the nine regime cells (wind regime within each wave block), grouped by tower and, inside each, by section (top, then base; the middle repeats the base), for the physics and the three models that beat the floor at the top; three-seed median (seed 0 for the physics), 50 epochs; red: few simulations within a factor of two.

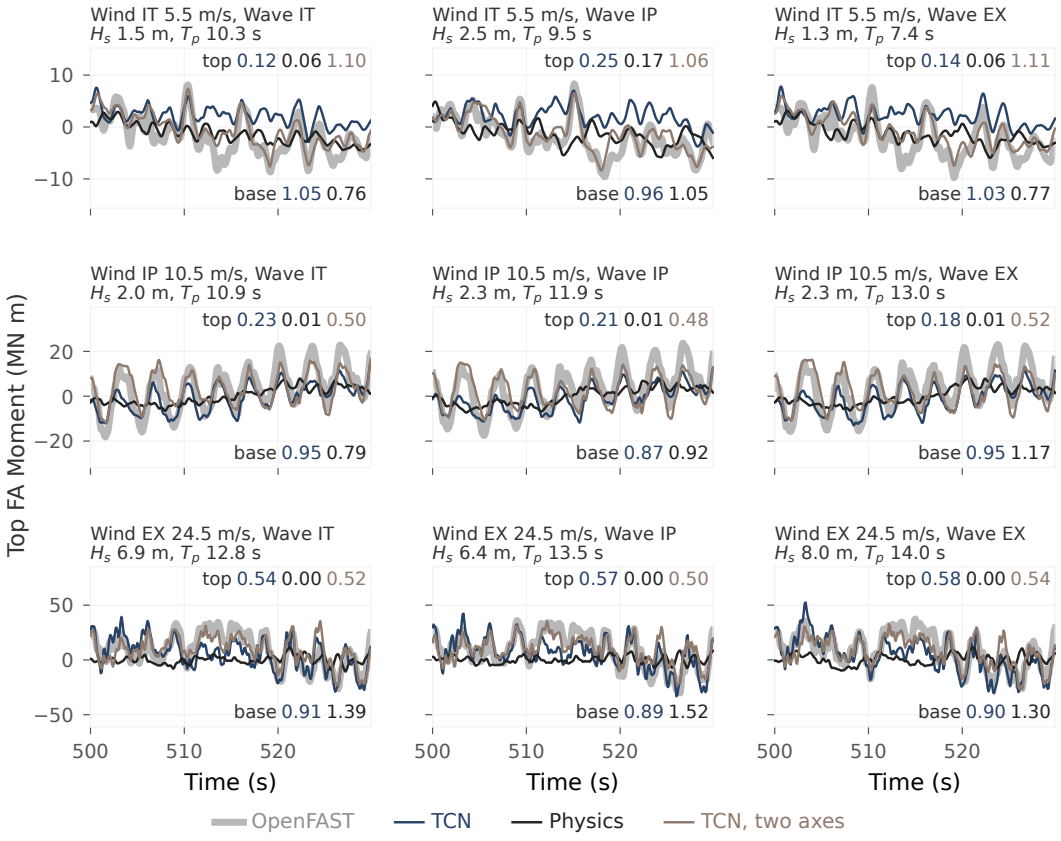


Figure 15: **Top moment across the nine regimes, OPT2.** Thirty seconds of one test simulation per regime cell, within tower, seed 0; rows: wind regime (IT 5.5, IP 10.5, EX 24.5 m/s); columns: wave regime at that wind. Lines: OpenFAST, TCN, physics and the two-axis TCN of the sensor ablation. Numbers: damage ratio $\hat{D}/D$ at the top (upper row) and, for TCN and physics, at the base of the same simulation (lower row).

J IS THE TOP LIMITED BY INFORMATION OR BY TRAINING?

J.1 LONGER TRAINING CLOSES MOST OF THE GAP FOR THE CONVOLUTIONAL MODELS

Every model trains for 50 epochs, a budget fixed before any result was seen (the fixed-budget track); a validation-tuned track (bounded per-model learning-rate and schedule search, validation checkpointing, three seeds for the contrasting models) will be added to the public leaderboard, and the longer probes here are convergence diagnostics, not entries. At 150 epochs (seed 0) the base does not move and the top rises: TCN 0.73 to 0.90, S4D 0.39 to 0.63, Prob-TCN 0.86 to 0.95 (Table 28); PatchTST stays negative (-0.23 at 300 epochs). FITS (top 0.72 to 0.63, z/H 0.78 0.81 to 0.70) and TimesNet (-1.08 to -2.60) get worse: their training is unstable at the recipe's fixed learning rate (final loss 29 to 35% and 23 to 25% above the minimum, against within 7% for the models that improve); Mamba, unstable on OPT2, gains at the top and loses lower down. Over nine runs (Table 27) the median TCN top goes from 0.77 to 0.93 and S4D from 0.45 to 0.64; one Prob-TCN run in nine diverges at 150 epochs (-0.54), none of the 27 at 50. On validation the TCN top reaches 0.655, 0.849 and 0.903 at 50, 150 and 300 epochs, with the lowest validation loss between epochs 230 and 300 in all nine runs: the recipe is short of convergence, not overfitted.

Table 27: **Training budget over the seeds.** Median R^2 of $\log_{10}$ damage over the three towers and the three training seeds, with the worst of the nine runs in parentheses, values below -1 clipped.

	base			top		
Model	50	150	300	50	150	300
TCN	0.99 (0.98)	0.99 (0.99)	0.99 (0.99)	0.77 (0.64)	0.90 (0.88)	0.93 (0.91)
Prob-TCN	0.96 (0.95)	0.99 (< -1)	0.99 (0.97)	0.85 (0.79)	0.91 (-0.54)	0.95 (0.74)
S4D	0.98 (0.97)	0.98 (0.97)	0.99 (0.99)	0.45 (0.27)	0.62 (0.52)	0.64 (0.58)

Table 28: **Training budget, selected models.** R^2 of $\log_{10}$ damage at 50, 150 and 300 epochs, seed 0, mean over the three towers, benchmark split; the 50-epoch column is the leaderboard run. 300 epochs only for TCN, Prob-TCN, S4D and PatchTST ("–": not run).

	base		z/H 0.78		top		top
Model	50	150	50	150	50	150	300
Prob-TCN	0.95	0.99	0.96	0.98	0.86	0.95	0.95
TCN	0.99	1.00	0.96	0.99	0.73	0.90	0.93
LSTM	0.98	0.99	0.93	0.96	0.42	0.62	–
S4D	0.98	0.98	0.92	0.92	0.39	0.63	0.60
FITS	0.76	0.77	0.81	0.70	0.72	0.63	–
PatchTST	0.99	0.99	0.75	0.87	< -1	-0.73	-0.23
TimesNet	0.86	0.89	0.67	0.62	< -1	< -1	–
Chronos	0.78	0.80	0.69	0.73	0.45	0.49	–
Mamba	1.00	0.98	0.92	0.89	-0.30	0.17	–
Hybrid	0.96	0.97	0.82	0.87	< -1	< -1	–

J.2 TRAINING FOR THE TOP ALONE, OR ON A DAMAGE LOSS, DOES THE SAME

Two controls change one element of the recipe, at seed 0: a network trained for the top alone, and the height task with a spectral damage proxy added to the loss (Table 29). Both are diagnostics, not benchmark entries. Trained for the top alone, TCN reaches 0.90 to 0.94 and Prob-TCN 0.97 to 0.98, against 0.64 to 0.87 for the shared model. The damage loss takes S4D from 0.32–0.44 to 0.88–0.94, which the top-only network does not (0.36 to 0.43). PatchTST stays negative under both controls on the redesigns, and Mamba stays at or below 0.17 when trained for the top alone. Part of the top gap is the objective, not the observation.

J.3 THE FAILING MODELS KEEP TOO LITTLE VARIANCE; ADDING IT BACK FIXES THE LEVEL, NOT THE REALIZATION

The spectral diagnostics use OPT2, seed 0, 50 epochs, and twelve test simulations for the spectra (Figure 16). For every model trained on a mean-squared error the damage ratio at the top follows

Table 29: **Controls at the top.** Within tower, seed 0, 50 epochs, task inputs; within 2, median ratio $\hat{D}/D$ and ρ_{wc}: mean over the three towers. *Top only*: trained for the top section alone, without the height channel. *Damage-aware loss*: all heights, damage proxy added to the loss (not run for Prob-TCN and Mamba). *Colored noise*: noise with the residual spectrum added to the TCN output at inference; spectrum and level estimated per height on 96 evenly spaced training simulations, never on the test split. Prob-TCN *sample*: the protocol output; *mean head*: its predicted mean.

		Top R^2			Top, mean over towers		
Model	Condition	REF	OPT1	OPT2	within 2	median ratio	ρ_{wc}
TCN	all heights (protocol)	0.83	0.64	0.71	0.47	0.54	0.69
	top only	0.90	0.94	0.93	0.74	0.68	0.82
	damage-aware loss	0.66	0.84	0.82	0.17	0.29	0.63
	colored noise	0.95	0.88	0.88	0.76	1.01	0.68
Prob-TCN	all heights, sample (protocol)	0.86	0.87	0.85	0.61	1.59	0.66
	all heights, mean head	0.73	0.68	0.82	0.26	0.42	0.73
	top only	0.97	0.98	0.98	0.93	1.08	0.81
S4D	all heights (protocol)	0.32	0.40	0.44	0.02	0.15	0.51
	top only	0.36	0.41	0.43	0.00	0.25	0.62
	damage-aware loss	0.88	0.93	0.94	0.61	0.59	0.71
PatchTST	all heights (protocol)	-0.94	<-1	<-1	0.00	0.02	0.48
	top only	-0.18	-0.79	-0.45	0.00	0.07	0.46
	damage-aware loss	0.89	-0.96	-0.12	0.24	0.19	0.48
Mamba	all heights (protocol)	-0.43	-0.27	-0.19	0.00	0.08	0.51
	top only	0.17	-0.03	-0.88	0.01	0.10	0.43

the variance ratio to a power of 1.5 to 2.5: TCN keeps 0.70 of the variance and 0.55 of the damage, Mamba 0.25 and 0.06, PatchTST 0.15 and 0.015. Prob-TCN is scored on one draw of its predictive distribution (the benchmark estimator: at every time step the predicted mean plus independent Gaussian noise with the predicted standard deviation, in standardized units, one draw per simulation and height, fixed random seed). Its mean head keeps 0.66 of the variance, like the TCN, and scores 0.73 / 0.68 / 0.82 at the top (Table 29); the sampled broadband variance restores the damage and overshoots it (ratio 1.70). Over twenty draws per simulation and height (seed 0) the top R^2 is 0.857 to 0.859 / 0.872 to 0.874 / 0.851 to 0.853, and the expected damage over the draws scores the same (0.858 / 0.873 / 0.852) and keeps the over-prediction (lifetime ratio 1.49 / 1.31 / 1.60): the overshoot is a bias of the predictive distribution, not Monte Carlo noise. Noise with the residual spectrum added to the TCN output does the same: the top R^2 rises from 0.83 / 0.64 / 0.71 to 0.95 / 0.88 / 0.88 (0.96 / 0.89 / 0.89 for the expected damage over 20 draws; single-draw spread at most 0.003) (Table 29), while the within-condition correlation shows the level restored, not the realization (Appendix I.2). Without the 3 Hz low-pass of the score the deterministic models are unchanged and the sampled Prob-TCN falls to 0.57 at the top of OPT2, since the rainflow count turns its broadband noise into cycles. Sampling beats the mean head at the top only at some seeds: at seed 1 the mean head is higher on REF (0.81 against 0.79) and OPT2 (0.90 against 0.87).

J.4 The missing variance is the 3P band below rated wind

At the top the moment is quasi-static thrust plus the 3P rotor harmonic (48% and 26% of the variance). Below rated wind the 3P window falls inside the wave band, where the coherence of the fore-aft acceleration with the top moment is 0.11 at 5.5 m/s and 0.38 at 9.5 m/s, against 0.86 to 0.91 above rated; the side-side axis keeps 0.67 to 0.75 at every wind (Figure 16c). The TCN keeps 0.22 of the 3P variance below rated and 1.03 above; with both axes its top R^2 below rated rises from -0.03 to 0.92. Figure 17 shows the physics side: scaled by its height factor, its gain is 14 times too low in the 3P window, a rotor load that the height factor does not model. PatchTST and Mamba lose the 3P band above rated wind too (ratio 0.03 to 0.12), where the fore-aft coherence is 0.9, so for them the top is a failure of the model, not of the information.

Summary. Part of the top gap is training and part is information. Longer training, training for the top alone or a damage loss each close most of it for the convolutional models. The failing models keep too little variance, and adding it back fixes the level, not the realization. The missing variance

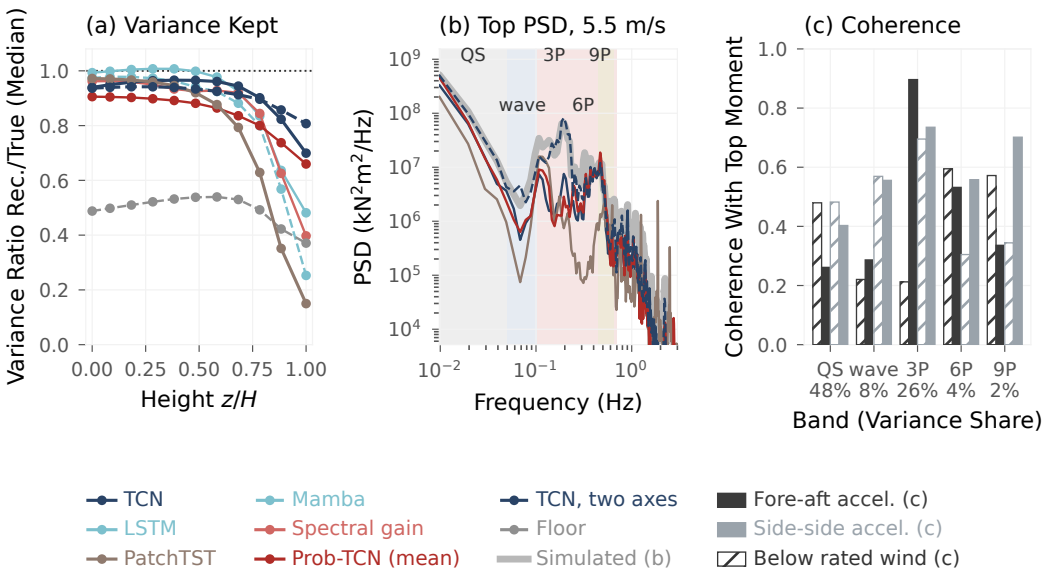


Figure 16: **Why the top fails, OPT2.** (a) Median reconstructed-to-true variance ratio along the tower, full test split. (b) Welch spectrum of the true top moment at 5.5 m/s with the reconstructions of TCN, PatchTST, Prob-TCN (mean) and the two-axis TCN; bands shaded. (c) Magnitude-squared coherence of the true top moment with the fore-aft and the side-side acceleration per band, below and above rated wind, with the variance share of each band.

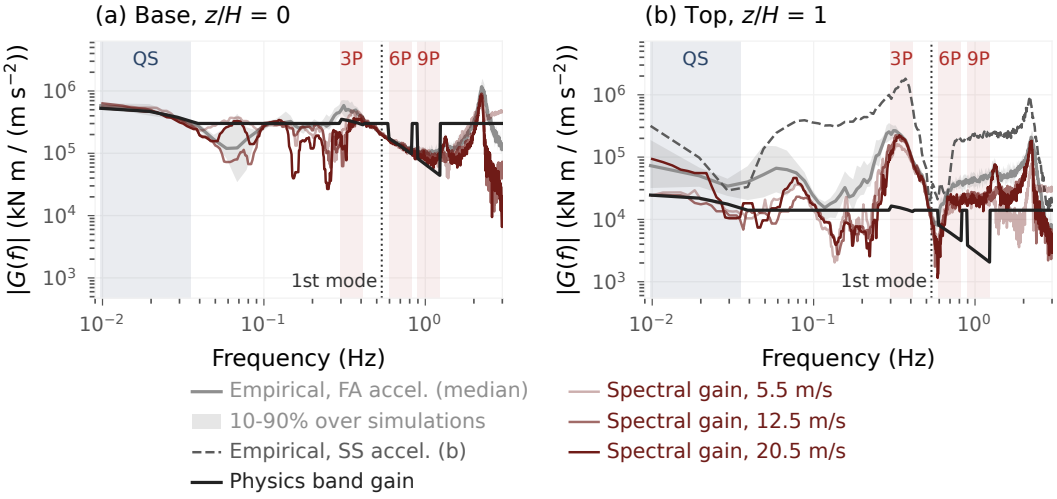


Figure 17: **Transfer functions from the tower-top acceleration to the moment, OPT2.** Empirical $\sqrt{S_{MM}/S_{aa}}$ (median and 10 to 90% band over twelve test simulations), the physics band gain (at the top scaled by its height factor) and the learned spectral gain at three wind speeds, at the base (a) and at the top (b), where the transfer function from the side-side acceleration is dashed. The quasi-static band and the 3P, 6P and 9P windows are shaded; the dotted line is the first fore-aft mode.

is the 3P band below rated wind, where the fore-aft acceleration is weakly coherent with the top moment and the side-side one is not.

K How should the top metrics and the transfer be read? (controls)

Six controls test how the metrics of Section 5 should be read. Within tower they use the test split (4,740 simulations: 790 operating points × six realizations per tower), the fore-aft moment and the three-seed median of Section 5 (†seed 0 only; protocol variants at seed 0). Intervals are 95% percentile cluster bootstraps over the 790 operating points (Appendix G.4; $B = 2,000$, six realizations resampled together, the seed median taken in each resample).

K.1 A condition-only predictor wins the top in aggregate R^2

Definitions. The condition-only oracle predicts, per test simulation and height, the arithmetic mean of the true damage over the six realizations of its operating point (leave-one-out variant: over the other five); it reads the test labels, so it is a bound on any condition-only predictor, not a model. The condition-only surrogate is an XGBoost (500 trees, depth 4, learning rate 0.05, subsample 0.8, seed 0) from wind speed U, H_s and T_p to $\log_{10}$ damage, trained as in FLOATBench on the 1,728 training simulations of the tower; without sensor input it has $\rho_{\mathrm{wc}} = 0$, and it is given the true H_s and T_p, which the FLOATSense sensor set lacks (in the field: a wave buoy or a hindcast).

Finding. The within-condition share of the log-damage variance is 2 to 6%, so the oracle scores 0.94 to 0.98 and the surrogate 0.91 to 0.96 (Table 30). At the base TCN, PatchTST and Mamba beat both (0.99 to 1.00, 99 to 100% within a factor of two); at the top every sensor model is below them, the best being the physics on REF (0.93) and Prob-TCN on the redesigns (0.85), and every one of the 22 entries of Tables 13 to 15 is below 0.95. A top R^2 is thus evidence of the level, not of the realization, which ρ_{wc} (Table 23) and the slope of Appendix K.2 measure (the leave-one-out oracle has ρ_{wc} of -0.95 to -0.98 by construction). The surrogate is weakest, like Prob-TCN, where the wind speed is interpolated (top R^2 0.47 to 0.89). Configurations off the declared protocol approach or reach the oracle: the two-axis TCN scores 0.90 / 0.96 / 0.96 at the top and the top-only Prob-TCN 0.97 / 0.98 / 0.98 (seed 0).

Table 30: **Condition-only controls.** R^2 of $\log_{10}$ damage and fraction within a factor of two at the base and the top, within tower, test split. Oracle and XGBoost surrogate as defined above (the surrogate fed the true H_s and T_p). Models: median of three seeds, †seed 0.

	R^2 of $\log_{10}$ damage						within a factor of two					
	base			top			base			top		
Predictor	REF	OPT1	OPT2	REF	OPT1	OPT2	REF	OPT1	OPT2	REF	OPT1	OPT2
Oracle, mean of six	0.97	0.94	0.94	0.97	0.98	0.98	0.97	0.91	0.92	0.94	0.92	0.91
Oracle, leave one out	0.96	0.92	0.92	0.96	0.97	0.97	0.95	0.87	0.88	0.90	0.88	0.88
XGBoost on (U, H_s, T_p)	0.96	0.91	0.91	0.95	0.95	0.95	0.93	0.89	0.89	0.93	0.88	0.86
Floor†	0.80	0.69	0.73	0.75	0.70	0.65	0.73	0.67	0.68	0.60	0.17	0.18
Physics†	0.97	0.94	0.96	0.93	<-1	<-1	1.00	0.94	0.97	0.73	0.00	0.00
TCN	1.00	0.99	0.99	0.77	0.77	0.73	1.00	0.99	0.99	0.60	0.59	0.60
Prob-TCN, sample	0.96	0.96	0.96	0.82	0.85	0.85	0.88	0.93	0.93	0.59	0.63	0.56
Prob-TCN, mean head†	1.00	0.98	0.98	0.73	0.68	0.82	1.00	0.97	0.99	0.32	0.02	0.45
PatchTST	1.00	0.99	0.99	-0.94	<-1	<-1	1.00	1.00	1.00	0.00	0.00	0.00
Mamba	1.00	0.99	0.99	-0.38	-0.27	-0.19	1.00	1.00	1.00	0.00	0.00	0.00

K.2 At the top every model attenuates the within-condition deviation

Definition. With δ_i the deviation of $\log_{10} D_i$ from the mean over the six realizations of its operating point and $\hat{\delta}_i$ the same for the reconstruction, the within-condition calibration slope is the OLS slope of $\hat{\delta}$ on δ, $b = \sum_i \delta_i \hat{\delta}_i / \sum_i \delta_i^2$ (both deviations have zero mean). $b = \rho_{\mathrm{wc}}\, \mathrm{sd}(\hat{\delta})/\mathrm{sd}(\delta)$; $b = 1$ is the right within-condition amplitude, $b = 0$ a condition-only prediction.

Finding. At the base TCN, PatchTST, Mamba and the Prob-TCN mean head have slopes of 0.92 to 1.06, the sampled Prob-TCN 0.80 to 0.82 and the floor 0.60 to 0.89 (Table 31). At the top every benchmark model is attenuated (0.31 to 0.85): the sampled Prob-TCN is lowest (0.31 / 0.51 / 0.47), below its own mean head (0.60 / 0.85 / 0.75) and the TCN (0.61 / 0.75 / 0.63), so the sampled noise

restores the damage level and dilutes the realization. Among the protocol variants the two-axis TCN comes closest to a unit slope at the top (0.92 / 1.00 / 0.83); the SCADA-only Prob-TCN has 0.18 to 0.34 at a top R^2 of 0.76 to 0.91 (level without realization); the top-only Prob-TCN (R^2 0.97 to 0.98) keeps 0.59 to 0.64.

Table 31: **Within-condition calibration slope** b [95% cluster-bootstrap interval], within tower. Upper block: benchmark models, median of three seeds ([†]seed 0). Lower block: protocol variants of Appendix J.2 and Appendix M, seed 0; "–": top-only models are not scored at the base.

	base			top		
Model	REF	OPT1	OPT2	REF	OPT1	OPT2
Floor[†]	0.89 [0.87, 0.91]	0.60 [0.58, 0.62]	0.62 [0.60, 0.65]	0.53 [0.50, 0.57]	0.52 [0.49, 0.54]	0.47 [0.45, 0.50]
Physics[†]	0.94 [0.93, 0.96]	0.76 [0.74, 0.78]	0.79 [0.77, 0.80]	0.57 [0.55, 0.60]	0.49 [0.46, 0.52]	0.46 [0.43, 0.49]
TCN	0.95 [0.94, 0.96]	0.96 [0.94, 0.97]	0.96 [0.95, 0.97]	0.61 [0.57, 0.66]	0.75 [0.70, 0.81]	0.63 [0.56, 0.70]
Prob-TCN, sample	0.80 [0.78, 0.82]	0.82 [0.81, 0.84]	0.82 [0.81, 0.84]	0.31 [0.28, 0.35]	0.51 [0.48, 0.53]	0.47 [0.44, 0.50]
Prob-TCN, mean head[†]	0.95 [0.94, 0.96]	0.92 [0.90, 0.94]	1.06 [1.05, 1.07]	0.60 [0.56, 0.64]	0.85 [0.82, 0.89]	0.75 [0.71, 0.79]
PatchTST	0.94 [0.93, 0.95]	0.92 [0.91, 0.93]	0.95 [0.94, 0.95]	0.51 [0.48, 0.55]	0.49 [0.46, 0.52]	0.53 [0.48, 0.56]
Mamba	0.96 [0.95, 0.97]	0.96 [0.96, 0.97]	0.93 [0.93, 0.94]	0.59 [0.56, 0.62]	0.81 [0.76, 0.86]	0.81 [0.77, 0.83]
TCN, two axes[†]	0.95 [0.94, 0.96]	1.05 [1.03, 1.06]	0.97 [0.97, 0.98]	0.92 [0.90, 0.94]	1.00 [0.99, 1.02]	0.83 [0.79, 0.86]
TCN, top only[†]	–	–	–	0.92 [0.89, 0.95]	0.83 [0.79, 0.88]	0.77 [0.70, 0.84]
TCN, SCADA only[†]	1.13 [1.11, 1.16]	1.14 [1.12, 1.16]	1.09 [1.07, 1.12]	0.52 [0.45, 0.60]	0.90 [0.82, 0.99]	0.80 [0.73, 0.88]
Prob-TCN, two axes[†]	0.80 [0.77, 0.82]	0.85 [0.84, 0.86]	0.76 [0.74, 0.78]	0.46 [0.44, 0.48]	0.57 [0.55, 0.59]	0.55 [0.52, 0.58]
Prob-TCN, top only[†]	–	–	–	0.64 [0.60, 0.67]	0.59 [0.54, 0.64]	0.63 [0.59, 0.67]
Prob-TCN, SCADA only[†]	0.49 [0.45, 0.54]	0.69 [0.67, 0.71]	0.65 [0.61, 0.68]	0.18 [0.16, 0.20]	0.34 [0.32, 0.36]	0.31 [0.28, 0.33]

K.3 THE LIFETIME BIASES ARE NOT SAMPLING NOISE

Definition. The ratio $\sum_i w_i \hat{D}_i / \sum_i w_i D_i$ over the test split with the FLOATBench lifetime weights w_i, as in Table 24 (seed 0); the condition-only oracle has ratio 1 by construction.

Finding. The intervals are never wider than ± 0.03 (Table 32), so the biases of Table 24 are not sampling noise: at the top the TCN keeps 0.61 to 0.66 of the lifetime damage, the Prob-TCN mean head 0.38 to 0.55, the sampled Prob-TCN over-predicts by 44 to 74%, and PatchTST and Mamba keep at most 8%; at the base the TCN is within 8%.

Table 32: **Lifetime-weighted damage ratio** [95% cluster-bootstrap interval], within tower, median of three seeds ([†]seed 0).

	base			top		
Model	REF	OPT1	OPT2	REF	OPT1	OPT2
Floor[†]	0.97 [0.96, 0.98]	0.80 [0.77, 0.83]	0.77 [0.75, 0.80]	0.70 [0.69, 0.71]	0.21 [0.20, 0.21]	0.18 [0.17, 0.18]
Physics[†]	1.44 [1.43, 1.45]	1.11 [1.10, 1.13]	1.10 [1.08, 1.11]	0.68 [0.68, 0.69]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
TCN	0.99 [0.99, 0.99]	0.98 [0.97, 0.98]	0.92 [0.92, 0.92]	0.66 [0.66, 0.67]	0.61 [0.61, 0.62]	0.63 [0.62, 0.63]
Prob-TCN, sample	1.08 [1.08, 1.09]	1.08 [1.07, 1.08]	1.11 [1.10, 1.11]	1.74 [1.71, 1.77]	1.44 [1.42, 1.45]	1.60 [1.57, 1.63]
Prob-TCN, mean head[†]	0.97 [0.97, 0.97]	0.86 [0.85, 0.86]	0.84 [0.84, 0.85]	0.49 [0.48, 0.49]	0.38 [0.37, 0.38]	0.55 [0.54, 0.55]
PatchTST	1.00 [1.00, 1.00]	0.96 [0.95, 0.96]	0.86 [0.85, 0.86]	0.04 [0.04, 0.04]	0.01 [0.01, 0.01]	0.02 [0.02, 0.02]
Mamba	1.07 [1.06, 1.07]	1.06 [1.05, 1.06]	1.05 [1.05, 1.06]	0.08 [0.07, 0.08]	0.07 [0.07, 0.07]	0.07 [0.07, 0.07]

K.4 ONE SCALE FACTOR PER HEIGHT FROM FIVE TARGET SIMULATIONS RECOVERS THE TOP

Definition. The source checkpoint is run unchanged (source normalization, scored window, 3 Hz low-pass, target S–N damage) on target training simulations, reproducing the stored zero-shot TCN damage exactly on the checked simulations (the sampled Prob-TCN draw is not bit-reproducible, so its scales use a fresh draw). Per height h, $s_h = \mathrm{median}_i(D_i/\hat{D}_i)$ over these calibration simulations rescales the stored zero-shot test predictions, with no gradient step. Calibration sets: the five simulations of the $N = 5$ adaptation (draw 0), the 18 designed simulations of Table 38, and the test split itself (oracle, the bound of any single scale). Intervals hold the scale fixed; the seed spread shows the calibration uncertainty.

Finding. One factor per height from five simulations brings the TCN top from its zero-shot value to 0.75 to 0.94 on every pair (Table 33; 0.91 in the seed median, Table 34), above its fine-tuning at the top on all six pairs at $N = 5$ (0.11 to 0.79), on the 18 designed simulations (0.38 to 0.90) and at every budget up to $N = 100$ (Table 39); the lowest, OPT2→REF, has the interval [0.73, 0.77]. Over pairs and seeds the five-simulation scale gives the TCN top 0.75 to 0.95 and Prob-TCN 0.48 (OPT1→REF, seed 0) to 0.97; the scaled floor also reaches 0.89 to 0.92 (seed 0), so the scaled top R^2 again scores mostly the level. At the base the scale matches or beats fine-tuning between the redesigns and from REF (0.97 to 0.99) but stays at 0.83 into REF, against 0.97 to 0.99 fine-tuned; even the oracle scale gives 0.84 to 0.85 there, so the residual error into REF is condition-dependent, which fine-tuning captures. Five calibration simulations all below rated wind (the first five designed ones) do worse (TCN top 0.79 against 0.91, seed median): the calibration set must span the wind speeds.

Table 33: **Scale-only transfer, TCN, seed 0.** R^2 of $\log_{10}$ damage on the target test split: zero-shot, one scale per height fitted on 5 ($N = 5$, draw 0), 18 (designed) or the test split (oracle), and fine-tuning on the same 5 and 18 (Table 38, Table 39).

	base						top					
		one scale per height			fine-tuning			one scale per height			fine-tuning	
Pair	zero-shot	5	18	oracle	5	18	zero-shot	5	18	oracle	5	18
REF→OPT1	0.67	0.97	0.97	0.97	0.80	0.94	0.81	0.94	0.93	0.94	0.11	0.90
REF→OPT2	0.72	0.98	0.98	0.98	0.86	0.91	0.76	0.94	0.94	0.94	0.52	0.89
OPT1→OPT2	0.99	0.99	0.99	0.99	0.89	0.96	0.87	0.89	0.90	0.89	0.79	0.74
OPT1→REF	−0.34	0.83	0.84	0.85	0.97	0.99	<-1	0.88	0.91	0.91	0.18	0.38
OPT2→REF	−0.14	0.83	0.83	0.84	0.98	0.99	<-1	0.75	0.83	0.83	0.14	0.59
OPT2→OPT1	0.97	0.99	0.99	0.99	0.94	0.97	0.56	0.86	0.84	0.85	0.57	0.77

Table 34: **Scale-only transfer, seed median.** Median over the six pairs of the three-seed median: R^2 of $\log_{10}$ damage and fraction within a factor of two, zero-shot and after one scale per height from 5, 18 or the test split (oracle).

		R^2 of $\log_{10}$ damage				within a factor of two			
Model	Height	zero-shot	5	18	oracle	zero-shot	5	18	oracle
TCN	base	0.70	0.97	0.97	0.97	0.29	0.99	0.99	0.99
	top	0.75	0.91	0.91	0.91	0.25	0.69	0.74	0.75
Prob-TCN	base	0.81	0.93	0.94	0.94	0.52	0.88	0.91	0.91
	top	0.83	0.90	0.91	0.92	0.45	0.71	0.74	0.73

K.5 THE TOP DAMAGE RANKING IS NOT A WAVEFORM RANKING

Definition. On 200 test simulations per tower drawn at random with a fixed generator seed, the true and reconstructed moment series (kN m, zero mean, 3 Hz low-pass as in the score, 400 to 1,000 s) give per simulation and height the Pearson r, the NRMSE $= \mathrm{RMS}(\hat{M} - M)/\mathrm{sd}(M)$ and the variance ratio $\mathrm{var}(\hat{M})/\mathrm{var}(M)$, as the median over simulations, then over seeds (three for TCN and Prob-TCN, seed 0 for PatchTST and Mamba), ranged over the three towers.

Finding. All five models track the base waveform ($r \geq 0.98$, NRMSE ≤ 0.19; Table 35). At the top only the TCN keeps r of 0.85 to 0.87 (NRMSE 0.52 to 0.54); the sampled Prob-TCN has a worse waveform than its own mean head (r 0.72 to 0.75 against 0.80 to 0.82) while scoring the best top damage, and PatchTST and Mamba fall to r 0.35 to 0.53 with 14 to 25% of the variance. TCN and the Prob-TCN mean head are in phase with the true moment in every band and height (median |phase| at most 8°); at the top they lose coherence in the wave band (0.29 to 0.32) and in the 3P band below rated (0.27 to 0.28), not phase. The physics baseline has 48 to 62° of phase error in the 3P window and 88° in the wave band at the top.

Table 35: **Waveform metrics**, within tower, 200 test simulations per tower: Pearson r and NRMSE at four heights, variance ratio at the top; median over simulations, then seeds; ranges: over the three towers.

	Pearson r				NRMSE				variance ratio
Model	base	z/H 0.48	z/H 0.78	top	base	z/H 0.48	z/H 0.78	top	top
TCN	0.99	0.98–0.99	0.95–0.97	0.85–0.87	0.14–0.19	0.15–0.21	0.27–0.32	0.52–0.54	0.73–0.78
Prob-TCN, mean head	0.99	0.98–0.99	0.95–0.97	0.80–0.82	0.13–0.16	0.15–0.20	0.27–0.33	0.58–0.61	0.65–0.68
Prob-TCN, sample	0.98–0.99	0.98–0.99	0.93–0.95	0.72–0.75	0.15–0.19	0.17–0.23	0.32–0.39	0.69–0.72	0.81–0.83
PatchTST	0.98	0.96–0.97	0.81–0.89	0.35–0.37	0.18	0.23–0.29	0.47–0.59	0.93–0.94	0.14–0.15
Mamba	0.98–0.99	0.96–0.97	0.86–0.90	0.52–0.53	0.18	0.23–0.26	0.44–0.52	0.86	0.24–0.25

K.6 Sensor errors do not change the base or the top

Definition. Inference only, no retraining: the inputs of TCN and the Prob-TCN mean head (seed 0) are perturbed before normalization, on 100 evenly spaced test operating points per tower with all six realizations (600 simulations). Perturbations: white noise on the gravity-corrected fore-aft acceleration at 5 and 10% of its window standard deviation; a 0.02 m/s^2 bias plus a drift of 0.02 m/s^2 over the run; the hub wind replaced by a trailing 60 s mean plus white noise of 0.5 m/s, a nacelle-anemometer proxy; and all three together. The degraded wind also enters the condition vector. *Finding.* The base stays saturated ($R^2 \geq 0.976$, changes of at most 0.006) and the top stays the hard, under-predicted height (median ratio 0.34 to 0.73; ρ_{wc} changes of at most 0.05). Bias and drift lower the TCN top on every tower (by 0.01 to 0.03); white noise raises the top damage only by offsetting its under-prediction; the degraded hub wind moves every metric by at most 0.03, consistent with SCADA setting the level and the accelerometer the realization.

Table 36: **Sensor-error stress test.** Base R^2 of $\log_{10}$ damage and, at the top, R^2, fraction within a factor of two and within-condition correlation, 600 test simulations per tower, seed 0, inference only.

Tower	Model	Metric	clean	noise 5%	noise 10%	bias+drift	hub wind	all
REF	TCN	base R^2	0.996	0.996	0.997	0.995	0.996	0.995
		top R^2	0.83	0.83	0.85	0.82	0.84	0.84
		top within 2	0.59	0.59	0.60	0.59	0.60	0.59
		top ρ_{wc}	0.70	0.70	0.70	0.69	0.68	0.67
	Prob-TCN (mean head)	base R^2	0.998	0.998	0.998	0.997	0.997	0.996
		top R^2	0.74	0.75	0.76	0.75	0.76	0.78
		top within 2	0.30	0.34	0.45	0.34	0.33	0.40
		top ρ_{wc}	0.66	0.66	0.65	0.67	0.66	0.68
OPT1	TCN	base R^2	0.979	0.979	0.981	0.980	0.977	0.979
		top R^2	0.65	0.66	0.71	0.62	0.65	0.64
		top within 2	0.26	0.30	0.45	0.18	0.29	0.25
		top ρ_{wc}	0.72	0.73	0.73	0.67	0.71	0.67
	Prob-TCN (mean head)	base R^2	0.976	0.978	0.981	0.976	0.977	0.979
		top R^2	0.69	0.70	0.74	0.68	0.71	0.71
		top within 2	0.03	0.03	0.08	0.03	0.05	0.05
		top ρ_{wc}	0.82	0.82	0.82	0.81	0.81	0.80
OPT2	TCN	base R^2	0.989	0.990	0.991	0.990	0.990	0.990
		top R^2	0.72	0.73	0.75	0.71	0.72	0.72
		top within 2	0.57	0.58	0.60	0.53	0.56	0.55
		top ρ_{wc}	0.59	0.59	0.60	0.56	0.61	0.56
	Prob-TCN (mean head)	base R^2	0.980	0.982	0.986	0.977	0.979	0.979
		top R^2	0.83	0.84	0.86	0.83	0.84	0.85
		top within 2	0.44	0.49	0.61	0.48	0.47	0.54
		top ρ_{wc}	0.71	0.71	0.71	0.69	0.71	0.71

Summary. At the top aggregate R^2 scores the condition level: both condition-only controls sit above every sensor model, and every model attenuates the within-condition deviation. The lifetime biases are not sampling noise, one scale per height from five target simulations recovers the TCN top on every pair, the top damage ranking is not a waveform ranking, and sensor errors change neither the base nor the top.

L DOES THE WINNER SURVIVE A CHANGE OF TOWER? (CROSS TOWER)

L.1 ZERO-SHOT KEEPS THE RANKING AND MISSES THE LEVEL

Table 37 shows how zero-shot fails: the within-condition correlation stays at 0.84 to 0.95 on every pair, so the models rank the realizations of the target tower correctly, while the damage level is off by a factor of about 2.5 into REF and about 0.3 out of it.

Table 37: **Zero-shot diagnosis at the base.** Per ordered pair, median over the 20 learned models of each model's median over its seeds (three for thirteen models, seed 0 for seven): R^2 of $\log_{10}$ damage, median damage ratio $\hat{D}/D$ and within-condition correlation ρ_{wc}.

Pair	R^2	median ratio	ρ_{wc}
OPT1→OPT2	0.93	0.93	0.92
OPT2→OPT1	0.84	0.70	0.95
OPT1→REF	0.75	2.59	0.84
OPT2→REF	0.71	2.45	0.85
REF→OPT1	0.41	0.30	0.92
REF→OPT2	0.48	0.34	0.92

L.2 FINE-TUNING ON FIVE TARGET SIMULATIONS RECOVERS THE BASE; A SCALE PER HEIGHT DOES BETTER AT THE TOP

Figure 18 follows adaptation up the tower: ten target simulations bring the base and the mid tower to within 0.15 of the within-tower level on every pair, and the mean-squared-error models stop at z/H 0.78. Table 38 and Table 39 give the curve in N; adaptation runs 50 epochs at batch 4, so the steps grow with N and the curve does not separate the two. At the base the curve of the six sequence models (TCN, Prob-TCN, LSTM, S4D, PatchTST, Mamba) is flat from $N = 5$ on (0.90 to 0.99, mean over the five pairs of Figure 3c; 0.87 to 1.00 on OPT2→OPT1), and the draw spread at a fixed N (up to 0.16 for PatchTST at $N = 10$) is larger than the gain from $N = 5$ to 100; the other four need more simulations: the hybrid rises from 0.22 at $N = 5$ to 0.97 at 100, FITS from 0.18 to 0.58, TimesNet and Chronos stay between 0.55 and 0.88; at the top recovery (a top R^2 within 0.05 of the model's own within-tower value, three-seed median; curves at seed 0, draw 0) depends on model, pair and budget: Prob-TCN, already at 0.80 to 0.89 zero-shot into the redesigns (seed 0; 0.74 to 0.91 over three seeds), recovers them at $N = 5$ (0.84 to 0.89) and 25 (0.86 to 0.93) but not at 100 (0.67 to 0.73; OPT2→OPT1: 0.88 zero-shot, 0.71, 0.88 and 0.79 at $N = 5$, 25 and 100), and REF only at some budgets of 50 or more; the TCN recovers some pairs and budgets (0.89 to 0.90 on the designed set from REF); no model recovers every pair at every budget. What transfers is the source model: the TCN trained from scratch on the same ten target simulations scores 0.00 to 0.14 at the base of the redesigns, against 0.91 to 0.99 after adaptation. One scale factor per height fitted on the same five simulations, with no gradient step, does better at the top than every fine-tuning budget of the TCN (Appendix K.4). On every pair of Figure 18 the TCN comes within 0.04 of its within-tower level at the base and mid tower, and Prob-TCN carries its profile to the top into the redesigns, not into REF.

The designed sea state of Table 38 (H_s index 3, or 5 at 15.5, 16.5 and 23.5 m/s; T_p index 6 of 7; first realization) is a training one, not the grid middle (index 4) of the review subset: `sim_id` 314, 608, 902, 1196, 1490, 1784, 2372, 2666, 2960, 3254, 3562, 3856, 4430, 4724, 5018, 5312, 5606 and 5914, all, like every random draw, in the target training split and none in its test split.

Recalibrated on the same target simulations, the physics reaches 0.94 to 0.97 at the base of the redesigns from $N = 10$ (with five, the harmonic damping cannot be identified on two of three draws) and stays below -1 at their top at every N, as with the full training split: its top failure is the model form, not the calibration budget.

Summary. Zero-shot keeps the ranking and misses the level. Ten target simulations bring the base and the mid tower close to the within-tower level on every pair, but the mean-squared-error models stop at z/H 0.78, and one scale per height does better at the top; the recalibrated physics fails at the top of the redesigns at every N, a failure of model form, not of calibration budget.

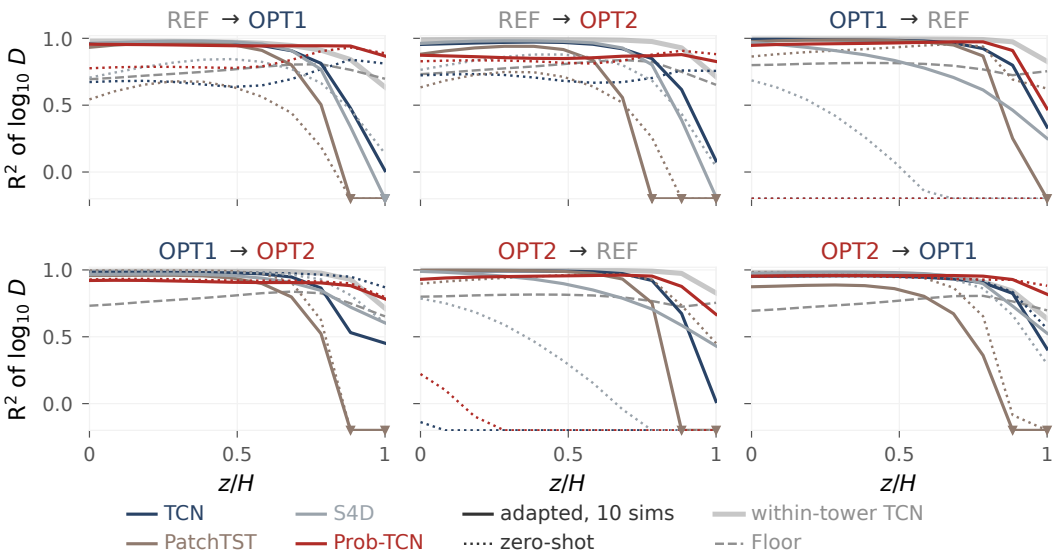


Figure 18: **Adaptation along the tower.** R^2 of $\log_{10}$ damage against height on the target test split, one panel per ordered pair (source $\rightarrow$ target, names in the tower colors): zero-shot (dotted) and after adaptation on ten target simulations (solid, draw 0) for TCN, PatchTST, S4D and Prob-TCN; thick gray: within-tower TCN of the target; dashed gray: floor; triangles: values below the axis.

Table 38: **Adaptation curve, base.** R^2 of $\log_{10}$ damage on the target test split after adapting a 50-epoch source model on N target simulations (50 epochs, batch 4, so the steps in the second header row grow with N); $N = 0$: zero-shot; $N = 10$: the ten-shot protocol. "18 (designed)": a fixed set of 18 target training simulations, one per training wind speed at the median sea state, instead of a random draw (`sim_id` list below the table). "–": run not completed for this release (on OPT2→OPT1 only TCN and Prob-TCN have the full curve). Seed 0, draw 0.

	Target simulations N						
Model	0 (zero-shot)	5	10	25	50	100	18 (designed)
Adaptation steps	0	50	100	300	600	1,250	200
REF → OPT1							
TCN	0.67	0.80	0.96	0.98	0.87	0.98	0.94
Prob-TCN	0.78	0.95	0.96	0.98	0.92	0.88	0.87
FITS	0.72	0.44	0.72	0.67	0.74	0.80	0.66
LSTM	0.47	0.91	0.90	0.91	0.95	0.95	0.97
S4D	0.71	0.87	0.96	0.96	0.97	0.97	0.97
Chronos	0.02	–	0.54	–	–	–	–
TimesNet	0.63	0.34	0.34	0.91	0.92	0.60	0.44
Hybrid	0.27	–	0.82	0.86	0.90	0.96	0.83
PatchTST	0.54	0.96	0.93	0.86	0.94	0.94	0.96
Mamba	0.81	0.97	0.99	0.96	0.97	1.00	0.98
REF → OPT2							
TCN	0.72	0.86	0.95	0.98	0.95	0.98	0.91
Prob-TCN	0.83	0.94	0.87	0.98	0.86	0.91	0.89
FITS	0.71	0.44	0.71	0.67	0.73	0.79	0.64
LSTM	0.54	0.94	0.88	0.91	0.96	0.96	0.98
S4D	0.76	0.89	0.96	0.98	0.97	0.97	0.97
Chronos	0.19	0.40	0.63	0.84	0.55	0.69	0.59
TimesNet	0.69	0.49	0.54	0.86	0.85	0.93	0.84
Hybrid	0.35	–	0.84	0.88	0.90	0.97	0.85
PatchTST	0.63	0.97	0.88	0.90	0.91	0.92	0.97
Mamba	0.86	0.98	0.99	0.95	0.97	1.00	0.99
OPT1 → OPT2							
TCN	0.99	0.89	0.98	0.98	0.96	0.99	0.96
Prob-TCN	0.93	0.95	0.92	0.97	0.98	0.92	0.96
FITS	0.62	0.55	0.65	0.60	0.65	0.60	0.50
LSTM	0.98	0.97	0.94	0.96	0.92	0.96	0.99
S4D	0.99	0.94	0.97	0.99	0.98	0.97	0.97
Chronos	0.80	0.69	0.81	0.92	0.79	0.86	0.81
TimesNet	0.82	0.79	0.71	0.71	0.11	0.95	0.81
Hybrid	0.98	–	0.96	0.97	0.94	0.96	0.96
PatchTST	1.00	0.99	0.96	0.99	0.93	0.96	0.99
Mamba	1.00	0.99	0.99	0.99	0.98	0.98	0.98
OPT1 → REF							
TCN	−0.34	0.97	0.99	0.99	0.98	0.99	0.99
Prob-TCN	−0.29	0.89	0.95	0.95	0.95	0.94	0.91
FITS	0.02	−0.35	0.06	0.12	0.30	0.33	0.00
LSTM	0.78	0.98	0.98	0.99	1.00	1.00	1.00
S4D	0.69	0.93	0.97	0.99	1.00	0.99	0.99
Chronos	0.83	–	0.97	–	–	–	–
TimesNet	0.81	0.91	0.93	0.94	0.94	0.98	0.89
Hybrid	0.03	0.47	0.79	0.92	0.99	0.98	0.96
PatchTST	0.86	1.00	0.97	0.99	0.98	1.00	0.96
Mamba	0.79	0.98	0.99	1.00	1.00	1.00	1.00
OPT2 → REF							
TCN	−0.14	0.98	0.99	0.99	1.00	0.98	0.99
Prob-TCN	0.22	0.83	0.93	0.92	0.94	0.94	0.91
FITS	0.15	−0.17	0.17	0.22	0.38	0.38	0.11
LSTM	0.68	0.98	0.99	1.00	0.97	0.99	0.99
S4D	0.79	0.96	0.99	1.00	1.00	0.99	1.00
Chronos	0.80	–	0.97	–	–	–	–
TimesNet	0.91	0.84	0.96	0.98	0.94	0.96	0.98
Hybrid	−0.30	−0.03	0.85	0.78	0.95	0.99	0.89
PatchTST	0.90	1.00	1.00	0.98	0.97	1.00	1.00
Mamba	0.84	1.00	1.00	1.00	1.00	1.00	1.00
OPT2 → OPT1							
TCN	0.97	0.94	0.95	0.97	0.96	0.99	0.97
Prob-TCN	0.95	0.95	0.95	0.97	0.96	0.92	0.98
FITS	0.74	0.72	0.76	0.72	0.75	0.67	0.64
LSTM	0.98	0.97	0.97	0.98	0.92	0.98	0.98
S4D	0.96	0.95	0.98	0.98	0.98	0.98	0.95
Chronos	0.56	0.58	0.76	0.84	0.70	0.78	0.68
TimesNet	0.73	0.21	0.87	0.87	0.94	0.96	0.83
Hybrid	0.91	–	0.97	0.97	0.95	0.97	0.97
PatchTST	0.98	0.95	0.87	0.95	0.97	0.97	0.98
Mamba	0.99	0.97	0.99	0.99	0.99	0.98	1.00

Table 39: **Adaptation curve, top.** As Table 38, at the top of the tower; <-1: below -1.

	Target simulations N						
Model	0 (zero-shot)	5	10	25	50	100	18 (designed)
Adaptation steps	0	50	100	300	600	1,250	200
REF → OPT1							
TCN	0.81	0.11	0.01	0.31	0.19	0.43	0.90
Prob-TCN	0.89	0.84	0.87	0.91	0.90	0.73	0.34
FITS	0.65	0.54	0.74	0.58	0.67	0.61	0.64
LSTM	0.11	0.06	−0.43	0.41	0.55	0.33	0.43
S4D	0.14	−0.50	−0.24	0.36	0.25	0.45	−0.16
Chronos	−0.34	–	0.13	–	–	–	–
TimesNet	−0.77	<-1	<-1	<-1	<-1	<-1	<-1
Hybrid	<-1	–	<-1	<-1	<-1	<-1	<-1
PatchTST	<-1	<-1	<-1	−0.11	<-1	<-1	−0.89
Mamba	−0.12	−0.69	−0.62	−0.50	<-1	−0.59	−0.94
REF → OPT2							
TCN	0.76	0.52	0.08	0.35	0.54	0.66	0.89
Prob-TCN	0.88	0.89	0.83	0.86	0.75	0.67	0.50
FITS	0.61	0.53	0.70	0.56	0.63	0.57	0.61
LSTM	−0.05	−0.16	−0.25	0.53	0.68	0.54	0.39
S4D	0.04	−0.35	−0.19	0.43	0.29	0.48	0.04
Chronos	−0.33	−0.23	0.13	0.28	−0.03	0.15	−0.03
TimesNet	−0.92	<-1	<-1	<-1	<-1	<-1	<-1
Hybrid	<-1	–	<-1	<-1	<-1	<-1	<-1
PatchTST	<-1	−0.85	<-1	−0.11	<-1	−0.97	−0.55
Mamba	−0.19	−0.57	−0.54	−0.40	<-1	−0.26	−0.70
OPT1 → OPT2							
TCN	0.87	0.79	0.45	0.74	0.65	0.69	0.74
Prob-TCN	0.80	0.88	0.78	0.93	0.85	0.71	0.78
FITS	0.66	0.56	0.71	0.61	0.70	0.62	0.67
LSTM	0.72	0.62	0.59	0.63	0.67	0.40	0.54
S4D	0.60	0.44	0.60	0.62	0.35	0.53	0.43
Chronos	0.27	0.09	0.31	0.45	0.23	0.38	0.26
TimesNet	<-1	<-1	<-1	<-1	<-1	<-1	<-1
Hybrid	<-1	–	<-1	<-1	<-1	<-1	<-1
PatchTST	<-1	−0.63	<-1	−0.11	−0.70	<-1	−0.34
Mamba	−0.06	−0.03	−0.35	−0.07	−0.08	−0.18	−0.14
OPT1 → REF							
TCN	<-1	0.18	0.34	0.51	0.05	0.65	0.38
Prob-TCN	<-1	0.15	0.47	0.68	0.77	0.80	0.60
FITS	0.36	−0.01	0.36	0.32	0.52	0.45	0.32
LSTM	−0.58	−0.31	−0.54	−0.24	0.18	0.28	−0.06
S4D	<-1	0.59	0.25	0.46	0.18	0.15	0.34
Chronos	0.89	–	0.92	–	–	–	–
TimesNet	−0.12	<-1	<-1	<-1	0.03	<-1	<-1
Hybrid	0.58	0.48	0.60	0.91	0.75	0.47	0.89
PatchTST	0.62	−0.22	<-1	0.45	−0.07	<-1	−0.13
Mamba	0.01	−0.17	−0.01	0.14	−0.42	0.02	−0.41
OPT2 → REF							
TCN	<-1	0.14	0.02	0.45	0.59	0.47	0.59
Prob-TCN	<-1	−0.08	0.67	0.74	0.86	0.73	0.47
FITS	0.38	0.04	0.38	0.33	0.52	0.44	0.35
LSTM	−0.91	−0.24	−0.10	0.47	0.65	0.30	−0.01
S4D	−0.58	0.69	0.43	0.54	0.08	0.24	0.31
Chronos	0.85	–	0.92	–	–	–	–
TimesNet	0.20	<-1	−0.98	−0.29	<-1	<-1	<-1
Hybrid	0.52	0.22	0.67	0.90	0.84	0.58	0.92
PatchTST	0.45	−0.28	<-1	0.32	0.38	<-1	0.68
Mamba	0.21	−0.00	−0.11	−0.03	−0.29	−0.46	−0.69
OPT2 → OPT1							
TCN	0.56	0.57	0.41	0.60	0.52	0.73	0.77
Prob-TCN	0.88	0.71	0.82	0.88	0.90	0.79	0.87
FITS	0.67	0.59	0.74	0.61	0.71	0.63	0.68
LSTM	0.28	0.48	0.41	0.63	0.44	0.54	0.55
S4D	0.30	0.34	0.53	0.67	0.28	0.59	0.38
Chronos	0.22	0.16	0.39	0.46	0.28	0.44	0.28
TimesNet	<-1	<-1	<-1	<-1	<-1	<-1	<-1
Hybrid	<-1	–	<-1	<-1	<-1	<-1	<-1
PatchTST	<-1	−0.42	<-1	−0.18	−0.52	<-1	−0.77
Mamba	−0.39	−0.42	0.25	−0.14	−0.23	−0.61	−0.36

M WHAT DOES EACH SENSOR ADD? (SENSOR ABLATION)

Table 19 (Appendix H.4) scores the ten selected models under the five input sets of Section 4.2 at 50 epochs. Chronos and the hybrid read the fore-aft acceleration alone, so the other sets are not defined for them.

M.1 THE ACCELEROMETER CARRIES THE REALIZATION; SCADA CARRIES THE LEVEL

With SCADA alone Prob-TCN keeps 0.94 at the base and 0.86 at the top, because its sampled amplitude restores the level without the realization (Appendix I.2). Among the mean-squared-error models TCN reaches 0.88 at the base, and at the top only FITS passes 0.28 (0.61). The accelerometer alone is within 0.07 of the task inputs at the base for TCN, LSTM, S4D and TimesNet: SCADA adds the operating point, which the acceleration already implies (Table 19).

M.2 THE SECOND AXIS HELPS THE TOP; PLATFORM MOTIONS ADD LITTLE BEYOND IT

The side-side acceleration is neutral at the base (± 0.02 for TCN, PatchTST and S4D) and, at 50 epochs, lifts the top for every model that reads it except FITS: TCN 0.76 to 0.94, LSTM 0.44 to 0.75, S4D 0.44 to 0.59. Along the tower the two-axis curve separates from the task curve above $z/H \approx 0.7$ (Figure 19). The six platform motions give the TCN the same top score as the second axis (0.94 against 0.94) and less for LSTM, S4D and TimesNet; after 300 epochs the TCN reaches 0.98 with them against 0.97 with both axes. Up to $z/H \approx 0.7$ the three sets with the accelerometer

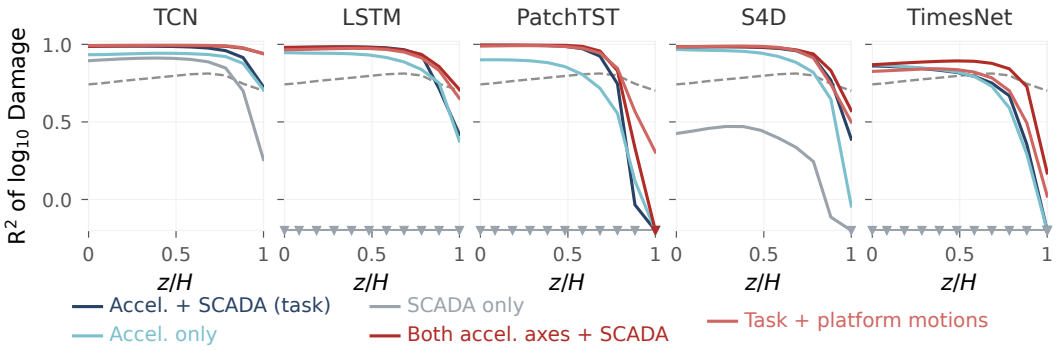


Figure 19: **Sensor ablation along the tower.** R^2 of $\log_{10}$ damage against height z/H under the five input sets of Table 19, one panel per model (TCN, LSTM, PatchTST, S4D, TimesNet), seed 0, 50 epochs, mean over the three towers; dashed gray: floor; triangles: values below the axis.

and SCADA are indistinguishable in Figure 19; SCADA alone stays about 0.1 below them for TCN and far below for the other models. Above it the second axis (red) and the platform set lift the curve for every model shown.

M.3 LONGER TRAINING KEEPS THE ORDER OF THE INPUT SETS

Table 40 repeats the input sets with the TCN at 50, 150 and 300 epochs, nine runs per cell; the five sets keep the same order at every budget. The poorer the input, the more the budget buys: between 50 and 300 epochs SCADA alone gains 0.46 at the top, the accelerometer alone 0.16, and both axes and the platform set 0.04. The top gap of the second axis shrinks from 0.17 to 0.05 and 0.04 and stays positive in all twenty-seven paired comparisons (0.025 to 0.052 at 300 epochs); on REF it is comparable to the seed spread.

M.4 GENERATOR POWER ADDS NOTHING, AND FIELD SCADA KEEPS THE CONCLUSIONS

Adding the generator power as a fourth SCADA series leaves the TCN base unchanged (0.990 against 0.991) and moves the top within the seed spread (0.78 against 0.76). Figure 20 shows why: power is torque times rotor speed, and the controller sets the generator torque from the rotor speed,

Table 40: **Training budget, input sets.** R^2 of $\log_{10}$ damage of the TCN under the five input sets of Section 5.4 at 50, 150 and 300 epochs, mean over the three towers of the three-seed median, the rule of Section 5 and Figure 3e; nine runs in every cell.

	base			z/H 0.78			top		
Input set	50	150	300	50	150	300	50	150	300
SCADA only	0.89	0.93	0.93	0.87	0.91	0.92	0.28	0.62	0.74
accelerometer only	0.96	0.98	0.98	0.92	0.97	0.97	0.68	0.78	0.84
task (one axis + SCADA)	0.99	0.99	0.99	0.98	0.99	0.99	0.76	0.91	0.93
both axes + SCADA	0.99	1.00	1.00	0.99	0.99	0.99	0.93	0.96	0.97
+ platform motions	0.99	0.99	0.99	0.99	0.99	1.00	0.94	0.97	0.98

so power is a function of an input the model already reads; over the 19,404 simulations their within-simulation correlation has median 0.995 and minimum 0.952. It is released and not used. The field configuration replaces the three 10 Hz SCADA channels by their window mean and standard deviation (Table 41). The convolutional models lose at most 0.03 of R^2 at the base and stay within their seed spread at the top; S4D loses its top score (0.39 to 0.04, seed 0). The conclusions of Section 5 hold under the field configuration.

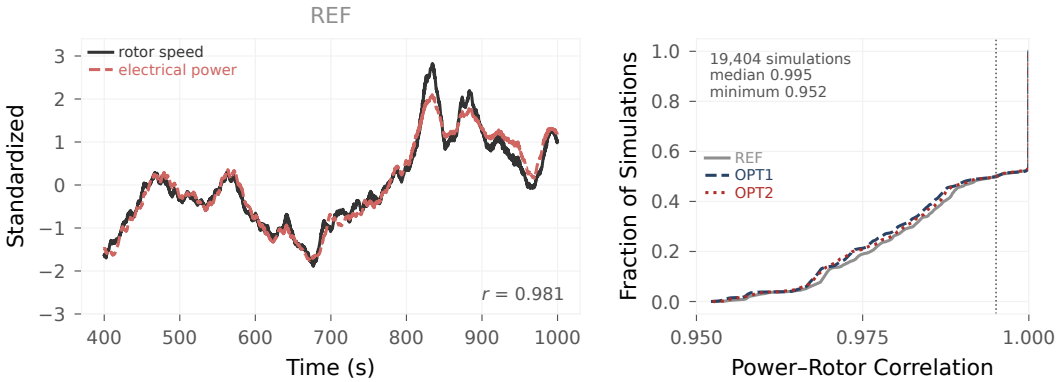


Figure 20: **The generator power is the rotor speed again.** Left: one simulation of REF over the scored window (400 to 1,000 s), rotor speed and electrical power standardized on one axis (the other towers overlap it). Right: cumulative distribution of their within-simulation correlation over all 19,404 simulations, per tower (dotted: median 0.995); the three towers overlap.

Table 41: Input configuration, within tower, R^2 of $\log_{10}$ damage, mean over the three towers (and over three training seeds where marked [3]; otherwise seed 0). Task: accelerometer and three SCADA channels at 10 Hz. Field: accelerometer at 10 Hz, SCADA as window mean and standard deviation. Two axes adds the side-side accelerometer.

Model	Input	base	z/H 0.48	z/H 0.78	top
TCN	task[3]	0.991	0.991	0.973	0.764
	task + power[3]	0.990	0.990	0.977	0.779
	field[3]	0.966	0.964	0.928	0.690
	two axes[3]	0.991	0.993	0.987	0.935
	two axes + power[3]	0.993	0.993	0.988	0.929
	two axes, field[3]	0.980	0.980	0.975	0.940
Prob-TCN	task[3]	0.960	0.959	0.956	0.846
	field[3]	0.967	0.958	0.952	0.853
	two axes[3]	0.971	0.969	0.970	0.924
	two axes, field[3]	0.975	0.974	0.977	0.929
S4D	task[3]	0.983	0.984	0.916	0.443
	field	0.965	0.951	0.813	0.044
	two axes, field	0.975	0.966	0.889	0.460

Summary. The accelerometer carries the realization and SCADA the level. The second axis lifts the top for every model that reads it except FITS, the platform motions add little beyond it, longer training keeps the order of the input sets, and the generator power, a function of the rotor speed through the controller, adds nothing.

N DATASHEET FOR DATASETS

This appendix follows the Gebru et al. (2021) template. Headings preserved verbatim where possible; full Q&A reproduced below.

N.1 MOTIVATION

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. FLOATSense was created to serve as a public, benchmark-quality reference for load reconstruction on floating offshore wind turbine (FOWT) towers, with a shared dataset, protocol and fatigue-damage metric. The specific task in mind is the reconstruction of the fore-aft bending-moment history at 11 tower heights from a tower-top accelerometer and SCADA signals (Section 4.1). The specific gap is that prior load-reconstruction methods are evaluated at only one or a few instrumented sections, and prior wind-fatigue datasets label operating points rather than release sensor and load time series (Table 1), so no public resource tests reconstruction along the tower, across operating regimes, tower geometries and sensor configurations. The release supports both engineering re-use (virtual sensing for structural monitoring) and ML re-use (an inverse time-series task with simulator ground truth at every height).

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)? The dataset was created by the authors of this paper; their identities and affiliations are withheld for anonymous review.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number. Funding is disclosed in the final version (withheld for anonymous review).

Any other comments? None.

N.2 COMPOSITION

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. Each instance is one 1,000 s aero-hydro-servo-elastic OpenFAST simulation of one of three FOWT tower geometries at one operating point and one realization: 10,001 `float32` samples at 10 Hz of 37 channels (Table 5). The channels are the measurable signals, namely tower-top accelerations, raw and gravity-corrected ($\times$4), SCADA ($\times$4: rotor speed, blade pitch, generator power, hub wind speed), platform motions ($\times$6) and wave elevation ($\times$1), and the simulator reference fore-aft and side-side bending moments at 11 heights ($\times$22). A second, smaller type of instance is the parked run (rotor parked, waves only), used to calibrate the physics baseline (Appendix F). Each simulation also carries per-channel statistics over the scored window, the reference fore-aft damage at the 11 heights, the section geometry used for the stress, and a metadata row with the operating point, the realization, the split, the regime labels and the lifetime weight `damage_weight` (the number of ten-minute periods its operating point is assigned over a 25-year life; Appendix E.3).

How many instances are there in total (of each type, if appropriate)? 19,404 simulations (3 towers $\times$ $1,078$ operating points $\times$ 6 realizations), for 717,948 channel time series and about 24.0 GB, plus 66 parked runs (22 per tower).

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? The released grid is a deliberate sample of a larger continuous operating envelope: 22 wind speeds and, at each, a 7×7 grid of significant wave height H_s and peak period T_p, for 1,078 operating points per tower, with six realizations per point that differ in the turbulence seed. It is the complete FLOATBench campaign, which samples the wind and wave distribution of one site under IEC 61400-3-2 DLC 1.2 (normal power production) with aligned wind and waves; the full environmental specification is that of FLOATBench (Section 3). The dataset is closed under DLC 1.2 and aligned-only wind–wave assumptions; misaligned conditions, parked or idling DLCs (beyond the 66 calibration runs) and fault DLCs are out of scope.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Raw simulator time series, unfiltered, except the two gravity-corrected tower-top accelerations, which are released next to their raw counterparts (Appendix F). The derived damage labels and per-channel statistics are released as separate files (Appendix E.3).

Is there a label or target associated with each instance? If so, please provide a description. Yes. The target is the fore-aft bending-moment history at the 11 heights over the scored window (400 to 1,000 s), and `damage.parquet` gives its reference fatigue damage at each height, computed as in the evaluation (Appendix G.3). The side-side moments are released but not scored.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing. No.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. Yes. `sim_id` is the FLOATBench simulation identifier: the same operating point and realization carries the same `sim_id` on the three towers and in FLOATBench. The grid indices (`wind_speed_id`, `wave_hs_id`, `wave_tp_id`) tie together the six realizations of the same operating point, and `wind_seed_id` identifies which realization a run is.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them. Yes: the regime-aware partition of FLOATBench, identical on the three towers, with 1,728 training and 4,740 test simulations per tower and all six realizations of an operating point on the same side (Section 4.2, Appendix E.2). Each test simulation carries a wind and a wave regime label (in-train, interpolation or extrapolation), giving nine regime cells (Table 4). The rationale is that a random split places the test set in the interior of the training envelope and hides the extrapolation behavior that matters in deployment. For model selection, 20% of the training operating points are held out as a validation split (Appendix G.5).

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. The signals are simulator output and carry no sensor noise. Aleatoric variability from the turbulence realization is the dominant source of spread and is represented by the six realizations per operating point. The first 400 s of each run are start-up transient and are excluded from scoring.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, are there guarantees that they will exist and remain constant over time; are there official archival versions; are there any restrictions (e.g., licenses, fees) that apply to a future user? Self-contained as released: the operating point, the physical wind and wave values, the split, the regime labels and the lifetime weight are in `metadata.parquet`, and `sim_id` links each run to its FLOATBench row for users who want the tabular labels. The upstream OpenFAST inputs (turbine, controller, platform) derive from the IEA-22-280-RWT release (Zahle et al., 2024), which is hosted on a public GitHub repository under Apache-2.0 (Appendix B); the redesigned towers are released with Ribeiro et al. (2026).

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description. No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. No: the dataset consists of numerical outputs of physics simulations and contains no human-generated content.

Does the dataset identify any subpopulations (e.g., by age, gender)? N/A: instances are physics simulations, not human data.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. No: the dataset contains no information about natural persons.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union

memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. No.

Any other comments? None.

N.3 Collection process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. Indirectly derived from physics simulation: each instance is the output of an aero-hydro-servo-elastic OpenFAST run of the IEA 22 MW turbine on a three-column semi-submersible under DLC 1.2 with aligned wind and waves (Section 3). The damage labels are derived from the moment histories by a 3 Hz low-pass filter, rainflow counting and Miner summation against the DNV-RP-C203 curve E S–N curve in air (bilinear, $m = 3/5$, with thickness correction; Appendix G.3). Validation: OpenFAST is an open-source NREL solver with code-to-code and code-to-experiment validation in the open literature, and the lifetime damage sums recomputed from the released series agree with the FLOATBench labels to within 7.1% at every scored height (Appendix I.3).

What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated? OpenFAST v3.5 (Jonkman, 2013) with turbulent wind fields from TurbSim (Jonkman & Buhl, 2006), executed in batch on a commercial cloud HPC platform; the OpenFAST setup follows the IEA-22-280-RWT release (Zahle et al., 2024). The time series were packed to Parquet and the damage labels computed with Python scripts released with the benchmark code; rainflow counting uses the `rainflow` package and the S–N curve follows DNV-RP-C203 (DNV, 2024). The same damage pipeline scores every model and the physics baseline.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Deterministic cartesian grid: 22 wind speeds $\times$ 7 significant wave heights H_s $\times$ 7 peak periods T_p, giving 1,078 operating points per tower, and six realizations per point with different turbulence seeds; the parked runs are one per wind speed at its median sea state.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? The authors only; no crowdworkers or paid annotators. Compensated through their academic positions.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created. The time series come from the FLOATBench simulation campaign, which required about 19,400 CPU-hours across the three towers (about 2.2 core-years); packing, labeling and benchmark preparation followed. Because the data is generated by simulation, the collection timeframe equals the creation timeframe.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. N/A: no human subjects, no animals, no PII.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? N/A.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself. N/A: no individuals.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. N/A: no human-subjects data.

Any other comments? None.

N.4 Preprocessing / cleaning / labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section. Yes: 37 of the 88 recorded channels are written to Parquet, one row group per simulation; the two tower-top accelerations are also released gravity-corrected (Appendix F); per-channel statistics are computed over the 400–1,000 s scored window; and the reference fore-aft damage at the 11 heights is computed by a 3 Hz low-pass filter, rainflow counting and Miner summation against the DNV-RP-C203 curve E S–N curve (Appendix G.3). No instance was removed.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. The full OpenFAST output (all 88 channels, about 190 GB) is retained by the authors and not released for storage reasons; the released channels are unfiltered simulator output.

Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point. Yes, released with the benchmark code under the MIT license; the code repository is referenced in the URL appendix (Appendix B).

Any other comments? None.

N.5 Uses

Has the dataset been used for any tasks already? If so, please provide a description. This paper, which evaluates a physics baseline, a one-gain floor and twenty learned models on the three towers under the within-tower, cross-tower and sensor-ablation protocols; see Section 5.

What (other) tasks could the dataset be used for? (i) virtual sensing and load reconstruction benchmarking; (ii) fatigue estimation from measurable signals; (iii) distribution-shift and transfer studies across regimes and tower geometries; (iv) sensor-placement studies, using the released accelerations, SCADA, platform motions and wave elevation; (v) probabilistic and multi-output models, neural operators and multi-fidelity learning; (vi) joint tabular and time-series studies with FLOATBench through `sim_id`.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. Will be maintained on the dataset repository page (anonymized link in Appendix B).

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups, or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a future user could do to mitigate these risks or harms? Yes. The dataset is restricted to (i) simulation only, with no sensor noise and no measured load series; (ii) one 22 MW turbine on one platform; (iii) DLC 1.2 (normal power production); (iv) aligned wind and waves only, which leaves the side-side moment without direct wave excitation; (v) mid-fidelity

OpenFAST modeling of the tower as a beam; and (vi) the DNV-RP-C203 curve E S–N curve (bilinear, $m = 3/5$) for the damage labels (Section 6). Reconstruction models trained on FLOATSense should not be expected to generalize without retraining or validation to measured signals, other turbines or platforms, misaligned conditions, other DLCs or other S–N classes. Dataset consumers planning to use FLOATSense for certification-adjacent decisions must couple it with the full DLC matrix specified in IEC 61400-3-2 (IEC TS 61400-3-2:2019, 2019), now its 2025 edition, and validate the reconstruction on the target structure.

Are there tasks for which the dataset should not be used? If so, please provide a description. Yes: the dataset is *not certification-grade* and should not drive safety decisions on a real structure without validation on that structure (Section 6).

Any other comments? None.

N.6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description. Yes, publicly via a public dataset repository.

How will the dataset be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? Parquet files (one directory per tower, the time series in 17 shards, plus one shared parked-run file; Table 6) with a Croissant 1.1 metadata record, a fixed subset hosted at an anonymized link during review (Appendix B) and the full dataset on the public dataset repository with the final version. A DOI will be released with the final version.

When will the dataset be distributed? A fixed subset at the time of submission through the anonymized link, the full release to reviewers on request during review, and the full dataset publicly with the final version.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. CC-BY-4.0 for the dataset (`https://creativecommons.org/licenses/by/4.0/`); MIT for the accompanying code (Appendix C). No fees apply.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. The IEA-22-280-RWT reference turbine and floating support structure (Zahle et al., 2024) on which the simulation campaign is built, and the OpenFAST (Jonkman, 2013) and TurbSim (Jonkman & Buhl, 2006) software used to run it, are released under Apache-2.0; FLOATSense is a downstream derivative and inherits no usage restriction beyond Apache-2.0 attribution. No other third party imposes restrictions on the released files.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. None.

Any other comments? None.

N.7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The authors (disclosed in the final version).

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? Through the corresponding author (withheld for anonymous review), or by opening an issue on the dataset repository page.

Is there an erratum? If so, please provide a link or other access point. No erratum at submission time. Errata, if any, will be tracked through the version history of the dataset repository and a corresponding `ERRATA.md` file in the dataset repository.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)? Yes. Updates will be issued by the authors as needed (no fixed cadence) as versioned releases, communicated through the dataset repository page and its version history. The authors commit to a v1.x release accompanying the paper and to logging changes in the repository. Planned additions include higher-fidelity (FEA, CFD) re-runs of selected conditions (Section 6), and, on the public leaderboard, a validation-tuned track (per-model learning rate, validation checkpointing, three seeds), a protocol supervised at the base gauge only, and modal and Kalman (joint input-state) baselines.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced. N/A: no human-subjects data.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users. Yes. All released versions remain hosted on the dataset repository with version-pinned access, so the version used in this paper stays available and downstream work that pins a specific release remains reproducible after later releases.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Yes: the released packing scripts accept further towers and load cases, and contributions produced with them can be released as additional versions under the same CC-BY-4.0 license with attribution, not as overrides of an existing version. The authors will check contributed files against the released schema and will flag (rather than reject) contributions whose simulation provenance is not reproducible from the linked code.

Any other comments? None.

O AI USE STATEMENT

In this work, we used generative AI tools (Claude and ChatGPT) to implement methods, namely the data-processing, training, evaluation and scoring code, to provide feedback on the research methodology and on the design of experiments, such as the sensor ablation, the few-shot protocol and the training-budget controls, to support data analysis and the interpretation of results, to refine the framing of the benchmark, and to clean and reformat the dataset for release. We have not used generative AI tools to generate data. Every time series comes from an OpenFAST simulation run by the authors, and the tools only helped write the post-processing scripts. Formulating mathematical claims and writing proofs are not applicable to this work. Additionally, we used generative AI tools to create and modify figures, to create and edit code, to draft and edit text for readability, to suggest the paper structure, title and keywords, and to identify, summarize and check related work and format references. We have reviewed all AI-assisted work. Every number in the paper was re-derived from the released per-simulation damage files by the released scorer and scripts, reference metadata were checked against publisher records, and research decisions, protocols and claims were made by the authors. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.