SOL ENGINEERING SERVICES

Uncertainty-Aware Neuro-Symbolic AI for Engineering Decisions

Hybrid physics and learned models, epistemic and aleatoric uncertainty, detection of operation outside the training envelope, regime switching, engineer-in-the-loop governance and alignment with ARL-BAA-0038

Prepared for DEVCOM Army Research Laboratory, the Ground Vehicle Systems Center, and university and industry partners of the CF-JET effort

Classification	Unclassified. Prepared from open-source information.
Date	23 November 2026

Executive Summary

A hybrid model that adds a learned discrepancy term to a codified physics rule predicted thermal resistance outside its training region with 3.1 percent mean absolute error in a benchmark computed for this paper, against 7.0 percent for a purely learned Gaussian process, 15.1 percent for an ensemble of five networks and 15.7 percent for the rule alone (60 training points, Table 7). The result supports the architecture described for CF-JET on a synthetic problem.[12] Three conditions limit it. The published hardware evidence for neuro-symbolic cooling design is thin (Table 1). The hybrid 95 percent intervals covered 70.9 percent of far-region observations. The benchmark rule error is smooth, which favors a hybrid model.

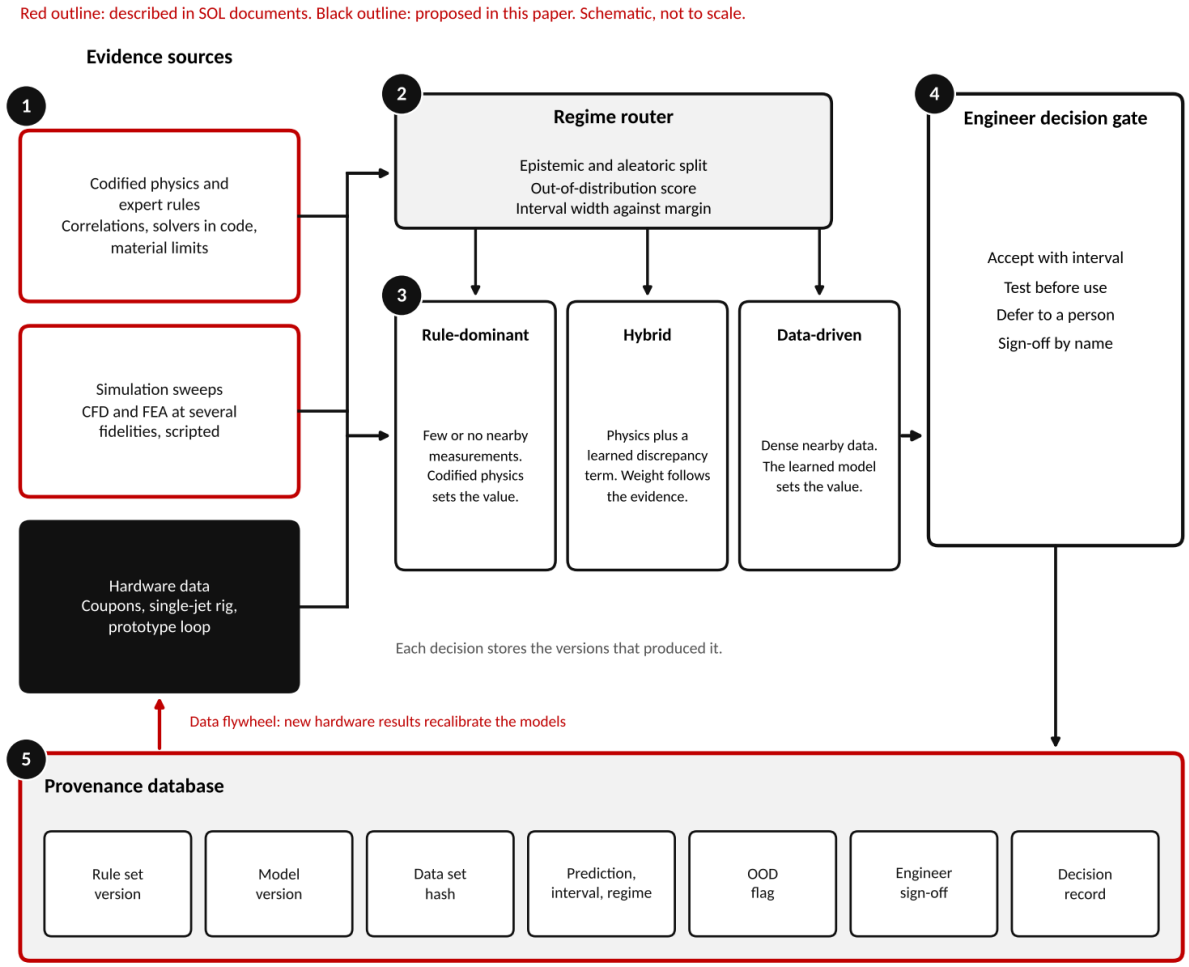
- **Established methods.** Physics-informed machine learning is a reviewed field, and physics-informed networks have been applied to heat transfer inverse problems with sparse measurements.[34] A probabilistic digital twin framework was calibrated with data from a physical aircraft.[5] No retrieved study validates a neuro-symbolic model on a jet-impingement cooler (Table 1).
- **Separating uncertainty.** A Gaussian process with a noise term recovered the injected 2.0 percent noise to within 0.1 percentage point at 50 to 400 training points (computed, Table 3, Figure 2). Ensemble disagreement rose from 4.73 to 9.48 percent in the far region as training points grew from 25 to 400, so its spread did not track what more data would resolve.
- **Calibration.** At 60 training points the 95 percent intervals covered 93.1 percent of in-region observations for the Gaussian process, 85.6 percent for the hybrid and 56.1 percent for the ensemble, and coverage fell in every method outside the region (Table 7). Published studies report the same loss of calibration under dataset shift.[67]
- **Detecting the envelope.** Mahalanobis distance on the inputs separated out-of-region points with an area under the curve of 0.947 near the region and 0.999 far from it, against 0.829 and 0.947 for ensemble disagreement (Table 6). At a fixed 5 percent false-alarm rate its detection rates were 73.4 percent near and 99.4 percent far (computed). The benchmark region is defined in input space, which favors input-space detectors.
- **Regime switching.** A weight computed from the epistemic variance of the hybrid discrepancy term routed 31.7 percent of in-region points to the data-driven regime and 78.6 percent of far-region points to the rule-dominant regime (Table 8). An interval-width guardrail cut far-region errors over 10 percent for the Gaussian process from 22.9 to 6.0 percent of points and accepted 51.5 percent of them (Figure 4).
- **Army publications.** ARL-BAA-0038 states the sparse-data problem and names domain rules and uncertainty-aware neuro-symbolic processing as expected elements.[8] ARL-authored work addresses sparse-exemplar adaptation and physics-guided networks. No retrieved ARL publication addresses regime detection for thermal hardware (Table 9, Table 10).
- **Governance.** DoD, NIST and W3C documents supply principles, functions and a provenance data model, and the ASME credibility standard targets medical devices.[9][10][11][12] No retrieved source prescribes engineer sign-off for a learned surrogate of thermal hardware. A decision-record chain assembled from the cited guidance is proposed in this paper (Table 11, Figure 5).

1. Introduction and Scope

The CF-JET proposal replaces a conventional jet-impingement cold plate with a three-dimensional printed body that has interleaved return ports and a carbon fiber target plate with perpendicular pitch fiber

inserts, sealed by a thin electroplated coolant face. The design point is 1.5 kW on a 60 mm by 60 mm footprint with 50/50 water-glycol entering at 65 °C at 4 to 8 L/min, a pressure drop of 30 kPa or less and a case-to-coolant resistance of 0.02 K/W.[13] Over 36 cm² the heat flux averages 41.7 W/cm², and the resistance budget equals a 30 K rise from case to coolant at full load (derived). The engineering decisions that follow from this concept rest on models of several kinds: published correlations, simulation at several fidelities, and measurements from coupons, a single-jet rig and a prototype loop. Each kind has a region of validity and an error that varies across the design space.

Figure 1. Operational concept of uncertainty-aware neuro-symbolic inference for CF-JET design decisions



Red outlines mark elements that SOL documents describe: codified physics and rules, simulation sweeps, the provenance database and the data flywheel.[14][15] *Black outlines mark elements proposed in this paper. The diagram is schematic and not to scale.*

SOL Engineering Services describes a software pipeline, SOL Aquifer, that codifies engineering steps as code: scripted condition sweeps, physics solvers in code, a SQL database that records provenance, and surrogate training with uncertainty quantification.[16] The same documents describe uncertainty-aware neuro-symbolic inference, which combines codified physics and rules with learned surrogates, and a data flywheel that recalibrates the models from coupon, single-jet and prototype test data.[17] The approach originated in building HVAC performance modeling done with Alabama A&M University students under the name Code as Architecture.[18] These are capabilities that SOL proposes or has demonstrated. The HVAC

work, carried out with Alabama A&M University students through Georgia Tech funded internships, is documented in SOL's 2026 poster, Modeling HVAC System Performance: Code as Architecture Approach to AI, and the Forest-to-Desert Framework is defined in SOL's topic development report for ARL-BAA-0107. Hardware validation of the approach has not yet been published; the CF-JET program is meant to supply it. The Army research topic on extremely sparse data names the same pairing of domain rules and uncertainty-aware processing as a research need.[19]

Figure 1 shows the loop that this paper examines. Evidence from rules, simulation and hardware enters a router that estimates two kinds of uncertainty and a score for operation outside the training envelope. The router sends each query to a rule-dominant, hybrid or data-driven regime. An engineer decides whether to accept, test or defer, and the decision record stores the versions that produced it. The paper answers five questions. Which neuro-symbolic and physics-informed methods are established for engineering design, and which have been validated on hardware? How do Gaussian processes, ensembles and evidential models separate epistemic from aleatoric uncertainty, and how is calibration measured? Which methods detect that a model operates outside its training envelope? What does the Army Research Laboratory publish on sparse-data AI, regime detection and edge deployment? Which governance practices appear in Department of Defense responsible AI and digital engineering guidance?

The literature search covers the five questions. A benchmark computed for this paper adds measurements that the literature does not supply. It uses a synthetic jet-impingement resistance with five inputs (jet diameter, pitch, gap, flow rate and through-thickness conductivity), a rule model with a deliberate 16 to 17 percent error, and training data restricted to the central half of each input range. Test points fall in the training region, near it or far from it. The benchmark tests software behavior and the sample sizes that uncertainty methods need. It makes no statement about CF-JET hardware.

2. Established Methods and Hardware Validation

2.1 Physics-informed and neuro-symbolic methods are established in the literature, and hardware validation is reported for few of them

Karniadakis and colleagues review physics-informed machine learning and describe three ways to build physical knowledge into a learned model: through the data that the model sees, through the architecture, and through the loss function.[20] Cai and colleagues apply physics-informed neural networks to heat transfer problems, including inverse problems in which temperature and velocity measurements are sparse, and the abstract of their study refers to industrial applications related to power electronics.[21] Garcez and Lamb argue that combining neural learning with symbolic reasoning defines a third wave of AI, and their review gives the neurosymbolic field its general vocabulary.[22] These three sources establish the methods. None of them reports a jet-impingement cooler.

The oldest statistical form of the hybrid idea is Bayesian calibration of a computer model, in which a Gaussian process represents the discrepancy between the simulator and reality.[23] The benchmark in this paper follows that form: a Gaussian process learns the residual between a rule model and the observations. Duraisamy, Iaccarino and Xiao review the use of data in turbulence modeling, where a physical closure is corrected by learning from high-fidelity data.[24] Mandler and Weigand examine how far such learned closures generalize and report limits to that generalization.[25] The finding applies to CF-JET because the same limit applies to any learned correction trained on a bounded set of conditions.

Three coupling patterns appear in this literature, and each changes what the model can report about its

own uncertainty. In the first, physics enters through the loss function, as in physics-informed networks, and the residual of the governing equation is penalized during training.[2627] In the second, a rule model or simulator supplies a baseline and a learned term corrects its error, as in Bayesian calibration with a discrepancy term and in learned turbulence closures.[2829] In the third, symbolic rules gate or constrain the output of a learned model, which is the pattern that the neurosymbolic reviews emphasize.[30] The CF-JET documents name codified physics and rules together with learned surrogates and do not state which pattern applies.[3132] The benchmark in this paper tests the second pattern, because its discrepancy term has an explicit uncertainty. A network trained with a physics loss returns a single value unless it is wrapped in an ensemble or a Bayesian layer, and a rule gate returns a decision and no interval. Table 1 lists the methods and the validation that each source reports.

Table 1. Neuro-symbolic and physics-informed methods and their hardware-validation status

Method	Source	Role of physics or rules	Reported validation	Hardware validation status	Evidence
Physics-informed machine learning	Karniadakis et al. 2021[33]	Physics enters through data, architecture or loss function	Review of applications across fields	Not a hardware study	Peer-reviewed
Physics-informed neural networks for heat transfer	Cai et al. 2021[34]	Governing equations in the loss	Forward and inverse heat transfer problems with sparse temperature and velocity data	Abstract cites power electronics applications; measurements on hardware not confirmed in the record read	Peer-reviewed
Bayesian calibration with a discrepancy term	Kennedy and O'Hagan 2001[35]	Simulator is the rule model; a Gaussian process models its error	Statistical framework	Application-specific	Peer-reviewed
Learned corrections to turbulence closures	Duraisamy et al. 2019; Mandler and Weigand 2025[3637]	Physical closure corrected with high-fidelity data	Reviewed gains; reported limits on generalization	Not a hardware study	Peer-reviewed
Probabilistic graphical model of a digital twin	Kapteyn et al. 2021[38]	Physics-based component models in a graph	Calibrated with experimental data from a physical aircraft	Yes, for the aircraft; not a thermal system	Peer-reviewed
Physics-guided neural network	Army Research Laboratory report, tech brief ARL-9655[3940]	Physical constraints on derivatives enforced in custom loss functions	R^2 of 0.991 for the physics-guided network against 0.884 for a conventional network, melt pool width in directed energy deposition	Predicts a build process with sensing; not a thermal system	Government report abstract; trade summary
Neurosymbolic AI, general	Garcez and Lamb 2023[41]	Symbolic reasoning combined with neural learning	Research review	Not a hardware study	Peer-reviewed
Uncertainty-aware neuro-symbolic inference, CF-JET	SOL documents[4243]	Codified physics and rules with learned surrogates	Described; no results in the documents read	Proposed	Proposal material

The status column states what the cited record reports. The abstracts of Cai et al. and of the Army Research Laboratory report were read without the full reports.

2.2 No retrieved study validates a neuro-symbolic model on a jet-impingement cooler

Kapteyn, Pretorius and Willcox calibrate a probabilistic graphical model of a digital twin with experimental data from a physical unmanned aircraft, which is the clearest hardware case among the retrieved sources.[44] The approach transfers to a cooler only by analogy, because the retrieved record describes an aircraft and not a thermal system. The Army Research Laboratory physics-guided network for melt pool width demonstrates the constraint idea on additive manufacturing data and reports an R^2 of 0.991 against 0.884 for a conventional network, a result for a manufacturing process and not for a thermal system.[45] A search for hardware validation of a physics-informed or hybrid model on a jet-impingement cold plate returned no record. This is a gap in the retrieved literature, and it applies equally to purely learned surrogates of jet arrays.

The gap sets the evidence standard for CF-JET. A claim that hybrid inference works for a printed cold plate with a carbon fiber target plate rests on the benchmark in this paper and on measurements that SOL will take. The benchmark shows what a hybrid model can do when the rule error is smooth, and Table 7 gives the numbers. A rule that fails abruptly, such as a change of flow regime between jets, falls outside that test.

3. Separating Epistemic and Aleatoric Uncertainty

3.1 A Gaussian process separates the two types with an explicit noise term, and ensembles and evidential models infer them from spread or priors

Kendall and Gal define aleatoric uncertainty as noise inherent in the observations and epistemic uncertainty as the uncertainty in the model that more data explains away.[46] They also show that an aleatoric estimate alone cannot identify inputs that lie outside the data, which is the reason a design decision needs both. A Gaussian process separates them in one fitted object: the white-noise term estimates the aleatoric part and the posterior variance of the latent function estimates the epistemic part.[47] A deep ensemble trains several networks that each predict a mean and a variance, and takes the average of the variances as the aleatoric estimate and the disagreement of the means as the epistemic estimate.[48] Evidential regression places a normal-inverse-gamma prior over the parameters of the output distribution, so a single network predicts both parts in one forward pass, and its authors report that it needs neither sampling nor out-of-distribution training examples.[49] The classification form of the idea has a co-author at the Army Research Laboratory.[50] Conformal prediction produces an interval with a coverage guarantee and does not separate the two types.[51] Table 2 compares the methods.

Table 2. Uncertainty-separation methods compared

Method	Aleatoric estimate	Epistemic estimate	Reported or expected behavior outside the data	Cost	Source
Gaussian process	White-noise kernel term	Posterior variance of the latent function	Posterior variance grows with distance from the data	Cubic in training points for exact inference[52]	Textbook
Deep ensemble	Mean of member variances	Disagreement of member means	Performed strongly under dataset shift among the methods tested[53]	Several trainings; one forward pass per member[54]	Peer-reviewed

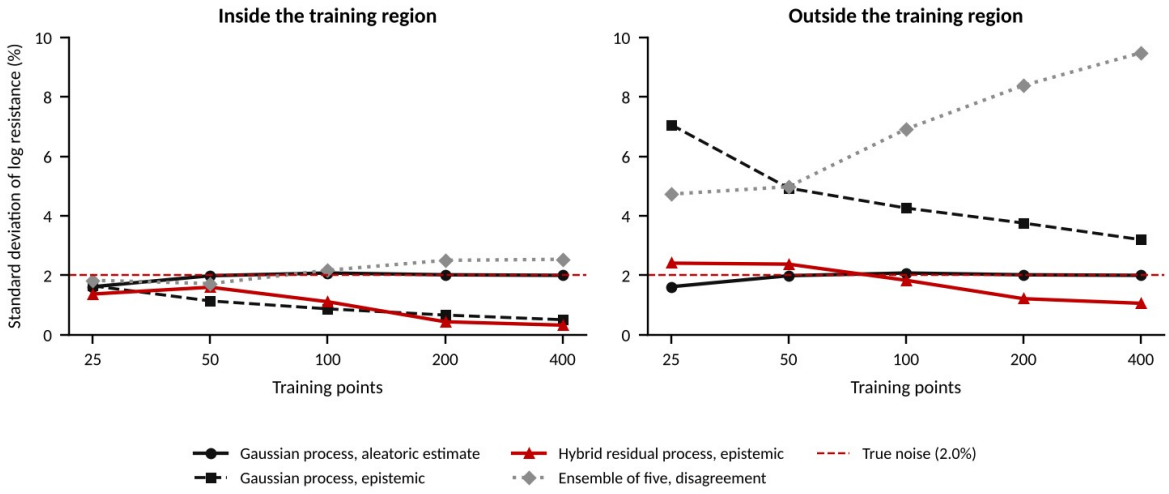
Method	Aleatoric estimate	Epistemic estimate	Reported or expected behavior outside the data	Cost	Source
Evidential regression	Expected noise from the prior	Variance of the predicted mean from the prior	Reported to need no sampling and no out-of-distribution training data[55]	One forward pass	Peer-reviewed
Hybrid residual process	Noise term of the residual process	Posterior variance of the discrepancy	Reverts to the rule model far from the data (computed here)	Cubic in training points	Computed in this paper
Conformal prediction	No separation	No separation	Coverage guarantee assumes exchangeable data[56]	Held-out calibration set	Review

The Gaussian process, ensemble and hybrid rows are computed in the benchmark of this paper. Evidential regression was not computed, and its behavior on thermal data is untested here.

3.2 In a computed benchmark the Gaussian process recovered the 2.0 percent noise from 50 points, and ensemble spread did not fall with data

The benchmark injected Gaussian noise with a standard deviation of 2.0 percent of the resistance and trained on the central half of each input range. Table 3 reports the estimates averaged over six training sets at each size, and Figure 2 plots them. The Gaussian process estimate of the noise was 1.99 percent at 50 points, 2.08 at 100, 2.02 at 200 and 2.00 at 400, and it was 1.61 percent at 25 points with a spread of 0.42 across training sets. The model therefore identified the aleatoric part once the sample was about ten times the number of inputs. The identification rests on the smoothness of the underlying function. A function with fine structure would absorb part of the noise into the signal.

Figure 2. Estimated aleatoric and epistemic standard deviation against training size, inside and outside the training region



Computed in this paper on a synthetic five-input problem with 2.0 percent injected noise; values are means of six training sets and are not CF-JET results.

Table 3. Computed uncertainty estimates against training size (percent standard deviation of log resistance)

Training points	Aleatoric estimate (spread)	GP epistemic, in region	GP epistemic, far	Hybrid epistemic, in region	Hybrid epistemic, far	Ensemble, in region	Ensemble, far

Computed in this paper. The injected noise is 2.0 percent. The spread is the standard deviation of the aleatoric estimate over six training sets. Ensemble columns report disagreement of five networks.

The epistemic estimates behaved differently across the three methods. The Gaussian process epistemic deviation inside the region fell from 1.64 to 0.51 percent as training points grew from 25 to 400, and in the far region it fell from 7.06 to 3.20 percent. The hybrid epistemic deviation was smaller in both places, 0.32 percent in the region and 1.06 percent far from it at 400 points, because the rule model already supplies most of the structure. The ensemble disagreement moved in the opposite direction. It rose from 1.83 to 2.54 percent in the region and from 4.73 to 9.48 percent in the far region. The benchmark does not explain the trend. The observation limits the use of raw ensemble spread as an estimate of what more data would resolve, and it agrees with the reported need to recalibrate ensembles.[57]

The separation has a practical limit for CF-JET data. Simulation output from a converged solver has near-zero aleatoric noise, and the scatter between nominally identical runs reflects mesh and convergence differences. Hardware data has measurement noise that includes sensor error, loop repeatability and assembly variation. A model trained on both kinds of data needs a noise term for each source, and replicate measurements at identical settings are the only direct way to estimate the hardware term without relying on smoothness.

CF-JET combines three kinds of evidence, and each needs its own treatment in the uncertainty model. Correlations contain a systematic error that does not shrink with more evaluations, and a discrepancy term absorbs it only where measurements exist. Simulation output varies with mesh, closure and convergence, and its error against hardware is a bias that a Gaussian process can model as a smooth function of the inputs. Hardware data contains both a bias relative to the physics and measurement noise. A single noise term fitted to the pooled data would average these sources and misstate each. The Bayesian calibration framework handles the problem by giving the simulator discrepancy and the observation noise separate terms, and the benchmark follows that structure.[58]

3.3 Calibration is measured against held-out observations, and interval coverage is the quantity that a design decision uses

Guo and colleagues define expected calibration error for classifiers by binning predictions by confidence and comparing confidence with accuracy, and they show that temperature scaling repairs the miscalibration of modern networks on in-distribution data.[59] Kuleshov, Fenner and Ermon extend the idea to regression by comparing the predicted cumulative distribution with the observed frequency at each quantile level.[60] Ovadia and colleagues test predictive uncertainty under dataset shift and report that post-hoc calibration fell short as the shift grew, while methods that average over several models, including ensembles, performed strongly.[61] A Google Research summary of the study states the same finding for practitioners.[62] Table 4 lists the metrics and the use of each in this paper.

Table 4. Calibration and selective-prediction metrics

Metric	Quantity measured	Applies to	Limit	Source
Expected calibration error	Gap between confidence and accuracy over bins	Classification	Defined for class probabilities; not used here	Guo et al. 2017[63]
Regression calibration error	Gap between nominal and observed frequency over quantile levels	Regression	Used here with ten levels from 0.1 to 0.95	Kuleshov et al. 2018[64]
Interval coverage	Share of observations inside the nominal 95% interval	Regression	Used here; scored against new noisy observations	Computed in this paper

Metric	Quantity measured	Applies to	Limit	Source
Conformal coverage	Finite-sample guarantee on interval coverage	Regression and classification	Guarantee assumes exchangeable data[65]	Angelopoulos and Bates 2023
Risk and coverage of selective prediction	Error among accepted predictions against share accepted	Classification in the source	Used here as share accepted and errors over 10% among accepted	Geifman and El-Yaniv 2017[66]

The regression calibration error in the benchmark is the mean absolute gap between nominal and observed coverage over ten central intervals.

The benchmark scores every interval against new noisy observations drawn at the test inputs, because a design decision meets a noisy measurement. Scoring against the noise-free truth would count the aleatoric part as over-coverage. Table 7 reports the results. A calibration error of 0.3 means that the observed coverage differs from the nominal coverage by 30 percentage points on average over the ten levels.

4. Detecting Operation Outside the Training Envelope

4.1 Detection methods fall into input-space, model-based and post-hoc families

Heim and Frank of the Software Engineering Institute describe out-of-distribution detection as a last line of defense and group the approaches into data-only methods, uncertainty-aware methods and post-hoc methods, and they note that edge devices limit what a detector can compute.[67] Lee and colleagues score inputs with the Mahalanobis distance to class-conditional Gaussian distributions fitted on the features of a trained classifier.[68] Liu and colleagues derive an energy score from the logits of a classifier and report that it separates in-distribution from out-of-distribution inputs better than the softmax confidence.[69] Geifman and El-Yaniv let a classifier abstain when its confidence falls under a threshold, which turns detection into a selective prediction with a measurable risk.[70] Regression has no logits and no classes, so the energy score and the class-conditional distance have no direct counterpart. Detectors for regression use the distance of the input to the training inputs or the epistemic variance of the model. Table 5 lists the families.

Table 5. Out-of-distribution detection methods

Method	Score	Needs	Applicability to regression	Source
Input-space distance	Mahalanobis or nearest-neighbor distance of the input to the training inputs	Training inputs only	Direct; no model required	Lee et al. 2018[71] for the class-conditional form
Model epistemic variance	Posterior deviation of a Gaussian process or hybrid discrepancy	A probabilistic model	Direct	Computed in this paper
Ensemble disagreement	Deviation of member predictions	Several trained models	Direct	Lakshminarayanan et al. 2017[72]
Energy score	Log-sum-exponential of classifier logits	A classifier	No direct form	Liu et al. 2020[73]
Selective prediction	Abstain when confidence falls under a threshold	Calibrated confidence	Interval width serves as the confidence	Geifman and El-Yaniv 2017[74]

The input-space, epistemic and ensemble rows were computed here. The energy score is listed for completeness and was not computed.

4.2 Input-space distance separated out-of-region points best in the benchmark, with an area under the curve of 0.947 near the region and 0.999 far from it

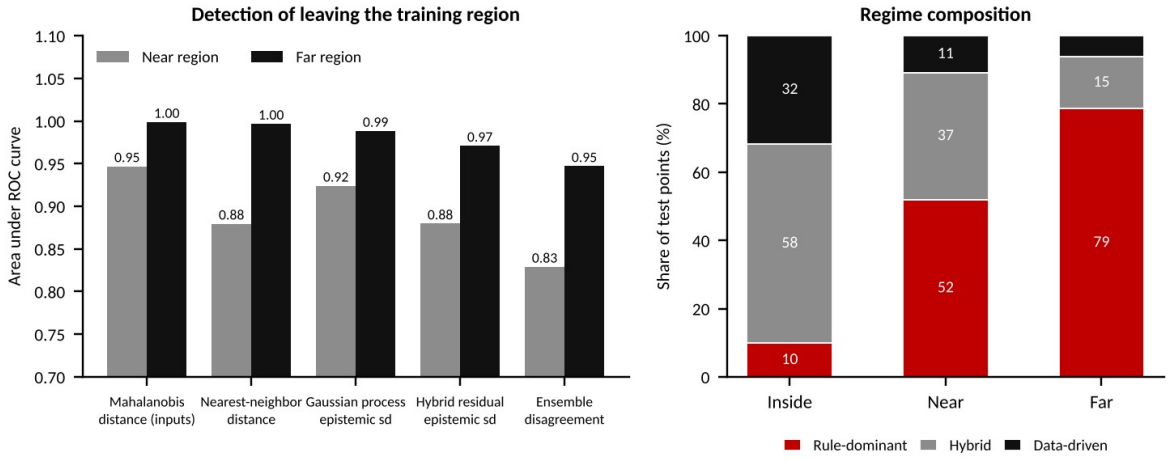
The benchmark scored five detectors by the area under the receiver operating characteristic curve for separating in-region test points from near points and from far points. Table 6 reports the results at three training sizes. At 60 training points the Mahalanobis distance reached 0.947 near and 0.999 far, the nearest-neighbor distance 0.880 and 0.997, the Gaussian process epistemic deviation 0.924 and 0.988, the hybrid epistemic deviation 0.880 and 0.971, and the ensemble disagreement 0.829 and 0.947. The ordering held at 30 and 120 points. Separation was lower in the near zone, which lies within one eighth of the input range from the region, than in the far zone for every detector, and the Mahalanobis distance led in both zones. Figure 3 shows the areas at 60 points.

Table 6. Computed area under the curve for detecting points outside the training region

Detector	Basis	Near, 30 points	Near, 60 points	Near, 120 points	Far, 30 points	Far, 60 points	Far, 120 points
Mahalanobis distance	Inputs	0.922	0.947	0.952	0.995	0.999	0.999
Nearest-neighbor distance	Inputs	0.860	0.880	0.897	0.993	0.997	0.999
Gaussian process epistemic deviation	Model output	0.905	0.924	0.934	0.985	0.988	0.991
Hybrid epistemic deviation	Model output	0.850	0.880	0.889	0.953	0.971	0.970
Ensemble disagreement	Model output	0.791	0.829	0.837	0.915	0.947	0.959

Computed in this paper; means of 20 training sets per size. The training region is the central half of each of five input ranges, and an area of 0.5 means no separation. The synthetic problem defines the region in input space, which favors input-space detectors.

Figure 3. Detection of operation outside the training region, and the regime composition that follows from the epistemic weight



Computed in this paper at 60 training points on a synthetic five-input problem. Bars show the area under the curve for each detector and the share of test points in each regime; the values are not CF-JET results.

The result needs a qualification. The benchmark defines the training region as a hyper-rectangle in input space, so a distance on the inputs measures exit from that region almost directly. The model-based scores respond to the same exit through a longer chain, since the posterior variance depends on fitted length scales. A real envelope for CF-JET follows manufacturing limits, regime boundaries of the flow and combinations of variables. In that setting a model-based score can detect a boundary that the input

distance misses, for example a combination of high flow and small jet diameter at which the flow regime changes. The benchmark supports the use of input distance as a low-cost first gate and does not show that it suffices.

4.3 A detection threshold needs a stated cost of a missed detection

A detector yields a score, and a decision needs a threshold. The threshold converts the score into a rate of missed detections and a rate of false alarms, and the two have different costs for CF-JET. The benchmark sets each threshold at the 95th percentile of the scores of in-region test points, which fixes the false-alarm rate at 5 percent, and counts the share of near and far points that score over it (computed). At 60 training points the Mahalanobis distance flagged 73.4 percent of near points and 99.4 percent of far points, the Gaussian process epistemic deviation 70.0 and 95.5 percent, the nearest-neighbor distance 60.4 and 98.4 percent, the hybrid epistemic deviation 64.1 and 91.8 percent, and the ensemble disagreement 41.5 and 76.6 percent. At 30 points the Gaussian process epistemic deviation flagged 61.5 percent of near points and the Mahalanobis distance 61.0 percent, so the model-based score matched the input distance where data were scarce. Even the best detector at 60 points left 26.6 percent of near points unflagged. A deployed model has no labeled in-region test set, so the in-region scores for the threshold would come from cross-validation on the training data. A missed detection lets a prediction of unknown error enter a design decision. A false alarm sends a sound prediction to an engineer or to a test, which costs time. Heim and Frank describe the detector as a last defense, so the threshold is a risk decision that belongs in the record of the decision and not in the code alone.[75] Table 8 and Figure 4 show how a threshold on interval width, set at 10 percent, behaves for each model.

5. Regime Switching and Guardrails

5.1 A hybrid model predicted far-region resistance with 3.1 percent mean error against 7.0 percent for a learned model and 15.7 percent for the rule

The benchmark compares four predictors at 30, 60 and 120 training points, and Table 7 gives the results at 60 points. The rule model has a fixed 16 to 17 percent error by construction, with an exponent and a constant that differ from the truth and no gap term. The Gaussian process learns the resistance directly. The hybrid learns the residual between the observations and the rule. The ensemble is five networks with two hidden layers of 32 units. In the region all three learned predictors reached mean errors under 3 percent, and the Gaussian process (0.9 percent) and the hybrid (1.0 percent) were equal for practical purposes. The difference appeared outside the region. The hybrid held 1.9 percent near the region and 3.1 percent far from it, while the Gaussian process fell to 2.6 and 7.0 percent and the ensemble to 6.7 and 15.1 percent.

Table 7. Computed accuracy, calibration and guardrail results at 60 training points (percent)

Zone	Model	Mean absolute error	Within ±5%	Within ±10%	95% interval coverage	Calibration error (fraction)	Accepted by guardrail	Errors over 10%, all points	Errors over 10%, accepted points
In region	Rule only	17.4	0.0	0.0	Not applicable	Not applicable	Not applicable	100.0	Not applicable
In region	Gaussian process	0.9	100.0	100.0	93.1	0.04	100.0	0.0	0.0
In region	Hybrid	1.0	99.8	100.0	85.6	0.08	100.0	0.0	0.0

Zone	Model	Mean absolute error	Within ±5%	Within ±10%	95% interval coverage	Calibration error (fraction)	Accepted by guardrail	Errors over 10%, all points	Errors over 10%, accepted points
In region	Ensemble of five	2.7	85.8	98.5	56.1	0.28	98.5	1.5	1.3
Near	Rule only	16.9	0.0	2.0	Not applicable	Not applicable	Not applicable	98.0	Not applicable
Near	Gaussian process	2.6	85.9	98.3	87.7	0.08	95.6	1.7	1.3
Near	Hybrid	1.9	95.1	100.0	83.9	0.11	100.0	0.0	0.0
Near	Ensemble of five	6.7	48.9	76.8	53.9	0.29	82.3	23.2	17.0
Far	Rule only	15.7	0.3	14.0	Not applicable	Not applicable	Not applicable	86.0	Not applicable
Far	Gaussian process	7.0	52.8	77.1	77.5	0.14	51.5	22.9	6.0
Far	Hybrid	3.1	79.5	99.4	70.9	0.18	100.0	0.6	0.6
Far	Ensemble of five	15.1	22.0	41.4	40.3	0.36	51.3	58.6	27.6

Computed in this paper. Averages over 20 training sets with 400 test points per zone. Coverage and calibration error are scored against new noisy observations, and accuracy against the noise-free truth. The guardrail accepts a prediction when 1.96 times the predictive deviation is 10% or less. Error shares are percentages of all test points in the zone.

The scope of the benchmark limits the reading of Table 7. The truth is one analytic resistance function with a convection term, a conduction term and a fixed contact term. The rule model has the right structure and wrong coefficients, which makes the discrepancy smooth and low in amplitude. Training points come from a Latin hypercube in the central half of each input range, and the test points are three fixed sets of 400. Accuracy and calibration use 20 training sets at each size, and the separation study uses six. The Gaussian process uses a Matérn kernel with a separate length scale for each input and a white-noise term, and the networks have two hidden layers of 32 units trained by a quasi-Newton method. Other kernels, network sizes and training procedures would change the numbers, and the ordering of the methods needs confirmation on hardware data before it guides a design.

The hybrid needed fewer points than the learned model for the same far-region accuracy. At 30 training points the hybrid reached 3.0 percent mean error in the far region, and the Gaussian process needed 120 points to reach 6.1 percent, which is still twice the hybrid result at 30 points. At 120 points the hybrid placed 90.6 percent of far-region predictions within ±5 percent, against 60.4 percent for the Gaussian process. This is the behavior that the Bayesian calibration literature describes: the discrepancy process returns to its prior mean far from the data, so the prediction reverts to the rule model and the error approaches the rule error.[76] The reversion is the benefit and also the limit, because a rule error that varies sharply outside the data is invisible to the learned term.

All three learned predictors under-covered. At a nominal 95 percent the hybrid covered 85.6 percent of in-region observations at 60 points and 70.9 percent of far-region observations, and the coverage improved to 93.2 and 82.2 percent at 120 points. The Gaussian process covered 93.1 and 77.5 percent at 60 points, and the ensemble 56.1 and 40.3 percent. Under a normal error model the hybrid far-region coverage at 60 points corresponds to an interval of about 1.06 predictive deviations, which means that the predictive deviation would need to be multiplied by about 1.9 to reach 95 percent (derived). The corresponding factor is about 1.6 for the Gaussian process. A hybrid model that reports accurate values with intervals that are too narrow misleads a decision more than a wide interval does, so interval recalibration with held-out data is a requirement.

5.2 A weight from the epistemic variance of the discrepancy routes each query to a rule-dominant, hybrid or data-driven regime

The hybrid model supplies a natural routing signal. The weight is one minus the ratio of the epistemic variance of the discrepancy process at the query to the fitted prior variance of that process. A weight near zero means that no nearby data constrain the discrepancy, and the prediction is the rule value. A weight near one means that the data determine the discrepancy, and the learned term dominates. The benchmark assigns a query to the rule-dominant regime when the weight is under 0.25, to the data-driven regime when it exceeds 0.75, and to the hybrid regime otherwise. Table 8 defines the regimes and gives their composition at 60 training points, and Figure 3 plots the composition.

Table 8. Regimes, release rules and computed composition at 60 training points

Regime	Weight on the learned term	Value released and flag	Release condition	In region (%)	Near (%)	Far (%)
Rule-dominant	Under 0.25	Rule value with the interval of the discrepancy process; flagged as extrapolation	Engineer review and a test before use	9.9	52.0	78.6
Hybrid	0.25 to 0.75	Rule value plus learned discrepancy; flagged as partly constrained	Interval width of 10% or less, otherwise engineer review	58.4	37.2	15.3
Data-driven	Over 0.75	Rule value plus learned discrepancy; no flag	Interval width of 10% or less, then release	31.7	10.8	6.1

Computed in this paper. The benchmark reports the composition of the weight. The predictions scored in Table 7 are the hybrid predictions in every regime, and the benchmark does not score a router that switches among separately trained predictors.

The composition follows the evidence. At 60 points 31.7 percent of in-region queries fell in the data-driven regime, 10.8 percent of near queries and 6.1 percent of far queries. The rule-dominant share rose from 9.9 percent in the region to 52.0 percent near it and 78.6 percent far from it. The weight also depends on the training size and on the fitted amplitude of the discrepancy process. The in-region data-driven share was 50.8 percent at 30 points, 31.7 percent at 60 points and 88.0 percent at 120 points, so the share did not rise steadily with data. The benchmark does not explain the dip at 60 points. The weight is a useful ordering of queries and a poor absolute scale, and the thresholds of 0.25 and 0.75 need calibration for each problem.

The router combines two signals that can disagree. The weight reports how well the data constrain the discrepancy at the query, and the out-of-distribution flag reports the distance of the query from the training inputs. When the two disagree the more conservative one governs: a query flagged as outside the envelope is released as rule-dominant whatever its weight, and a query with a low weight is treated as rule-dominant whatever its distance. This precedence is proposed in this paper and the benchmark did not test it.

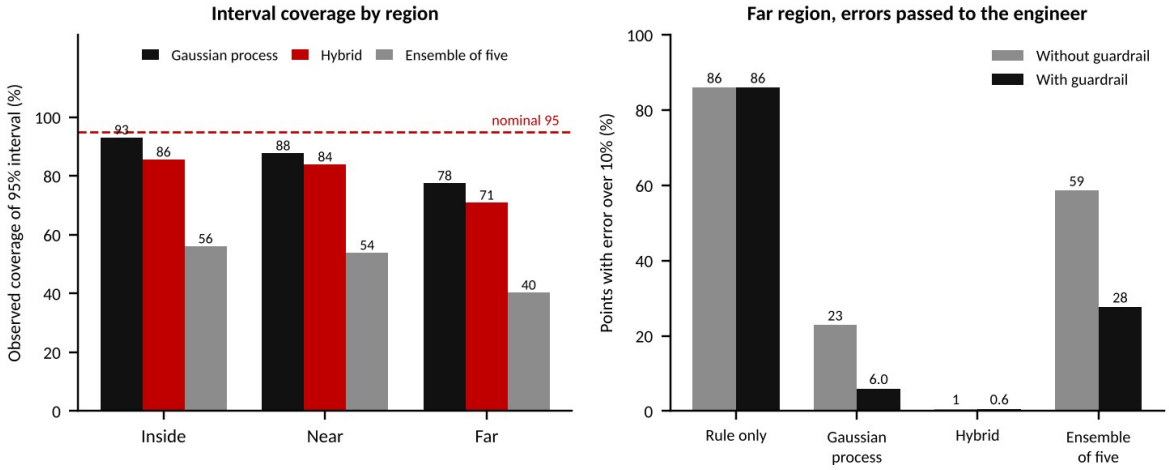
A rule-dominant result is a usable result. In the benchmark the rule model alone was wrong by 15.7 percent far from the region, and the hybrid in the same regime was wrong by 3.1 percent. The regime flag tells the engineer how much of the answer came from codified physics and how much from learned data, and the flag changes the action: a rule-dominant value goes to a test, and a data-driven value goes to release when the interval is narrow.

5.3 An interval-width guardrail cut far-region errors over 10 percent for the Gaussian process

from 22.9 to 6.0 percent of points

The guardrail accepts a prediction when 1.96 times its predictive deviation is 10 percent or less and sends the rest to an engineer. The rule is a selective prediction in the sense of Geifman and El-Yaniv.[77] Figure 4 shows the result at 60 points. In the far region the Gaussian process accepted 51.5 percent of points, and the share of all points that were accepted and also wrong by more than 10 percent was 6.0 percent, against 22.9 percent with no guardrail. The ensemble accepted 51.3 percent and left 27.6 percent against 58.6 percent. The hybrid accepted 100.0 percent and its errors over 10 percent were 0.6 percent whether or not the guardrail applied. At 120 points the guardrail cut the Gaussian process far-region figure from 18.4 to 8.4 percent and the ensemble figure from 52.8 to 11.8 percent.

Figure 4. Interval coverage by zone and far-region errors over 10 percent with and without the guardrail



Computed in this paper at 60 training points on a synthetic five-input problem; values are not CF-JET results. The guardrail threshold is the interval width defined for Table 7.

The guardrail removed errors in proportion to the quality of the intervals. For the Gaussian process the intervals under-covered in the far region at 77.5 percent, and 6.0 percent of far-region points were still accepted and wrong. The guardrail works only as well as the calibration of the interval it reads. The hybrid result shows the limit from the other side: its intervals were narrow enough to accept every point, and its errors were small in this problem, but its coverage was 70.9 percent, so the acceptance rested on intervals that were too narrow. The two results together show why calibration is a precondition of any threshold.

6. Army Research Laboratory Publications on Sparse Data, Regime Detection and Edge Deployment

6.1 ARL publications address sparse data and uncertainty, and no retrieved publication addresses regime detection for thermal hardware

ARL-BAA-0038 states that the Army has limited data with which to train or adapt AI and ML systems, expects domain-expert rules to reduce the need for data, and expects uncertainty-aware processing over neuro-symbolic architectures to clarify the limits of data-driven inference.[78] The topic lists five research goals, which this paper labels T1 to T5 in the order of the topic text: optimal learning and inference with limited supervised data, computationally efficient approximation, criteria for the sufficiency of

conventional AI and ML, efficient processing on edge devices, and distributed learning with graph neural networks. Fossaceca and Young of the Laboratory describe online nonparametric learning that adapts to variable distributions in environments with sparse exemplars.[79] A 2017 Laboratory presentation lists efficient learning with quantified uncertainty for situational understanding at the edge as a challenge.[80] Sensoy, Kaplan and Kandemir, with a co-author at the Laboratory, published the evidential method for classification.[81] The Laboratory physics-guided network for melt pool width enforces physical constraints through custom loss functions and its abstract reports a higher R^2 than a conventional network.[82] Table 9 lists these records with the goals they bear on.

Table 9. Publications on sparse-data AI, uncertainty and edge deployment, mapped to ARL-BAA-0038 goals

Source	Subject	Goals	Evidence
ARL-BAA-0038, Current Research Topics[83]	Extremely sparse data; domain rules; uncertainty-aware neuro-symbolic processing; five research goals	T1 to T5	Government solicitation
Fossaceca and Young 2018[84]	Online nonparametric learning that adapts to variable distributions with sparse exemplars	T1, T3	Peer-reviewed conference paper
Army Research Laboratory 2017[85]	Efficient learning with quantified uncertainty for situational understanding at the edge	T1, T4	Government presentation
Sensoy, Kaplan, and Kandemir 2018[86]	Evidential deep learning for classification uncertainty; one co-author at the Laboratory	T1, T3	Peer-reviewed
Army Research Laboratory physics-guided network, tech brief ARL-9655[8788]	Physical constraints in the loss function; R^2 of 0.991 against 0.884 for a conventional network, directed energy deposition	T1	Government report abstract; trade summary
Glaz et al. 2018[89]	Army science planning meetings on reasoning about complex and uncertain worlds	T1, T3	Government report
ARL Public Affairs 2021[90]	Uncertainty and robustness in data fusion systems of the Internet of Battlefield Things; no method or quantitative result named in the article	T3	Government press release
Sandia National Laboratories 2023[91]	Probabilistic transfer learning for sparse and noisy data; not an Army laboratory	T1	Government laboratory report

The goal labels T1 to T5 number the five research goals in the order of the topic text. The mapping of each record to goals is an interpretation made for this paper. The Glaz et al. report was read for scope, and the Sandia report is listed as an adjacent source.

The retrieved publications supply methods for learning with sparse data and for quantifying uncertainty, and none reports a method that detects a change of regime in thermal hardware. Fossaceca and Young come nearest, with adaptation to distributions that vary. The Laboratory press release on uncertainty for the Internet of Battlefield Things names no method and no quantitative result, so it supports the need and not the technique.[92] The Sandia report on transfer learning for sparse and noisy data is a government laboratory result outside the Army.[93] Heim and Frank name the limits of edge devices for detection, which is the only retrieved statement on edge deployment of an out-of-distribution detector.[94]

6.2 Goals T1 to T3 correspond to results in this paper, and goals T4 and T5 have no matching element in the CF-JET program description

Table 10. ARL-BAA-0038 goals T1 to T5 mapped to CF-JET elements and results of this paper

Goal	CF-JET element	Result in this paper	Gap	Status
T1. Optimal learning and inference with limited supervised data	Recalibration of surrogates from coupon, single-jet and prototype data[9596]	Far-region mean error 3.0% for the hybrid and 9.1% for the Gaussian process at 30 training points (computed)	Synthetic problem with smooth rule error; no hardware data	Direct
T2. Computationally efficient approximation	Surrogates that replace repeated conjugate CFD runs	No inference time measured in this paper	Latency of the hybrid on field outputs untested	Direct, not tested
T3. Criteria for the sufficiency of conventional AI and ML	Out-of-distribution gate and regime router	Area under the curve 0.947 near and 0.999 far for input distance (computed)	The benchmark envelope is a box, and the real envelope has an irregular form	Direct
T4. Efficient processing on edge devices	None in the program description	None	Edge limits named only in a practitioner blog[97]	No matching element
T5. Distributed learning with graph neural networks	None in the program description	None	None	No matching element

Goal labels as in Table 9. The correspondence to CF-JET elements is an interpretation made for this paper.

Goal T4 has no counterpart in the program description, and the benchmark supplies no timing. The cost of the methods can be stated from their structure. An exact Gaussian process stores the training inputs and a factorization of the kernel matrix, evaluates the mean with one kernel evaluation per training point, and evaluates the variance at a cost that grows with the square of the number of training points.[98] At the 30 to 400 points of the benchmark these costs are small, and a campaign with thousands of simulation points would need sparse or approximate methods. Heim and Frank identify limits of memory and compute for detectors on edge devices.[99] A measurement of inference time on the target device is the missing evidence.

7. Governance for Engineer-in-the-Loop Use

7.1 DoD and NIST guidance names traceability, reliability and governability, and no retrieved source prescribes engineer sign-off for a learned surrogate of thermal hardware

A Chief Digital and Artificial Intelligence Office presentation lists five Department of Defense AI ethical principles: responsible, equitable, traceable, reliable and governable, and it describes the tracking of issues through statements of concern.[100] The Office released the Responsible AI Toolkit on 14 November 2023.[101] A law firm summary of the release describes a six-step process called SHIELD: set foundations, hone operationalizations, improve and innovate, evaluate status, log for traceability, and detect through continuous monitoring.[102] The presentation read for this paper does not name the steps. DoD Instruction 5000.97 on digital engineering designates an authoritative source of truth for data and models.[103] The NIST AI Risk Management Framework organizes risk work into four functions: govern, map, measure and manage.[104] The W3C provenance data model describes the entities, activities and agents that produced a piece of data.[105]

ASME V&V 40-2018 sets out a risk-informed credibility framework for computational models, built around the context of use and a set of credibility factors, and a publisher summary counts thirteen of them.[106] The standard addresses medical devices, so its use for a cooler is an analogy. The ASME standard V&V 20 addresses verification and validation in computational fluid dynamics and heat transfer, which makes it

the nearer standard for the simulation tiers.[107] None of these sources prescribes sign-off by an engineer for a learned surrogate of thermal hardware, or a versioned rule set for physics rules. The practices in Table 11 are assembled from the cited elements, and the assembly is a proposal.

Table 11. Governance guidance mapped to CF-JET decision-record elements

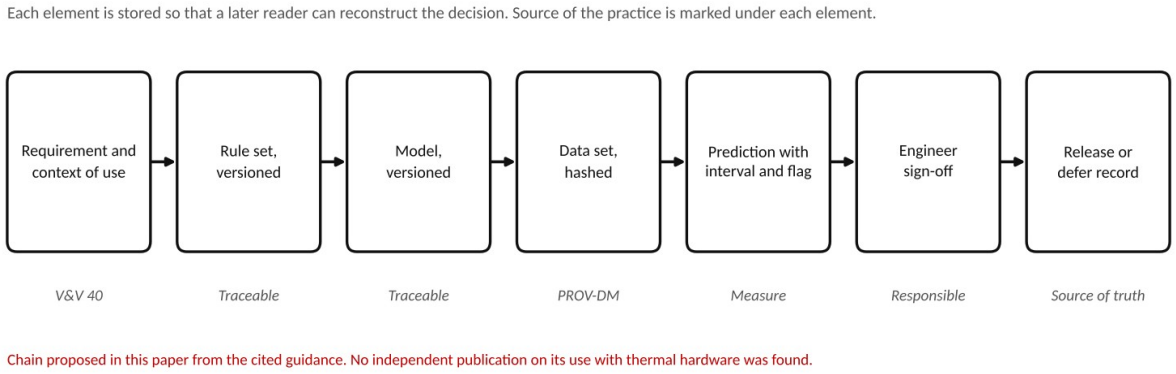
Source	Element	CF-JET decision-record element	Status of the correspondence
DoD AI ethical principles[108]	Traceable, reliable, governable, responsible, equitable	Versions of rule set, model and data linked to each prediction; coverage tests; deferral by the gate; sign-off by a named engineer	Proposed; the equitable principle has no hardware-design counterpart
SHIELD steps[109]	Log for traceability; detect through continuous monitoring	Provenance database; drift check at each hardware batch	Proposed; law firm source for the step names
DoD Instruction 5000.97[110]	Authoritative source of truth for data and models	The provenance database as the single record of decisions	Proposed
NIST AI RMF 1.0[111]	Govern, map, measure, manage	Rule versioning and roles; context of use; coverage and calibration; deferral and recalibration	Proposed
ASME V&V 40-2018[112]	Context of use; risk-informed credibility; credibility factors	Requirement and context of use stated before a model is used for release	Analogy; the standard targets medical devices
ASME V&V 20[113]	Verification and validation in CFD and heat transfer	Validation of the simulation tiers against single-jet data	Direct for simulation tiers
W3C PROV-DM[114]	Entities, activities, agents	Fields of the decision record	Proposed

The correspondences are an assembly made for this paper. The SHIELD step names come from a law firm summary, and the ASME V&V 40 description comes from a publisher blog; the standards were not read.

7.2 A decision-record chain assembled from the cited guidance links each design release to versions, intervals and a named engineer

Figure 5 shows the chain. A requirement and a context of use open the record. A versioned rule set, a versioned model and a hashed data set produce a prediction with its interval, regime and flag. A named engineer accepts, tests or defers, and the release or deferral enters the provenance database as the source of truth. The chain is proposed in this paper from the cited guidance. No independent publication on its use with thermal hardware was found. The SOL documents describe a SQL database with provenance, and they do not describe versioning of rules or engineer sign-off.[115]

Figure 5. Decision-record chain for a design release based on a learned surrogate



Bracketed labels name the guidance element that each step follows.[116117118119120]

The engineer who signs a release needs five items together: the predicted value, the interval, the regime and its weight, the out-of-distribution flag, and the versions that produced the prediction. The sign-off has three outcomes. Accept releases the value with its interval into the design record. Test sends the design point to the coming measurement batch with the flagged value as the prediction to be checked. Defer withholds the point until more evidence exists. Each outcome is a row in the provenance database that names the engineer, the date and the rationale in a short free-text field.

The versioning of rules deserves emphasis. In a hybrid model the rule set is part of the model, and a change to a correlation coefficient changes every prediction that depends on it. A record that stores the model version and omits the rule version cannot reproduce a past prediction. The same holds for the data set, which the chain stores by hash so that a recalibration after a new hardware batch can be separated from a prediction made with the old data.

8. Implications for the CF-JET Effort

8.1 Risks

Table 12. Risks to uncertainty-aware neuro-symbolic inference in CF-JET

Risk	Evidence	Mitigation	Evidence type
Hybrid intervals that are too narrow	Far-region coverage 70.9% at 60 points and 82.2% at 120 points against a nominal 95%	Recalibrate intervals on held-out hardware data before any acceptance threshold applies	Computed
Rule error that changes abruptly	The benchmark rule error is smooth; learned closures show limits of generalization[121]	Place hardware tests at suspected regime boundaries; test the hybrid on a rule with a discontinuity	Gap
Ensemble spread read as remaining ignorance	Far-region disagreement rose from 4.73% to 9.48% as training points grew from 25 to 400	Recalibrate ensemble intervals, or use a Gaussian process for the epistemic estimate	Computed
Input-distance gate misses an irregular envelope	The benchmark region is a box, which favors input distance	Combine input distance with the epistemic deviation of the model	Computed, with a stated limit
Aleatoric noise not identified	Estimate of 1.61% with spread 0.42 at 25 points against 2.0% injected	Replicate measurements at identical settings	Computed
Unstable regime weights	In-region data-driven share of 50.8%, 31.7% and 88.0% at 30, 60 and 120 points	Calibrate the weight thresholds for each problem and log the weight with each prediction	Computed
No governance standard for thermal surrogates	V&V 40 targets medical devices; no retrieved source prescribes sign-off for a learned surrogate[122]	Adopt the decision-record chain and state the context of use for each release	Gap
Edge deployment untested	No timing or memory measurement in this paper; edge limits named in a practitioner blog[123]	Measure inference time of the hybrid on the target device	Gap

Evidence type states whether the support is computed in this paper or is a gap in the retrieved literature.

8.2 Design implications

- **Replicate measurements.** The aleatoric term of hardware data is identified directly only by replicates at identical settings. About 13 replicates at a setting give a relative standard error of 20 percent in the estimated deviation (derived from normal theory), and a few settings across the flow range show whether the noise varies with flow.
- **Two-part gate.** Input-space distance supplied the best low-cost detection in the benchmark (area under the curve 0.999 far from the region). The epistemic deviation of the hybrid responds to regions that the training data constrain poorly, and a gate that uses both covers an irregular envelope.
- **Interval-width criterion tied to the budget.** A 10 percent interval on 0.02 K/W is 0.002 K/W, which equals 3 K of case-to-coolant rise at 1.5 kW (derived). The criterion for acceptance has to be set by the engineers from the margin that the budget leaves, and the benchmark value of 10 percent illustrates the mechanism.
- **Calibration before acceptance.** Interval coverage is measured on held-out points after each hardware batch, and a scale factor or conformal procedure restores nominal coverage before the guardrail applies.[124]
- **Test before train.** Each new hardware batch is scored against the current model before it enters the training set, so the coverage of the intervals and the share of flagged points measure the model that engineers used. The batch then joins the training set and the models are refitted.
- **Versioned rules and sign-off.** The rule set, model and data set are versioned and hashed, and each release records the name of the engineer who accepted it and the regime, interval and flag that applied (Table 11).
- **Regime flags in the release record.** The regime and its weight are stored with each value, so a reviewer can see how much of the answer came from codified physics and how much from learned data (Table 8).

8.3 Open research questions

- Does the hybrid advantage in Table 7 persist when the rule error has a discontinuity at a flow-regime boundary between jets?
- Which recalibration procedure restores 95 percent coverage in the far region at 30 to 60 training points, and how much does it widen the interval?
- Does a gate that combines input distance with the hybrid epistemic deviation outperform either alone when the envelope has an irregular form?
- How should coupon, single-jet and prototype data be weighted when each has its own noise and bias?
- How do evidential regression models compare with the Gaussian process and the ensemble on the benchmark in Table 7?
- Which credibility factors of a V&V 40-style assessment apply to a surrogate whose context of use is the release of a design for printing?

9. Conclusion

Hybrid inference that combines codified physics with a learned discrepancy term is established in the

literature and supported by the benchmark computed here: far from its training data the hybrid kept a mean error of 3.1 percent where the learned model reached 7.0 percent. The published hardware evidence for such models in cooling design is a gap, and the benchmark uses a smooth rule error that favors the hybrid. The claim for CF-JET therefore stands on a synthetic result and on measurements that remain to be taken.

Three conditions decide whether the uncertainty estimates support a design decision. The intervals have to be calibrated with held-out hardware data, since every method in the benchmark under-covered outside the region. The aleatoric term has to be identified with replicates. The gate for operation outside the training envelope has to combine an input-space distance with a model score. With those conditions met, a decision record that stores versions, interval, regime and the name of the accepting engineer gives each release a traceable basis, and the guidance cited here supplies the elements from which that record is built.

Bibliography

Akin Gump Strauss Hauer & Feld LLP. “Department of Defense Releases AI Toolkit.” AI Law and Regulation Tracker, 14 November 2023. https://www.akingump.com/en/insights/ai-law-and-regulation-tracker/department-of-defense-releases-ai-toolkit. Law firm commentary.

Amini, Alexander, Wilko Schwarting, Ava Soleimany, and Daniela Rus. “Deep Evidential Regression.” *Advances in Neural Information Processing Systems* 33 (2020). arXiv:1910.02600. https://arxiv.org/abs/1910.02600.

Angelopoulos, Anastasios N., and Stephen Bates. “Conformal Prediction: A Gentle Introduction.” *Foundations and Trends in Machine Learning* 16, no. 4 (2023): 494–591. https://doi.org/10.1561/2200000101.

U.S. Army DEVCOM Army Research Laboratory. “Artificial Intelligence and Machine Learning with Extremely Sparse Data (ARL-BAA-0038).” In *Current Research Topics for DEVCOM ARL BAA for Foundational Research*, 9 December 2024. https://files.simpler.grants.gov/opportunities/5e185327-5f25-4c09-889a-43615e1f54da/attachments/e4621f8a-b19c-40ec-8c59-6611cc4fa276/Current_Research_Topics_for_DEVCOM_ARL_BAA_9_Dec2024.pdf.

U.S. Army DEVCOM Army Research Laboratory Public Affairs. “AI Research Strengthens Certainty in Battlefield Decision-Making.” U.S. Army, 9 August 2021. https://www.army.mil/article/249169. Government press release.

U.S. Army Research Laboratory. “AI & ML Research Challenges.” Presentation slides, AdvML 2017 workshop, Stanford University. https://seclab.stanford.edu/AdvML2017/slides/pham.pdf.

American Society of Mechanical Engineers. *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*. ASME V&V 20-2009 (R2016). New York: ASME, 2009. https://webstore.ansi.org/Standards/ASME/ASME202009R2016.

American Society of Mechanical Engineers. “Bioengineering Blog: An Emerging Regulatory Framework for Medical Device Simulation.” ASME, n.d. https://www.asme.org/topics-resources/content/bioengineering-blog-an-emerging-regulatory-framework-for-medical-device-simulation. Publisher blog describing standard ASME V&V 40-2018; the standard itself was not read.

Cai, Shengze, Zhicheng Wang, Sifan Wang, Paris Perdikaris, and George Em Karniadakis. “Physics-Informed Neural Networks for Heat Transfer Problems.” *Journal of Heat Transfer* 143, no. 6 (2021). https://doi.org/10.1115/1.4050542.

Johnson, Matthew, and Michael Hanna. “DAGR, SHIELD, and the DoD Responsible AI Toolkit.” Presentation, Chief Digital and Artificial Intelligence Office, 20 February 2024. https://www.ai.mil/docs/ADOD24/DAGR_SHIELD_and_the_DoD_RAI_Toolkit.pdf.

Chief Digital and Artificial Intelligence Office. “Responsible AI Toolkit.” News release, 14 November 2023.

https://www.ai.mil/Latest/Blog/Article-Display/Article/3940314/responsible-ai-toolkit/.

U.S. Department of Defense. *Digital Engineering*. DoD Instruction 5000.97, 21 December 2023. https://irp.fas.org/doddir/dod/i5000_97.pdf.

Duraisamy, Karthik, Gianluca Iaccarino, and Heng Xiao. "Turbulence Modeling in the Age of Data." *Annual Review of Fluid Mechanics* 51 (2019): 357–377. https://doi.org/10.1146/annurev-fluid-010518-040547.

Fossaceca, John M., and Stuart H. Young. "Artificial Intelligence and Machine Learning for Future Army Applications." *Proceedings of SPIE* 10635 (2018): 1063507. https://doi.org/10.1117/12.2307753.

Garcez, Artur d'Avila, and Luís C. Lamb. "Neurosymbolic AI: The 3rd Wave." *Artificial Intelligence Review* 56, no. 11 (2023): 12387–12406. https://doi.org/10.1007/s10462-023-10448-w.

Geifman, Yonatan, and Ran El-Yaniv. "Selective Classification for Deep Neural Networks." *Advances in Neural Information Processing Systems* 30 (2017): 4878–4887. https://arxiv.org/abs/1705.08500.

Glaz, Bryan, Brett Piekarski, Brian Sadler, Lisa Troyer, Peter Khooshabehadeh, Brian J. Henz, Purush Iyer, Tien Pham, Ernest Chin, Jaroslaw Knap, Brent Kraczek, MaryAnn Fields, Brian Jalaian, and Alexander Kott. *Reasoning about Complex and Uncertain Worlds in Physical, Social, and Virtual Realms: A Report of the Army Science Planning and Strategy Meetings Held in Fiscal Year 2018 at the US Army Research Laboratory*. ARL-SR-0397. Aberdeen Proving Ground, MD: U.S. Army Research Laboratory, May 2018. https://apps.dtic.mil/sti/pdfs/AD1052445.pdf.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. "On Calibration of Modern Neural Networks." *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70 (2017): 1321–1330. https://proceedings.mlr.press/v70/guo17a.html.

Kapteyn, Michael G., Jacob V. R. Pretorius, and Karen E. Willcox. "A Probabilistic Graphical Model Foundation for Enabling Predictive Digital Twins at Scale." *Nature Computational Science* 1, no. 5 (2021). https://www.nature.com/articles/s43588-021-00069-0.

Karniadakis, George Em, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. "Physics-Informed Machine Learning." *Nature Reviews Physics* 3, no. 6 (2021): 422–440. https://doi.org/10.1038/s42254-021-00314-5.

Kendall, Alex, and Yarin Gal. "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" *Advances in Neural Information Processing Systems* 30 (2017): 5574–5584. https://arxiv.org/abs/1703.04977.

Kennedy, Marc C., and Anthony O'Hagan. "Bayesian Calibration of Computer Models." *Journal of the Royal Statistical Society: Series B* 63, no. 3 (2001): 425–464. https://doi.org/10.1111/1467-9868.00294.

Kuleshov, Volodymyr, Nathan Fenner, and Stefano Ermon. "Accurate Uncertainties for Deep Learning Using Calibrated Regression." *Proceedings of Machine Learning Research* 80 (2018): 2796–2804.

Lakshminarayanan, Balaji, Alexander Pritzel, and Charles Blundell. "Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles." *Advances in Neural Information Processing Systems* 30 (2017): 6402–6413.

Lee, Kimin, Kibok Lee, Honglak Lee, and Jinwoo Shin. "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks." *Advances in Neural Information Processing Systems* 31 (2018). arXiv:1807.03888. https://arxiv.org/abs/1807.03888.

Liu, Weitang, Xiaoyun Wang, John D. Owens, and Yixuan Li. "Energy-Based Out-of-Distribution Detection." *Advances in Neural Information Processing Systems* 33 (2020). arXiv:2010.03759. https://arxiv.org/abs/2010.03759.

Mandler, Hannes, and Bernhard Weigand. "Generalization Limits of Data-Driven Turbulence Models." *Flow, Turbulence and Combustion* 115 (2025): 1059–1094. https://doi.org/10.1007/s10494-024-00595-7.

Mock, Clara, Christopher Rinderspacher, and Brandon McWilliams. "Physics-Guided Neural Network for Regularization and Learning Unbalanced Data Sets: A Priori Prediction of Melt Pool Width Variation in Directed Energy Deposition." DEVCOM Army Research Laboratory technical report, March 2023. As summarized in tech brief ARL-

9655, *Mobility Engineering Technology*, 1 June 2023. https://www.mobilityengineeringtech.com/component/content/article/48330-arl-9655. Trade summary of the report; the abstract of the report was read through the Defense Technical Information Center record.

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST, January 2023. https://doi.org/10.6028/NIST.AI.100-1.

Ovadia, Yaniv, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift." *Advances in Neural Information Processing Systems* 32 (2019). arXiv:1906.02530. https://arxiv.org/abs/1906.02530.

"Physics-Guided Neural Network for Regularization and Learning Unbalanced Data Sets: A Priori Prediction of Melt Pool Width Variation in Directed Energy Deposition." Defense Technical Information Center, accession AD1196030. https://doi.org/10.21236/AD1196030. Authors, date and performing organization are not listed on the record page; the abstract was read and the full report was not.

Moreau, Luc, and Paolo Missier, eds. *PROV-DM: The PROV Data Model*. W3C Recommendation, 30 April 2013. https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

Rasmussen, Carl Edward, and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.

Sandia National Laboratories. "Trust-Enhancing Probabilistic Transfer Learning for Sparse and Noisy Data Environments." LDRD report SAND2023-10257R, October 2023. https://www.osti.gov/biblio/2430286.

Heim, Eric, and Cole Frank. "Out of Distribution Detection: Knowing When AI Doesn't Know." SEI Blog, Carnegie Mellon University Software Engineering Institute, 9 June 2025. https://www.sei.cmu.edu/blog/out-of-distribution-detection-knowing-when-ai-doesnt-know/.

Sensoy, Murat, Lance Kaplan, and Melih Kandemir. "Evidential Deep Learning to Quantify Classification Uncertainty." *Advances in Neural Information Processing Systems* 31 (2018). https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

Snoek, Jasper, and Zachary Nado. "Can You Trust Your Model's Uncertainty?" Google Research blog, 15 January 2020. https://research.google/blog/can-you-trust-your-models-uncertainty/.

SOL Engineering Services, LLC. *CF-JET: AI/ML-Enabled Design and Development of an Additively Manufactured, Liquid-Cooled Jet-Impingement Cold Plate with Optimized Carbon Fiber Target Plate*. Proposal, Rev Final 2, n.d.

SOL Engineering Services, LLC. *SOL Aquifer* software pipeline diagram. n.d.

SOL Engineering Services, LLC. "Modeling HVAC System Performance: Code as Architecture Approach to AI." Work carried out with Alabama A&M University students, n.d.

Notes

[1]SOL Engineering Services, LLC, *CF-JET: AI/ML-Enabled Design and Development of an Additively Manufactured, Liquid-Cooled Jet-Impingement Cold Plate with Optimized Carbon Fiber Target Plate*, proposal, Rev Final 2, n.d.

[2]SOL Engineering Services, LLC, *SOL Aquifer* software pipeline diagram, n.d.

[3]George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang, "Physics-Informed Machine Learning," *Nature Reviews Physics* 3, no. 6 (2021): 422–440, https://doi.org/10.1038/s42254-021-00314-5.

[4]Shengze Cai, Zhicheng Wang, Sifan Wang, Paris Perdikaris, and George Em Karniadakis, "Physics-Informed Neural Networks for Heat Transfer Problems," *Journal of Heat Transfer* 143, no. 6 (2021), https://doi.org/10.1115/1.4050542.

[5]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, no. 5 (2021), https://www.nature.com/articles/s43588-021-00069-0.

[6]Yaniv Ovadia et al., "Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift," *Advances in Neural Information Processing Systems* 32 (2019), arXiv:1906.02530, https://arxiv.org/abs/1906.02530.

[7]Jasper Snoek and Zachary Nado, "Can You Trust Your Model's Uncertainty?," Google Research blog, 15 January 2020, https://research.google/blog/can-you-trust-your-models-uncertainty/.

[8]U.S. Army DEVCOM Army Research Laboratory, "Artificial Intelligence and Machine Learning with Extremely Sparse Data (ARL-BAA-0038)," in *Current Research Topics for DEVCOM ARL BAA for Foundational Research*, 9 December 2024, https://files.simpler.grants.gov/opportunities/5e185327-5f25-4c09-889a-43615e1f54da/attachments/e4621f8a-b19c-40ec-8c59-6611cc4fa276/Current_Research_Topics_for_DEVCOM_ARL_BAA_9_Dec2024.pdf.

[9]Matthew Johnson and Michael Hanna, "DAGR, SHIELD, and the DoD Responsible AI Toolkit," presentation, Chief Digital and Artificial Intelligence Office, 20 February 2024, https://www.ai.mil/docs/ADOD24/DAGR_SHIELD_and_the_DoD_RAI_Toolkit.pdf.

[10]National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (Gaithersburg, MD: NIST, January 2023), https://doi.org/10.6028/NIST.AI.100-1.

[11]Luc Moreau and Paolo Missier, eds., *PROV-DM: The PROV Data Model*, W3C Recommendation (30 April 2013), https://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[12]American Society of Mechanical Engineers, "Bioengineering Blog: An Emerging Regulatory Framework for Medical Device Simulation," ASME, n.d., https://www.asme.org/topics-resources/content/bioengineering-blog-an-emerging-regulatory-framework-for-medical-device-simulation.

[13]SOL Engineering Services, *CF-JET*, Rev Final 2.

[14]SOL Engineering Services, *CF-JET*, Rev Final 2.

[15]SOL Engineering Services, *SOL Aquifer* pipeline diagram.

[16]SOL Engineering Services, *SOL Aquifer* pipeline diagram.

[17]SOL Engineering Services, *CF-JET*, Rev Final 2.

[18]SOL Engineering Services, LLC, "Modeling HVAC System Performance: Code as Architecture Approach to AI," work carried out with Alabama A&M University students, n.d.

[19]DEVCOM ARL, "ARL-BAA-0038," *Current Research Topics*.

[20]Karniadakis et al., "Physics-Informed Machine Learning."

[21]Cai et al., "Physics-Informed Neural Networks for Heat Transfer Problems."

[22]Artur d'Avila Garcez and Luís C. Lamb, "Neurosymbolic AI: The 3rd Wave," *Artificial Intelligence Review* 56, no. 11 (2023): 12387–12406, https://doi.org/10.1007/s10462-023-10448-w.

[23]Marc C. Kennedy and Anthony O'Hagan, "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society: Series B* 63, no. 3 (2001): 425–464, https://doi.org/10.1111/1467-9868.00294.

[24]Karthik Duraisamy, Gianluca Iaccarino, and Heng Xiao, "Turbulence Modeling in the Age of Data," *Annual Review of Fluid Mechanics* 51 (2019): 357–377, https://doi.org/10.1146/annurev-fluid-010518-040547.

[25]Hannes Mandler and Bernhard Weigand, "Generalization Limits of Data-Driven Turbulence Models," *Flow, Turbulence and Combustion* 115 (2025): 1059–1094, https://doi.org/10.1007/s10494-024-00595-7.

[26]Karniadakis et al., "Physics-Informed Machine Learning."

[27]Cai et al., "Physics-Informed Neural Networks for Heat Transfer Problems."

[28]Kennedy and O'Hagan, "Bayesian Calibration of Computer Models."

[29]Duraisamy, Iaccarino, and Xiao, "Turbulence Modeling in the Age of Data."

[30]Garcez and Lamb, "Neurosymbolic AI: The 3rd Wave."

[31]SOL Engineering Services, *CF-JET*, Rev Final 2.

[32]SOL Engineering Services, *SOL Aquifer* pipeline diagram.

[33]Karniadakis et al., “Physics-Informed Machine Learning.”

[34]Cai et al., “Physics-Informed Neural Networks for Heat Transfer Problems.”

[35]Kennedy and O’Hagan, “Bayesian Calibration of Computer Models.”

[36]Duraisamy, Iaccarino, and Xiao, “Turbulence Modeling in the Age of Data.”

[37]Mandler and Weigand, “Generalization Limits of Data-Driven Turbulence Models.”

[38]Kapteyn, Pretorius, and Willcox, “Probabilistic Graphical Model Foundation for Predictive Digital Twins.”

[39]Clara Mock, Christopher Rinderspacher, and Brandon McWilliams, “Physics-Guided Neural Network for Regularization and Learning Unbalanced Data Sets: A Priori Prediction of Melt Pool Width Variation in Directed Energy Deposition,” DEVCOM Army Research Laboratory technical report, March 2023, as summarized in tech brief ARL-9655, *Mobility Engineering Technology*, 1 June 2023, https://www.mobilityengineeringtech.com/component/content/article/48330-arl-9655.

[40]“Physics-Guided Neural Network for Regularization and Learning Unbalanced Data Sets: A Priori Prediction of Melt Pool Width Variation in Directed Energy Deposition,” Defense Technical Information Center, accession AD1196030, https://doi.org/10.21236/AD1196030.

[41]Garcez and Lamb, “Neurosymbolic AI: The 3rd Wave.”

[42]SOL Engineering Services, *CF-JET*, Rev Final 2.

[43]SOL Engineering Services, *SOL Aquifer* pipeline diagram.

[44]Kapteyn, Pretorius, and Willcox, “Probabilistic Graphical Model Foundation for Predictive Digital Twins.”

[45]“Physics-Guided Neural Network,” DTIC AD1196030.

[46]Alex Kendall and Yarin Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?,” *Advances in Neural Information Processing Systems* 30 (2017): 5574–5584, https://arxiv.org/abs/1703.04977.

[47]Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning* (Cambridge, MA: MIT Press, 2006).

[48]Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell, “Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles,” *Advances in Neural Information Processing Systems* 30 (2017): 6402–6413.

[49]Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus, “Deep Evidential Regression,” *Advances in Neural Information Processing Systems* 33 (2020), arXiv:1910.02600, https://arxiv.org/abs/1910.02600.

[50]Murat Sensoy, Lance Kaplan, and Melih Kandemir, “Evidential Deep Learning to Quantify Classification Uncertainty,” *Advances in Neural Information Processing Systems* 31 (2018), https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

[51]Anastasios N. Angelopoulos and Stephen Bates, “Conformal Prediction: A Gentle Introduction,” *Foundations and Trends in Machine Learning* 16, no. 4 (2023): 494–591, https://doi.org/10.1561/2200000101.

[52]Rasmussen and Williams, *Gaussian Processes for Machine Learning*.

[53]Ovadia et al., “Can You Trust Your Model’s Uncertainty?”

[54]Lakshminarayanan, Pritzel, and Blundell, “Deep Ensembles.”

[55]Amini et al., “Deep Evidential Regression.”

[56]Angelopoulos and Bates, “Conformal Prediction.”

[57]Ovadia et al., “Can You Trust Your Model’s Uncertainty?”

[58]Kennedy and O’Hagan, “Bayesian Calibration of Computer Models.”

[59]Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger, “On Calibration of Modern Neural Networks,” *Proceedings of the 34th International Conference on Machine Learning*, PMLR 70 (2017): 1321–1330, https://proceedings.mlr.press/v70/guo17a.html.

[60]Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon, “Accurate Uncertainties for Deep Learning Using Calibrated Regression,” *Proceedings of Machine Learning Research* 80 (2018): 2796–2804.

[61]Ovadia et al., “Can You Trust Your Model’s Uncertainty?”

[62]Snoek and Nado, “Can You Trust Your Model’s Uncertainty?”

[63]Guo et al., “On Calibration of Modern Neural Networks.”

[64]Kuleshov, Fenner, and Ermon, “Accurate Uncertainties for Deep Learning.”

[65]Angelopoulos and Bates, “Conformal Prediction.”

[66]Yonatan Geifman and Ran El-Yaniv, “Selective Classification for Deep Neural Networks,” *Advances in Neural Information Processing Systems* 30 (2017): 4878–4887, https://arxiv.org/abs/1705.08500.

[67]Eric Heim and Cole Frank, “Out of Distribution Detection: Knowing When AI Doesn’t Know,” SEI Blog, Carnegie Mellon University Software Engineering Institute, 9 June 2025, https://www.sei.cmu.edu/blog/out-of-distribution-detection-knowing-when-ai-doesnt-know/.

[68]Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin, “A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks,” *Advances in Neural Information Processing Systems* 31 (2018), arXiv:1807.03888, https://arxiv.org/abs/1807.03888.

[69]Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li, “Energy-Based Out-of-Distribution Detection,” *Advances in Neural Information Processing Systems* 33 (2020), arXiv:2010.03759, https://arxiv.org/abs/2010.03759.

[70]Geifman and El-Yaniv, “Selective Classification for Deep Neural Networks.”

[71]Lee et al., “Detecting Out-of-Distribution Samples and Adversarial Attacks.”

[72]Lakshminarayanan, Pritzel, and Blundell, “Deep Ensembles.”

[73]Liu et al., “Energy-Based Out-of-Distribution Detection.”

[74]Geifman and El-Yaniv, “Selective Classification for Deep Neural Networks.”

[75]Heim and Frank, “Out of Distribution Detection.”

[76]Kennedy and O’Hagan, “Bayesian Calibration of Computer Models.”

[77]Geifman and El-Yaniv, “Selective Classification for Deep Neural Networks.”

[78]DEVCOM ARL, “ARL-BAA-0038,” *Current Research Topics*.

[79]John M. Fossaceca and Stuart H. Young, “Artificial Intelligence and Machine Learning for Future Army Applications,” *Proceedings of SPIE* 10635 (2018): 1063507, https://doi.org/10.1117/12.2307753.

[80]U.S. Army Research Laboratory, “AI & ML Research Challenges,” presentation slides, AdvML 2017 workshop, Stanford University, https://seclab.stanford.edu/AdvML2017/slides/pham.pdf.

[81]Sensoy, Kaplan, and Kandemir, “Evidential Deep Learning.”

[82]“Physics-Guided Neural Network,” DTIC AD1196030.

[83]DEVCOM ARL, “ARL-BAA-0038,” *Current Research Topics*.

[84]Fossaceca and Young, “Artificial Intelligence and Machine Learning for Future Army Applications.”

[85]Army Research Laboratory, “AI & ML Research Challenges.”

[86]Sensoy, Kaplan, and Kandemir, “Evidential Deep Learning.”

[87]Mock, Rinderspacher, and McWilliams, “Physics-Guided Neural Network,” tech brief ARL-9655.

[88]“Physics-Guided Neural Network,” DTIC AD1196030.

[89]Bryan Glaz et al., *Reasoning about Complex and Uncertain Worlds in Physical, Social, and Virtual Realms: A Report of the Army Science Planning and Strategy Meetings Held in Fiscal Year 2018 at the US Army Research Laboratory*, ARL-SR-0397 (Aberdeen Proving Ground, MD: U.S. Army Research Laboratory, May 2018), https://apps.dtic.mil/sti/pdfs/AD1052445.pdf.

[90]U.S. Army DEVCOM Army Research Laboratory Public Affairs, “AI Research Strengthens Certainty in Battlefield Decision-Making,” U.S. Army, 9 August 2021, https://www.army.mil/article/249169.

[91]Sandia National Laboratories, “Trust-Enhancing Probabilistic Transfer Learning for Sparse and Noisy Data Environments,” LDRD report SAND2023-10257R (October 2023), https://www.osti.gov/biblio/2430286.

[92]ARL Public Affairs, “AI Research Strengthens Certainty in Battlefield Decision-Making.”

[93]Sandia National Laboratories, “Trust-Enhancing Probabilistic Transfer Learning.”

[94]Heim and Frank, “Out of Distribution Detection.”

[95]SOL Engineering Services, *CF-JET*, Rev Final 2.

[96]SOL Engineering Services, *SOL Aquifer* pipeline diagram.

[97]Heim and Frank, “Out of Distribution Detection.”

[98]Rasmussen and Williams, *Gaussian Processes for Machine Learning*.

[99]Heim and Frank, “Out of Distribution Detection.”

[100]Johnson and Hanna, “DAGR, SHIELD, and the DoD Responsible AI Toolkit.”

[101]Chief Digital and Artificial Intelligence Office, “Responsible AI Toolkit,” news release, 14 November 2023, https://www.ai.mil/Latest/Blog/Article-Display/Article/3940314/responsible-ai-toolkit/.

[102]Akin Gump Strauss Hauer & Feld LLP, “Department of Defense Releases AI Toolkit,” AI Law and Regulation Tracker, 14 November 2023, https://www.akingump.com/en/insights/ai-law-and-regulation-tracker/department-of-defense-releases-ai-toolkit.

[103]U.S. Department of Defense, *Digital Engineering*, DoD Instruction 5000.97 (21 December 2023), https://irp.fas.org/doddir/dod/i5000_97.pdf.

[104]NIST, *AI Risk Management Framework (AI RMF 1.0)*.

[105]Moreau and Missier, *PROV-DM*.

[106]ASME, “An Emerging Regulatory Framework for Medical Device Simulation.”

[107]American Society of Mechanical Engineers, *Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer*, ASME V&V 20-2009 (R2016) (New York: ASME, 2009), https://webstore.ansi.org/Standards/ASME/ASME202009R2016.

[108]Johnson and Hanna, “DAGR, SHIELD, and the DoD Responsible AI Toolkit.”

[109]Akin Gump, “Department of Defense Releases AI Toolkit.”

[110]U.S. Department of Defense, DoD Instruction 5000.97, *Digital Engineering*.

[111]NIST, *AI Risk Management Framework (AI RMF 1.0)*.

[112]ASME, “An Emerging Regulatory Framework for Medical Device Simulation.”

[113]ASME, *V&V 20-2009*.

[114]Moreau and Missier, *PROV-DM*.

[115]SOL Engineering Services, *SOL Aquifer* pipeline diagram.

[116]Johnson and Hanna, “DAGR, SHIELD, and the DoD Responsible AI Toolkit.”

[117]NIST, *AI Risk Management Framework (AI RMF 1.0)*.

[118]Moreau and Missier, *PROV-DM*.

[119]ASME, “An Emerging Regulatory Framework for Medical Device Simulation.”

[120]U.S. Department of Defense, DoD Instruction 5000.97, *Digital Engineering*.

[121]Mandler and Weigand, “Generalization Limits of Data-Driven Turbulence Models.”

[122]ASME, “An Emerging Regulatory Framework for Medical Device Simulation.”

[123]Heim and Frank, “Out of Distribution Detection.”

[124]Angelopoulos and Bates, “Conformal Prediction.”