Reliability assessment of guided wave damage localization with deep learning uncertainty quantification methods

Ishan D Khurjekar[a], Joel B. Harley[b]

[a]Scripps Institute of Oceanography, University of California San Diego, La Jolla, USA
[b]Electrical and Computer Engineering department, University of Florida, Gainesville, USA

ARTICLE INFO

Keywords:
Reliability assessment
Machine learning
Structural health monitoring
Guided waves
Uncertainty quantification

ABSTRACT

Guided wave-based structural health monitoring is an attractive option for detecting structural defects in an automated manner. In this work, we focus on the task of damage localization. Deep learning methods have been shown to have superior performance for damage localization. Yet, environmental variations introduce uncertainty in the system and reduce its reliability. For this reason, it is crucial to assess the reliability of estimates taken from structural health monitoring systems. In this work, we estimate the localization reliability from a single snapshot of sparse array guided wave measurements instead of reporting values averaged over an entire set of test measurements. The assessment strategy can be added to any deep learning localization model and produces both a localization and uncertainty estimate. The deep learning model is trained using only guided wave simulations. We assess the uncertainty using both simulated and experimental data with temperature variations. Multiple deep learning-based uncertainty quantification methods are analyzed. Results demonstrate correlations between uncertainty, temperature variations, and the presence of synthetic damage. We also compare with reliability derived from delay-and-sum localization. We find that a deep ensemble learning strategy provides the most reliable damage localization and uncertainty quantification.

1. Introduction and Background

Non-destructive testing (NDT) is routinely used across a wide variety of industries to monitor structures for defects [1]. NDT methods are usually employed through schedule-based human inspection. Structural health monitoring (SHM) is intended to supplement NDT through constant automated monitoring [2] to reduce manual labor and downtime through predictive maintenance.

Guided wave SHM is a popular technique for structural damage assessment [3]. Guided waves are usually transmitted into a structure by PZT (lead zirconate titanate) transducers, guided through the geometry of the structure, and then received by another collection of transducers. Variations in the data across time are analyzed to detect and locate damage. Most guided wave SHM methods depend on the measurement's similarity with a damage-free baseline measurement. For example, many localization algorithms process baseline subtracted data [4, 5]. Environmental and operational variations change signals, causing differences that are unrelated to damage. Temperature is one of the most commonly occurring environmental variations [6, 7]. It changes the velocity, amplitude, and phase [3]. Velocity changes approximately stretch or shrink measurements, and a small change in velocity can cause a large change in baseline subtraction. This leads to significant localization errors [8]. Common strategies for reducing baseline subtraction errors include optimal baseline selection [9] and baseline signal stretch [10]. Yet, these methods still produce artifacts and errors that reduce the system's performance.

The effects of the environmental variations on SHM system performance can be quantified by the reliability of the system. Reliability is the ability of the system to perform the same task without any failures for a given time span [11]. As SHM relies on constant and automated monitoring of structures, variations in operating conditions over time and space affect the performance. Measuring reliability in guided wave SHM is thus a difficult challenge. In this paper, reliability is assessed through uncertainty quantification to provide a metric of uncertainty associated with each localization estimate.

Initial efforts for guided wave SHM reliability assessment have focused on adopting NDT reliability assessment techniques [12, 13]. Challenges in reliability assessment of guided wave SHM systems due to external variabilities and contributing factors are discussed in [14]. Examples of SHM reliability assessment methods include probability density

ORCID(s):

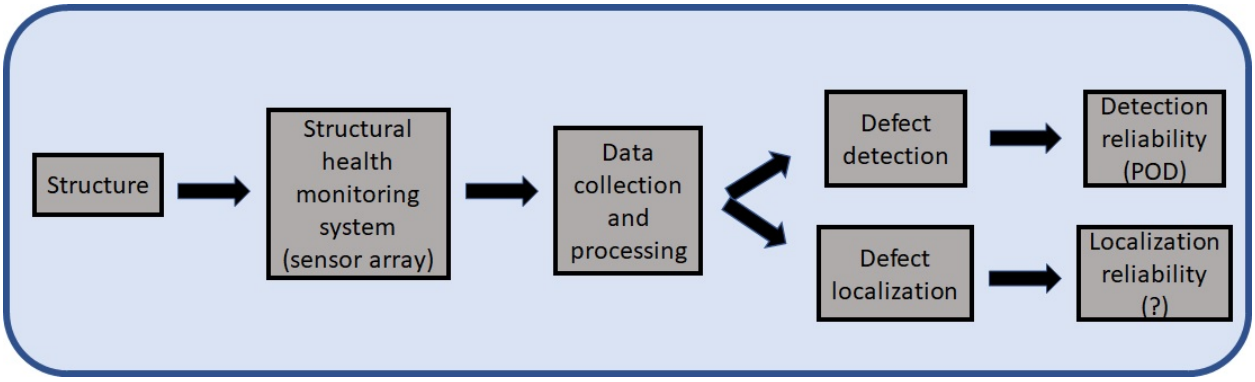

Figure 1: SHM reliability requirement

function (PDF) based methods [15] and model-assisted localization curves (MAPOL) [16] for damage localization. Many of these approaches require a computationally expensive simulation study. Existing approaches then perform an overall reliability assessment based on many different experimental measurements [16]. However, these methods do not provide a local reliability assessment for each localization estimate as illustrated in Fig. 1. Such an ability would enable a real-time and ad hoc assessment of uncertainty. Yet, with traditional statistics, this would require an impractical computational effort to map a more than 1000-dimensional measurement space to a scalar reliability metric.

Simultaneously, researchers have proposed machine learning (ML) and deep neural network(DNN) based methods for SHM [17] and NDE [18]. Empirical evidence shows that DNN methods have the ability to extract abstract representations of the input data [19], which is crucial for learning from the dispersive guided wave data. Yet, there is no existing discussion on guided wave-based damage localization reliability using deep learning methods.

In this work, we compare the ability of different DNN methods including ensemble, Monte-Carlo dropout, and Gaussian likelihood networks, to obtain reliability estimates conditioned on individual localization attempts. The only condition of the deep learning method is that it should be able to produce an uncertainty scalar estimate from the model prediction. Three deep learning methods and one traditional signal processing localization method are evaluated. In this paper, the deep learning methods are trained exclusively on simulated data. Each method is tested with simulated and experimental guided wave data from an aluminum plate exposed to non-uniform temperature variations. While any deep learning localization model can be used, we explicitly build on a previous robust deep learning-based localization framework for guided waves [20].

The main contribution of this work is an assessment of multiple deep learning-based reliability/uncertainty assessment methods for guided wave damage localization. Results demonstrate a correlation between our chosen reliability/uncertainty metric with external temperature variations and the inclusion of synthetic damage. Specifically, we derive three observations from our analysis:
1.) Deviation from the baseline temperature increases the uncertainty level in the localization estimate
2.) The localization uncertainty is lesser for measurements where damage is present as compared to the measurements without any damage.
3.) The DNN-Ensemble method provides better damage localization reliability metrics compared to other deep learning methods.

2. Guided Wave SHM with reliability

The presence of external temperature variations leads to uncertainty in the guided wave propagation. This uncertainty propagates through to the prediction of the damage location. Quantifying this predictive uncertainty is equivalent to assessing the method's reliability. That is, a system with low predictive uncertainty is highly reliable.

Methods for SHM reliability assessment have been borrowed from the NDT literature. In NDT, probability of detection (POD) curves [12] obtained from empirical measures and/or computational models [21] are the standard measures for reliability assessment. POD curves capture the relationship between $\hat{a}$ and a, where $\hat{a}$ is the measured sensor signal and a is the defect size [22]. The relationship between $\hat{a}$ and a is mathematically expressed by a regression

model that is fit using either empirical data or simulated model data [23] through a maximum likelihood procedure. The model often assumes the relationship is distorted by zero-mean normally distributed measurement noise.

The POD curve can then be expressed as $\text{POD}(a) = P(\hat{a} > \hat{a}_{th})$, where $\hat{a}_{th}$ is a detection threshold that can be determined through a receiver operator characteristic (ROC) curve [24]. This can be also be expressed as a cumulative distribution function (CDF) of a normal distribution $\phi\left((a - \hat{\mu})/\hat{\sigma}\right)$, where $\hat{\mu}$ and $\hat{\sigma}$ are parameters of a normal distribution. From this, the gold standard value of $\hat{a}_{90/95}$ can be obtained (flaw size can be detected with 90% probability at the 95% confidence level). The POD curves provide a prediction of the hit/miss probability as a function of the defect size, serving as a metric of reliability in NDT inspection.

PDF based approaches [15] require a complete sweep of model parameters, which is computationally expensive. Model-driven methods such as MAPOL [16] require access to multiple measurements, reducing real-time applicability. The reliability assessment approach should ideally:

1. Generate a reliability estimate conditioned on a single guided wave measurement.
2. Easily adapt to existing localization methods without major modifications.
3. Capture the effect of external and internal structural variations.

In the next section, we discuss several methods that satisfy these requirements.

3. Damage localization with deep learning

Deep learning methods have been proposed for SHM tasks, including damage detection [25] and localization [26]. Deep learning methods have shown superior performance under specific environmental conditions but without any performance guarantee. Deep learning models learn abstract representations through cascaded hidden layers without the need to manually craft features from inputs. This ability suits guided wave applications (dispersive input signals) that have relevant information spread temporally. A deep neural network (DNN) consists of a series of connected hidden layers with weights that are updated via multiple iterations of the optimization process [27].

Hidden layers can be fully connected dense layers, spatially-connected convolutional layers, or memory-based recurrent layers. Researchers have used convolutional neural networks (CNNs) for SHM tasks such as fault diagnosis from acoustic emission (AE) spectra [28] and guided wave damage localization [20]. We use a combination of dense and convolutional layers to learn the inverse mapping from the spatiotemporal guided wave data to the damage location. Every convolutional layer has filters (i.e., a matrix of trainable weights). The filter learns the same set of weights (weight-sharing property) and hence learns global features. The filter output is the convolution of the filter with the entire input to the convolutional layer. This is done separately for every input channel and the filter outputs are combined. The filter weights are learned during the optimization process where each filter learns distinct features.

3.1. Reliability assessment with deep learning

The deep learning-based damage localization methods do not provide any assessment of the localization reliability in the presence of external uncertainty sources. Assessing the reliability of damage localization is equivalent to quantifying the uncertainty in localization estimates. Uncertainty quantification (UQ) for deep learning is a growing research field. Many uncertainty estimation techniques in deep learning take a Bayesian approach [29, 30]. These techniques typically choose a prior distribution and develop a scheme for approximating the posterior distribution. The choice of prior is crucial and in cases where the underlying mathematical model is incomplete, such techniques fail to provide the right uncertainty estimates. Computational constraints also often limit the posterior approximation process.

Recent research in UQ for deep learning has focused on computationally scalable approaches. Researchers have proposed techniques such as Monte-Carlo dropout [31] and deep ensembling [32], among others, to provide uncertainty estimates. While not widely used in NDT or SHM, [33] recently used an ensemble approach for uncertainty quantification for plane-wave NDT imaging and [34] has used Monte Carlo Dropout and deep ensembles for assessing super-resolution guided wave imaging. We consider three deep learning methods that can provide reliability along with damage location estimates, see Fig. 2.

3.1.1. DNN-Ensemble

Ensembling is a popular approach to training robust machine learning models [35]. Consider an ensemble of N_{ens} models, $\mathcal{M}_{n=1}^{N_{\text{ens}}}$, outputting N_{ens} predictions $\hat{y}_{n=1}^{N_{\text{ens}}}$ for an input x_i. The mean of the ensemble predictions is taken as the

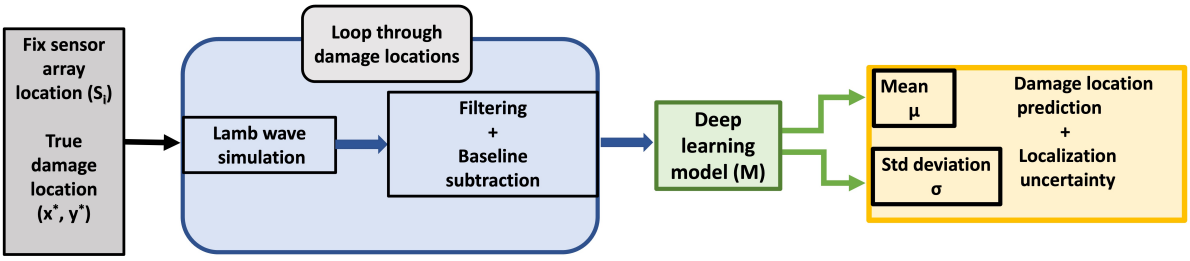

Figure 2: Setup illustrating guided wave data simulation, deep learning model training, and damage localization with reliability.

model prediction:

$$y_{\text{mean},i} = \frac{1}{N_{ens}} \sum_{n=1}^{N_{\text{ens}}} \mathcal{M}_n(x_i) \, . \tag{1}$$

The number of networks in the ensemble (N_{ens}) is a hyperparameter and can be chosen so as to maximize performance. The standard deviation of the ensemble predictions ($\hat{y}_{n=1}^{N_{\text{ens}}}$) can be used as a simple way to quantify uncertainty for the ensemble model predictions. Indeed, DNN-Ensemble (trained on the same dataset but initialized with different weights) is a simple method to provide an uncertainty estimate associated with the model predictions [32].

3.1.2. DNN-MCdropout

DNN-MCdropout is another easy way to quantify uncertainty with deep networks [31, 36]. Consider a model $\mathcal{M}$ trained on input data samples ($\mathbf{x}_i$) to predict the output y_i. For DNN-MCdropout, the model predictions are obtained by dropping out DNN nodes during both train and test times according to a preset probability rate. This is equivalent to obtaining predictions from an ensemble network ($\mathcal{M}_{m=1}^{N_{\text{MC}}}$) where the M networks have different random nodes dropped out. The number of Monte-Carlo runs (N_{MC}) is a hyperparameter and we set $N_{\text{MC}} = 100$. The mean of the Monte-Carlo predictions is taken as the model prediction:

$$y_{\text{mean},i} = \frac{1}{N_{MC}} \sum_{n=1}^{N_{\text{MC}}} \mathcal{M}_n(x_i) \, . \tag{2}$$

This method has also been shown to be a Bayesian approximation [31]. Similar to the ensemble case, the standard deviation of the MC predictions ($\hat{y}_{m=1}^{M}$) can be used for quantifying predictive uncertainty.

3.1.3. DNN-GaussianMLE

The DNN-GaussianMLE approach involves training a DNN to output mean (μ) and standard deviation (Σ) by assuming that the conditional output is Gaussian distributed:

$$\mathcal{M}(\hat{y}_i|x_i) \sim \mathcal{N}(\mu_i, \Sigma_i). \tag{3}$$

The joint Gaussian likelihood for the training dataset, $\mathbf{D}_{\text{tr}} = [\{x_1, y_1\} \ldots \{x_T, y_T\}]$, defined as

$$l = \prod_{t=1}^{T} p(\hat{y}_i|x_i), \tag{4}$$

is maximized. The network has output nodes for the mean (μ) and standard deviation (σ) predictions. In this work, we estimate only σ_x and σ_y (i.e., a diagonal covariance).

3.2. Simulation Framework

The training and validation data for all deep learning localization methods are generated with a Lamb wave model while the test data is obtained from an experimental setup, described in Sec. 4. This ensures that the deep learning

Table 1
Lamb wave model parameters

Variable	Description
Plate thickness	0.284 cm
Plate dimensions	1.2×1.2 m
Sampling frequency	$F_s = 1$ MHz
Frequencies	Q = 4000 (uniformly spaced between 0 to F_s MHz.)
Transducers	8 Transducers (M = 56 sensor pairs)
Input temporal length	$Q_t = 1000$ (1 ms)

methods learn a general mapping and do not overfit to a particular data distribution. For simulation data, the Lamb wave model for an unbounded plate at an angular frequency ω and travel distance of r is the linear superposition of different wave modes:

$$\mathbf{Z}(\omega, r) = \sum_n \sqrt{\frac{1}{\kappa_n(\omega) r}} \mathbf{S}(\omega) e^{-j\kappa_n(\omega) r} , \tag{5}$$

where $\kappa_n(\omega)$ is the wavenumber of the Lamb waves (obtained by solving the Rayleigh-Lamb equations [37]), $S(\omega)$ is the transmitted signal, and the value of n corresponds to a Lamb wave mode. In this work, we consider the zeroth order symmetric and anti-symmetric modes i.e., A0 and S0. However, after our filtering operation, the A0 mode is dominant in the guided wave signals. Table 1 lists the parameters chosen for the guided wave model. The parameters are chosen to match the guided wave experimental setup.

For deep learning methods, the transducer locations are assumed to be known beforehand and a random location within the plate dimensions (values in Table 1) is chosen as a damage location. The guided wave travels the distance between the transducer and the chosen damage location A fixed amplitude is assumed for the damage signal. As with traditional localization methods, the damage needs to be within the sensor array convex hull area for the best possible localization results. A common limitation of deep learning methods is that when the test damage location is outside the training data location distribution, the localization error also increases. Computing the localization error is not possible at test time as the true damage locations are not known at test time. As opposed to this, our method can quantify localization uncertainty without any ground truth requirement. Hence, using a combination of deep learning and a UQ component as proposed in this work, it is possible to reliably use deep learning methods for damage localization.

In (5), the frequency domain guided wave signal $\mathbf{Z}$ is a matrix of dimensions $Q \times M$, where Q is the number of equally spaced frequencies (range: 0 to F_s)and M is the number of sensor pairs (listed in Table 1). In our analysis, we utilize the input signal of length $Q_t = 1000$, corresponding to a time duration of 1 ms. This signal length contains temporal information for localization. The guided wave signal $\mathbf{Z}$ and corresponding damage location $\mathbf{d}$ form one data sample. The simulated guided wave data forms a training and validation set (listed in Table 2) and the experimental data forms the test set. Hence, the data is trained exclusively on simulated data. Note that boundary reflections in a real-world setup are an issue [38]. We describe a windowing operation to mitigate these effects in Sec. 4. All guided wave data (training, validation, and testing) is processed and filtered.

For damage localization, the received guided wave signal is assumed to be a superposition of the baseline signal and the damage signal. The baseline signal is the signal traveling directly from the transmitter to the receiver and the damage signal is the signal traveling from the baseline to the damage and then to the receiver. This is expressed as

$$\mathbf{Z}(\omega_q, r_m) = \mathbf{Z}_{bs}(\omega_q, r_m) + \beta \mathbf{Z}_d(\omega_q, r_m), \tag{6}$$

where $\mathbf{Z}_{bs}(\omega_q, r_m)$ is the baseline signal, $\mathbf{Z}_d(\omega_q, r_m)$ is the damage signal, and β is the reflection coefficient. We train the network with baseline subtracted data. Hence, $\mathbf{Z}_{bs}(\omega_q, r_m)$ is removed, but imperfectly due to temperature variations. Hence, β affects the results, but not crucially. For our simulated training data, we chose $\beta = 0.1$ and the damage is simulated at random locations across the plate.

The steps for guided wave simulation dataset generation (training data for DNN) are reiterated for clarity:
1.) The transducer locations are assumed to be known a priori and a random damage location within the plate

dimensions (see Table 1) is chosen.
2.) With the assumed damage location and guided wave parameters (as per Table 1), generate a guided wave signal using (5) that travels the fixed distance between transducers and random damage location.
3.) Guided wave signal recorded at the transducer is a linear superposition of the damage signal and baseline signal. Hence, we subtract the baseline signal (explained in Sec.4).
4.) Construct a simulation dataset by repeating steps 1-3 for many different damage locations sampled randomly (this forms the training data for DNN).

Separately, we also analyze the effect of temperature variations on simulated guided wave test data. We assume temperature only affects the velocity of propagation. The change in velocity is the predominant factor in reducing the effectiveness of baseline subtraction [6]. To change the velocity, we scale the wavenumber $\kappa_n(\omega)$ by a factor of α. The wavenumber is inversely proportional to the velocity. We will often refer to α as a stretch factor since a change in velocity is often described as stretching the data by that factor [10]. Additional details can be found in [20].

Note that we do not apply common temperature compensation methods (such as optimal baseline selection [39] or optimal signal stretch [10]) for the training data. These have been applied to a similar deep-learning architecture in prior work with success [20]. We purposely do not apply these methods because (1) the DNN performance with and without temperature compensation is relatively comparable and (2) this choice allows us to directly measure localization uncertainty with respect to temperature. Assessing the uncertainty with temperature compensation is much more difficult as the temperature compensation process affects the uncertainty. The goal of this paper is to assess the effectiveness of deep uncertainty quantification for guided wave localization rather than demonstrate the best possible localization algorithm.

3.3. Deep Learning Architecture

Table 2 specifies the network architecture used for all three comparison deep learning methods. This architecture has been previously used in [20]. We use max-pooling operation (pooling factor = 2) to aggregate information across multiple representations after every convolutional layer. We increase the number of filters ($12 \rightarrow 24$) to allow the network to learn a consecutively higher level of features. All sensor pairs share filters with kernel size = 3, and every sensor pair is treated as a separate channel (i.e., the outputs are computed separately and summed together). After two conv-1D layers, we have a flattening layer followed by three dense, fully connected layers. At the output, we have a dense layer with two nodes (one node for each output space dimension) and a linear activation function. Note that the objective of this work is to compare the potential for reliability assessment of different DNN-based methods and hence we are not solely concerned with optimizing the DNN architecture for the best possible damage localization results.

For the DNN-MCdropout and DNN-Ensemble methods, the root mean squared error,

$$\ell(x, y) = \sqrt{(x - \hat{x})^2 + (y - \hat{y})^2}\,, \tag{7}$$

is minimized, where $\mathbf{d} = [x, y]$ is the true location and $\widehat{\mathbf{d}} = [\hat{x}, \hat{y}]$ is the predicted damage location. While for the DNN-GaussianMLE, the joint Gaussian likelihood is maximized.

For the DNN-Ensemble method, we train $N_{\text{ens}} = 10$ networks (increasing the ensemble size does not produce a significant reduction in localization error while significantly increasing training time). Each neural network in the ensemble is trained separately with a different random weight initialization, sampled from a zero-mean Gaussian distribution with $\sigma = 0.01$. We train each network in the ensemble for 40 epochs. For all the deep learning-based methods, we use the dropout technique (dropout probability = 0.25) to counter over-fitting during training.

Let $\widehat{f}(\mathbf{Z})$ be the damage location prediction of a deep learning model trained on guided wave data ($\mathbf{Z}$). The UQ metric is calculated as,

$$UQ\left(\widehat{f}(\mathbf{Z})|\mathbf{Z}\right) = \frac{\sigma_x + \sigma_y}{2}, \tag{8}$$

where σ_x and σ_y are the standard deviations of the model predictions in the x and y output dimensions. As explained in Sec. 3.1, the standard deviation is calculated from the ensemble and Monte-Carlo predictions for the DNN-Ensemble and DNN-MCdropout, respectively. The standard deviation is predicted directly as a network output with DNN-GaussianMLE. Note that the UQ metric (predictive uncertainty) can be obtained individually for every test sample and not as a summary statistic. A higher UQ metric value signifies lower localization reliability.

Table 2
Deep learning methods implementation

Variable	Description
Inputs dim:	$Q_t \times M$
$\mathbf{D}_{\text{tr}}$	4000 training data points
$\mathbf{D}_{\text{val}}$	1000 validation data points
Hidden layers	Conv layer 1: 12x filters
	Conv layer 2: 24x filters
	Dense layer 1: 600
	Dense layer 2: 400
	Dense layer 3: 40
	Dense layer 4: 2
Activation function [40]	Conv layer 1-2: Rectified Linear Unit
	Dense layer 1-3: Sigmoid
	Dense layer 4: Linear
Opt. algorithm	Adam [41]
Train-validation data split (%)	80/20
Training iterations	200 (40 for DNN-ensemble method)
Train batch size	16

3.4. Comparison Methods

Four methods for damage localization reliability are compared in this work. Out of these, three are deep learning-based methods. The fourth method, delay-and-sum (DAS) is often used as a baseline approach due to its widespread use within the guided waves community [42] and other related fields [43].

1. DNN ensemble
2. DNN Monte-Carlo dropout
3. DNN Gaussian likelihood optimization
4. Delay-and-sum Monte Carlo

The deep learning methods are explained in Sec. 3.1. For delay-and-sum (DAS) imaging, a narrowband wave model with constant group velocity is used. The data is first passed through a filter and enveloped with the Hilbert transform to envelope the time series data and mitigate the effects of dispersion on phase. For our setup, the group velocity c_g is set to 1947.9 m/s. The value is computed as the inverse slope of the A0 mode dispersion curve at the desired filtering frequency f_c (due to A0 mode being dominant at lower frequencies in the experimental data). The filtered and enveloped signal received at transducers and the windowed narrow-band signal are described below:

$$S(\omega, r)_{env} = H(\omega)e^{-j\omega r/c_g} \tag{9}$$

$$Z(\omega, r) = W(\omega) * S(\omega, r)_{env}, \tag{10}$$

where ω is the frequency, r is the distance traveled by the wave, and c_g is the group wave velocity. $H(\omega)$ is the transmitted signal, $S(\omega, r)_{env}$ is the signal received at transducers. $W(\omega)$ is the window used to remove boundary reflections, $*$ is the convolution operation, and $Z(\omega, r)$ is the windowed narrow-band signal.

Note that the DAS model considers an unbounded plate without dispersion, as that is typically how it is defined. In guided wave SHM, the DAS model further incoherently processes the data by only analyzing the envelope of the time signals[42, 44]. As the baseline model, we expect this to perform poorly. While the data could be coherently processed by removing the envelope and introducing dispersion information, it requires us to accurately learn the dispersion curves[45] and do so in near real-time to address variable environmental conditions. This is very difficult with a small number of sensors [20]. As a result, we do not consider such methods in this paper. Rather, we compare DAS with the deep learning models trained with the simulation data generated using the Lamb wave model of an unbounded plate with a known dispersion curve. Hence, as with DAS, dispersion curves are never explicitly estimated. Prior work has shown that this approach achieves higher performance than dispersion curve learning frameworks when the number of sensors is low [20].

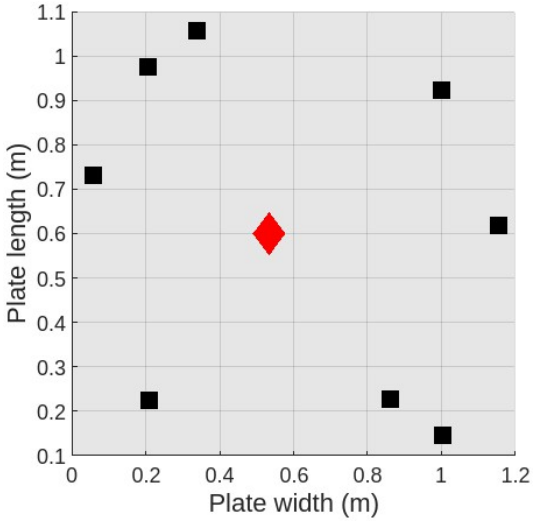

Figure 3: Experimental setup with 8 transducers(black squares) and damage location (red diamond).

The DAS forward model is computed for every possible grid point and compared with the data. This is mathematically represented as an ambiguity surface,

$$b_p = \frac{1}{QM}\left[\sum_{m=1}^{M}\sum_{q=1}^{Q} X(\omega_q, m) Z(\omega_q, r_m, p)\right]^2, \tag{11}$$

where b_p is the value of the ambiguity surface at pixel p, X is the experimental data, and $Z(\omega_q, r_m, p)$ is the model assuming a signal travels the distance corresponding to sensor pair m and the damage is at pixel p. The damage location estimate is given by

$$\hat{p} = \arg\max_p b_p \,, \tag{12}$$

i.e., the location of the ambiguity surface maxima.

To obtain an uncertainty estimate from DAS imaging, we perturb the group-wave velocity c_g randomly within a $\delta = 0.5\%$ range and generate $MC = 50$ Monte-Carlo DAS estimates. The δ value is chosen to approximately match the range of variations in the experimental data. This is equivalent to capturing parametric uncertainty. From the Monte Carlo estimates, we obtain the standard deviations (σ_x and σ_y). The UQ metric is then calculated as in (8).

4. Experimental setup

For testing the localization methods, 76 unique guided wave measurements are collected from an experimental setup with non-uniform temperature variations over $\approx$ 6 hours. The setup consists of an aluminum plate with square dimensions (1.2×1.2m). The aluminum material has an approximate density = 2700 kg/m^3 and Young's modulus of 69 GPa (gigapascal). PZT transducers are placed at random locations (array size = 8) to record guided wave measurements. Random transducer locations are chosen to avoid localization bias due to sensor placement. Each transducer transmits and receives signals in a round-robin manner and hence we have M = 56 total sensor pair signals.

The effect of structural damage is simulated by placing a mass at (0.5342, 0.6003 m) from the 37^{th} measurement onwards. The mass scatterer is a bronze cylinder of 1.5kg and 5cm in diameter. The mass is coupled to the plate with grease to allow the scattering of guided waves. While non-artificial damage would be preferable, the utilization of a mass scatterer is common within the guided waves community [46] to achieve experimental repeatability. Furthermore, its use should not significantly affect our assessment of uncertainty quantification methods in the presence of temperature variations. In this work, we only consider a single damage location. Multiple damage locations lead

Table 3
Experimental setup variations

Time (mins)	Record index	Heater	Damage
0 – 30	7	Low	Absent
30 – 61	14	Off	Absent
61 – 108	22	High	Absent
108 – 157	34	Off	Absent
157 – 166	36	Low	Absent
166 – 194	40	Low	Present
194 – 226	43	Off	Present
226 – 261	50	High	Present
261 – 291	58	Off	Present
291 – 321	65	Low	Present
321 – 381	76	Off	Present

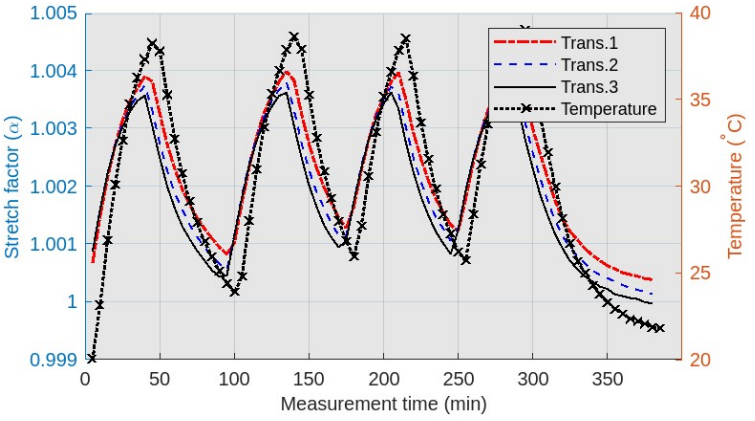

Figure 4: Stretch factor and temperature trends.

to an additional layer of uncertainty since some damage locations are more difficult to localize than others. Assessing location-specific uncertainty is sufficiently complex to merit its own paper and is a topic for future work.

A heating fan is placed inside a heat chamber beneath the plate specimen. The fan speed is toggled between 3 power levels periodically to create temporal temperature variations. The fan is kept pointed towards the top-left corner of the plate to create a spatial temperature gradient in addition to the temporal temperature variations. The temperature variations are measured at 5 distinct locations on the plate using a handheld laser IR thermometer. The heating fan is toggled 10 times creating 5 heating-cooling cycles. The heating fan variations are given in Table. (3). The temperature variations are noted to be between 4°C to 19°C above the room temperature of 20°C.

Every transducer transmits a 10V peak-to-peak chirp signal with a frequency sweep of 50 kHz to 500 kHz and a duration of 100 μs. Pulse compression is performed to remove dependence on the phase of guided wave signals and compress the measurement [47]. The transducers record signals for 4ms (4000 samples at a sampling rate of 1MHz). The first 40 μs are set to be zero to avoid cross-talk due to electromagnetic interference in the data collection setup. Finally, we pass the signal through a Gaussian filter. The Gaussian filter has a center frequency $f_c = 37.5$ kHz and bandwidth BW = 30 kHz. Note that although the filtered frequency is below the original chirp frequency range, there is significant energy at these frequencies due to side bands and weak attenuation at low frequencies. We use these frequencies for two reasons. First, they provide a relatively good resemblance between simulated and experimental guided wave signals. Second, the reflections from the mass are empirically strong at these low frequencies, improving our ability to locate the mass.

A deep learning model should be able to use the complete bandwidth. However, this requires a priori or learned knowledge of the dispersion curves, which includes the frequency-dependent wavenumbers, magnitudes, and phases for each mode). This knowledge is difficult to obtain, which is one reason DAS typically only considers narrowband,

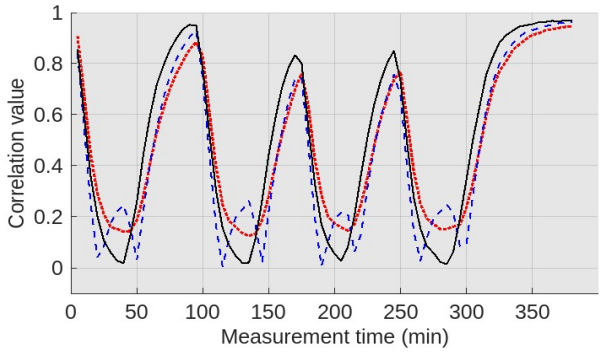

Figure 5: Correlation coefficient between signals recorded over the experiment duration and a reference signal (without temperature variations) corresponding to three distinct transducer pairs.

enveloped signals[42]. As previously mentioned, such information can be learned but requires a significant number of sensors. Our deep learning avoids the need to learn dispersion curves directly and instead aims to achieve invariance to the narrowband velocity.

The effect of temperature variations on guided waves is generally approximated as a stretch operation (i.e., higher temperature leads to a decrease in wave velocity, and hence the signal stretches in time). Note that the plate is unevenly heated and therefore produces a surface temperature gradient, and asynchronous temperature measurements were only taken during a single heating session. Therefore, we cannot provide a full description of the surface temperature. However, we can approximate the temperature at one point based on the minimum and maximum temperatures and the knowledge that the scale factor is linearly correlated to temperature [10]. This approximation is shown in Fig. 4 for a point near the heating fan. The figure shows the variations in stretch factor values (left Y axis) computed between three transducer pair signals and a reference signal (at ambient temperature) as well as the approximate temperature trends (curve with cross marks - right Y axis) over the entire experiment duration. We see the repeating pattern in stretch factor values indicating the heating fan toggling. Similarly inFig. 5, the correlation coefficient between signals recorded over the experiment duration and a reference signal (without temperature variations) corresponding to three distinct transducer pairs is shown. Any significant deviation from the ambient temperature leads to a significant change in the recorded guided wave signal. In the experiment, the heating fan is periodically toggled (refer to Table 3) leading to periodic temperature variations. This shows up as an apparent inversion in the correlation coefficient trends throughout the experiment duration.

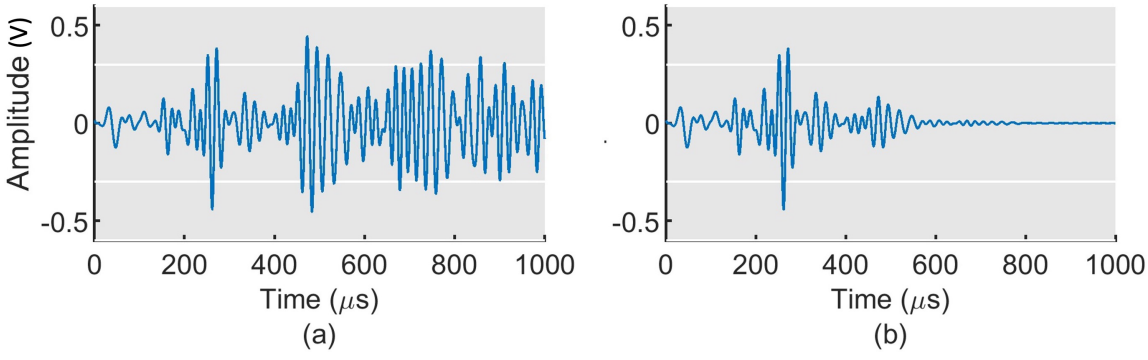

Figure 6: Post baseline subtraction after the filtering operation ($f_c = 37.5$ KHz)

: (a) Before velocity window, (b) after velocity window.

Baseline subtraction is often imperfect due to temperature variations. Hence, there are multiple strategies to compensate for temperature variations, such as optimal baseline subtraction (OBS) or baseline signal stretch (BSS). OBS chooses from a set of baseline measurements while BSS stretches the baseline by a factor to approximate the effect of temperature. We use a strategy that combines optimal baseline selection and baseline signal stretch using the

scale transform [10]. The scale transform, a variant of the Fourier transform has been used before to approximate the effect of temperature variations on guided waves. Two signals recorded at different temperatures can be denoted as $x(t)$ and $s(t) = x(\alpha t)$ where α is the stretch factor obtained using the scale transform. For implementation details, we refer the reader to [48]. The baseline subtraction strategy is described below:
1.) We form a template bank **T** by selecting a measurement each from the damaged set (template signal: $\overline{t_1}$) and undamaged set of experimental signals (global baseline: $\overline{t_0}$). This can be considered as an extension of the validation dataset.
2.) For each test measurement (x), the template bank signal that minimizes residual energy is chosen (t^*).
3.) Next, the stretch factor α_m between each transducer pair signal of the measurement ($x(m)$) and the corresponding transducer pair signal of the chosen template bank signal ($t^*(m)$) is calculated.
4.) Each transducer pair signal of the measurement is stretched using the scale transform $x^{\text{str}}(m)$[10] by α_m (calculated in Step 3) to best match the chosen template signal, (t^*).
5.) The global baseline signal ($\overline{t_0}$) is subtracted from the stretched signal to obtain the baseline subtracted signal.

Note that the temperature compensation strategies at the baseline stage still do not help in quantifying the effect of external variations on the localization performance. The Lamb wave model described in (5) is for an unbounded plate. In an experimental setup, there are multiple unmodeled boundary reflections. We mitigate the effect of these reflections by applying an exponentially tapering window to the guided wave signal. The velocity window is defined with a chosen velocity value (v_{win}) and an attenuation constant (a). The velocity window is a rectangular window that tapers exponentially with the chosen attenuation constant (a) after the arrival of a hypothetical signal with the assumed window velocity (v_{win}). In this work, we set $\text{v}_{\text{win}} = 1500$ m/s and an exponential decay factor of $\text{a} = 100$. The window velocity value is chosen based on the assumed group velocity of 1947 m/s for the experimental data.Fig. 6 shows the baseline subtracted and filtered signal before and after the application of the tapering window.

5. Results and discussion

In this section, we discuss the results of our four different uncertainty quantification methods using both simulation data and experimental data. We first demonstrate results with simulation data to establish relationships between uncertainty and velocity variations and then validate this with experimental data.

5.1. Simulation validity

The reliability of the localization methods is first evaluated with simulated data (Fig. 7). The temperature variations affect the wave group velocity (v_g), which can be modeled as a multiplicative perturbation (α) to the wavenumber (κ). To simulate this variation, a dataset with random damage locations within the plate dimensions and a repeating pattern of the wavenumber perturbation factor (α) is generated ($0.99 - 1.00 - 1.01$), see the red curve in Fig. 7. This trend simulates a repeating temperature variation (low: 0.99, no change: 1.0, and high: 1.01). With DNN-Ensemble and DNN-MCDropout, (Fig. 7(a-b)), when $\alpha = 1$, there are low uncertainties while deviations from this α value produce high uncertainties. But, DNN-Ensemble has a much bigger range of values (0.01 to 0.14) as compared to DNN-MCDropout (0.08 to 0.02). For DNN-Gaussian MLE and DAS-MC (Fig. 7(c-d)), there are a lot of fluctuations with no apparent trends.

Variability in the results shown in Fig. 7 comes from 2 sources apart from the wavenumber perturbations: random damage locations, and the uncertainty inherent in the methods themselves. Since DNN-Ensemble is a collection of deterministic models, the only variability in this case is due to the random damage locations. Whereas, with DAS, the method is run 50 times with perturbed group wave velocity values (discussed in Sec. 3.4). The standard deviation of the Monte-Carlo location estimates is used to calculate uncertainty. Similarly, DNN-MCDropout and DNN-Gaussian MLE rely on randomly dropped-out nodes and probabilistic loss functions to provide a UQ metric. Hence, DAS-MC, DNN-Gaussian MLE, and MCDropout have an added level of uncertainty. This explanation is validated by the approximately similar UQ metric values for the same perturbation levels with DNN-Ensemble (Fig. 7(a)) as compared to the other algorithms (Fig. 7(b-d)). DNN-Ensemble provides the most intuitive UQ metric correlating with the wavenumber perturbations as compared to the other methods. Note that, uncertainty due to random damage locations is an added level of complexity observed for all methods and is not discussed in this work.

While different methods provide different levels of uncertainty, lower levels do not necessarily imply better results. Ideally, the UQ metric should reflect the true uncertainty in the data. However, interpreting UQ metrics from the deep

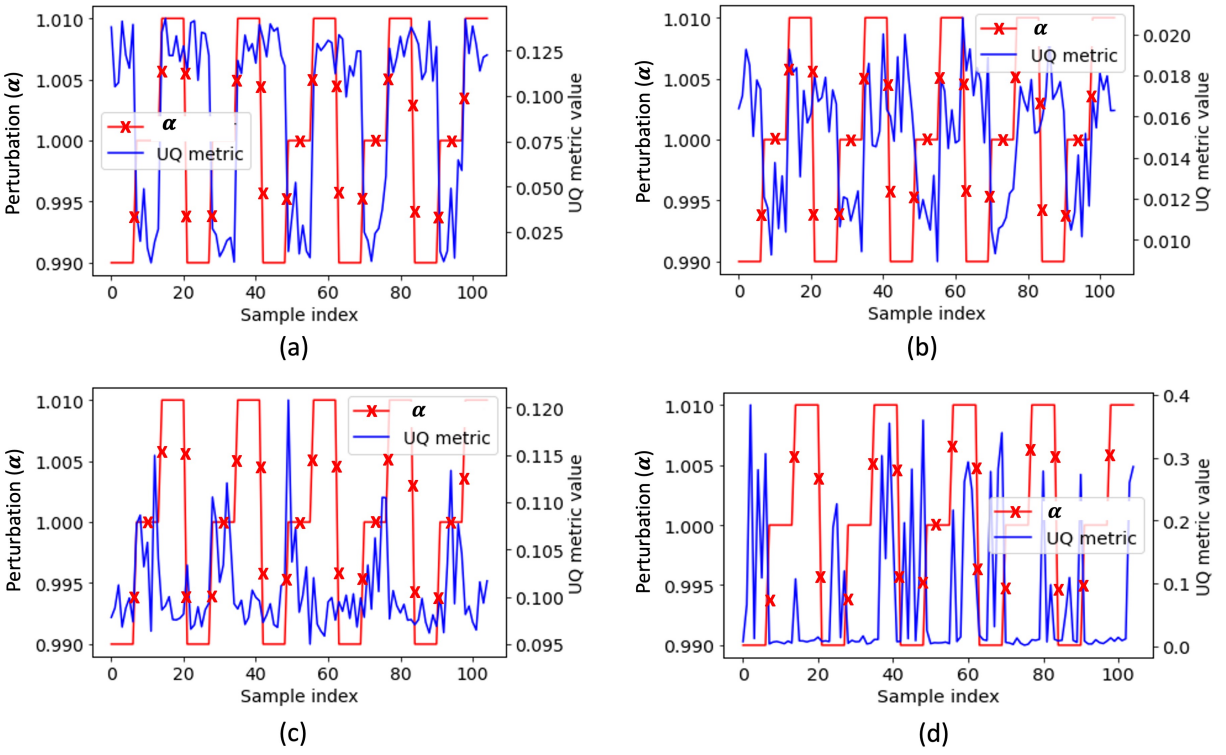

Figure 7: Localization reliability assessment for data with a simulated wavenumber perturbation for (a) DNN-Ensemble, (b) DNN-MCDropout, (c) DNN-Gaussian MLE, and (d) delay-and-sum. UQ metric is in meters since it describes the localization uncertainty.

learning strategies is a topic of considerable research [49]. We will discuss some interpretations in the next subsection when comparing the localization error with the UQ metric.

5.2. Experimental validity

To quantify the localization reliability, metrics are compared for all the methods on the experimental data. The experimental data has 36 measurements without damage followed by 41 measurements with damage (i.e., $T_{nd} = 36;\ T_d = 41$). The evaluation metrics include:

Average loc. error (m) ↓: Localization error averaged over all test samples with damage:

$$\text{ALE}(m) = \frac{1}{T_d}\sum_{i=1}^{T_d}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{13}$$

Average UQ metric (m): UQ metric value (as in (8)) averaged over all test samples with damage:

$$\text{UQ}_{avg} = \frac{1}{T_d}\sum_{i=1}^{T_d}\text{UQ}_i \tag{14}$$

Δ(UQ) value ↑: The percent difference between UQ metrics averaged over samples with and without damage:

$$\Delta(\text{UQ}) = \frac{1}{\text{UQ}_{avg}}\left[\left(\frac{1}{T_{nd}}\sum_{i=1}^{T_{nd}}\text{UQ}_i\right) - \text{UQ}_{avg}\right] \tag{15}$$

Table 4
Localization performance and reliability analysis for all comparison methods

Method	Loc. error (m) ↓	UQ_{avg} (m)	Δ(UQ) value ↑	Acc. ↑
DNN-Ensemble	0.092	0.142	39.4%	0.947
DNN-MCDropout	0.113	0.017	41.2%	0.880
DNN-Gaussian MLE	0.160	0.103	7.2%	0.947
DAS-MC	0.448	0.070	8.7%	0.520

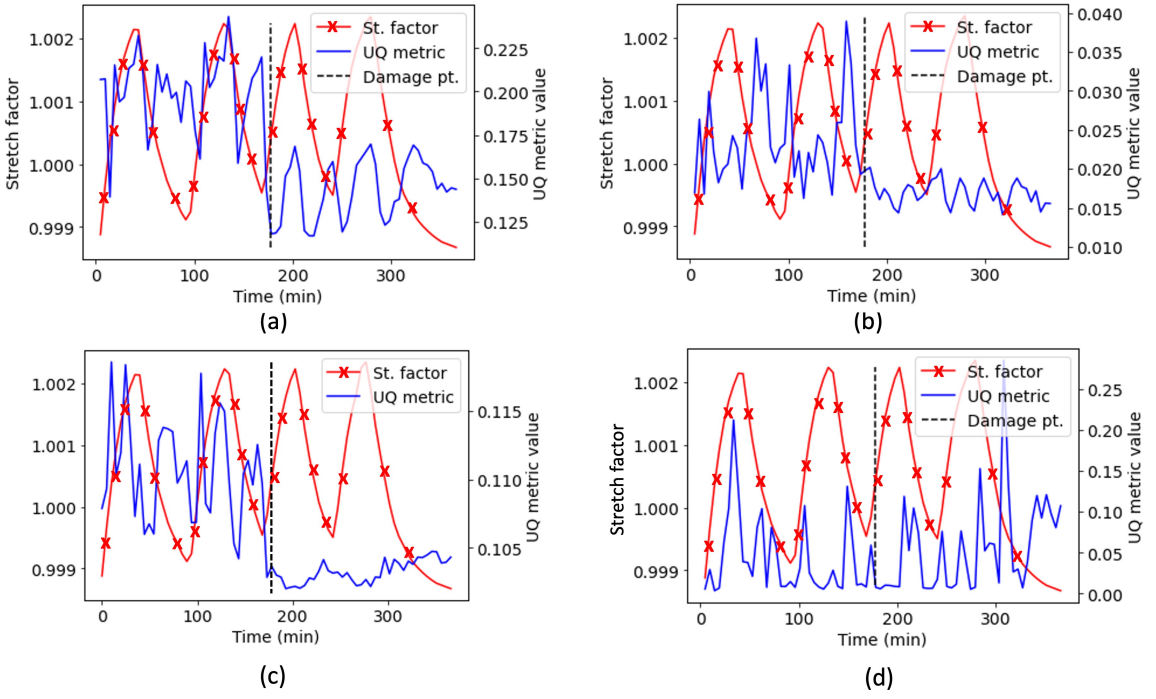

Figure 8: Localization reliability trends with experimental data for (a). DNN-Ensemble, (b). DNN-MCDropout, (c) DNN-Gaussian MLE, and (d) delay-and-sum Monte Carlo. The UQ metric has units of meters since it describes the uncertainty in localization.

Acc. ↑: The accuracy (the number of true positive plus true negative classifications divided by the total number of measurements) of detecting damage, given a UQ metric threshold that optimally separates the damage and no damage data. That is, if the UQ metric falls below this threshold, we identify damage. If the UQ metric falls above this threshold, we identify no damage.

The overall metrics from our experimental results are shown in Table 4. We discuss these metrics in depth in the following subsections.

5.2.1. Uncertainty over time

The methods are evaluated with the experimental data having repeating variations in external temperature (approximated by stretch factor), see the red curve in Fig. 8. Note that the stretch factor values fall in the range of (0.998 to 1.003), which correspond to temperature variations of 4°C to 19°C.

We also analyze the change in UQ metric before and after the damage is introduced. Ideally in the absence of any damage, a localization method should have a very high UQ metric as there is no damage to localize. Yet, due to the approximate nature of temperature compensation strategies, a residual signal remains, which produces erroneous localization results even when no damage is actually present. The UQ metric is expected to drop when the damage is

introduced (time ≈ 180 min denoted by the dashed black line in Fig. 8). This trend is observed with the deep learning-based methods, see Fig. 8(a-c). In contrast, the UQ metric based on delay-and-sum imaging shows no discernible change, see Fig. 8(d). Out of all the methods, DNN-Ensemble and DNN-MCDropout have the highest drop in UQ metric ($\Delta(UQ)$) after damage introduction (39.4% and 41.2%, respectively), whereas the DNN-Gaussian MLE and DAS-MC have significantly lower $\Delta(UQ)$ values (7.2% and 8.7%) respectively. Note that while both DNN-Ensemble and DNN-MCDropout have similar relative drops, the DNN-Ensemble UQ metric has a better absolute separation between the two classes.

Further, out of all methods, DNN-Ensemble shows sensitivity to the temperature variations after the damage is introduced, see Fig. 8(a). The UQ metric variations are correlated with the stretch factor variations. When the stretch factor deviates from a value of 1, the UQ metric rises and drops when the stretch factor is close to 1. This is similar to the trends we observed with the wavenumber perturbation simulation results in Fig. 7(a). These results demonstrate that the DNN-Ensemble method is the most capable of representing the effect of temperature variations on the localization performance for both simulated and experimental data.

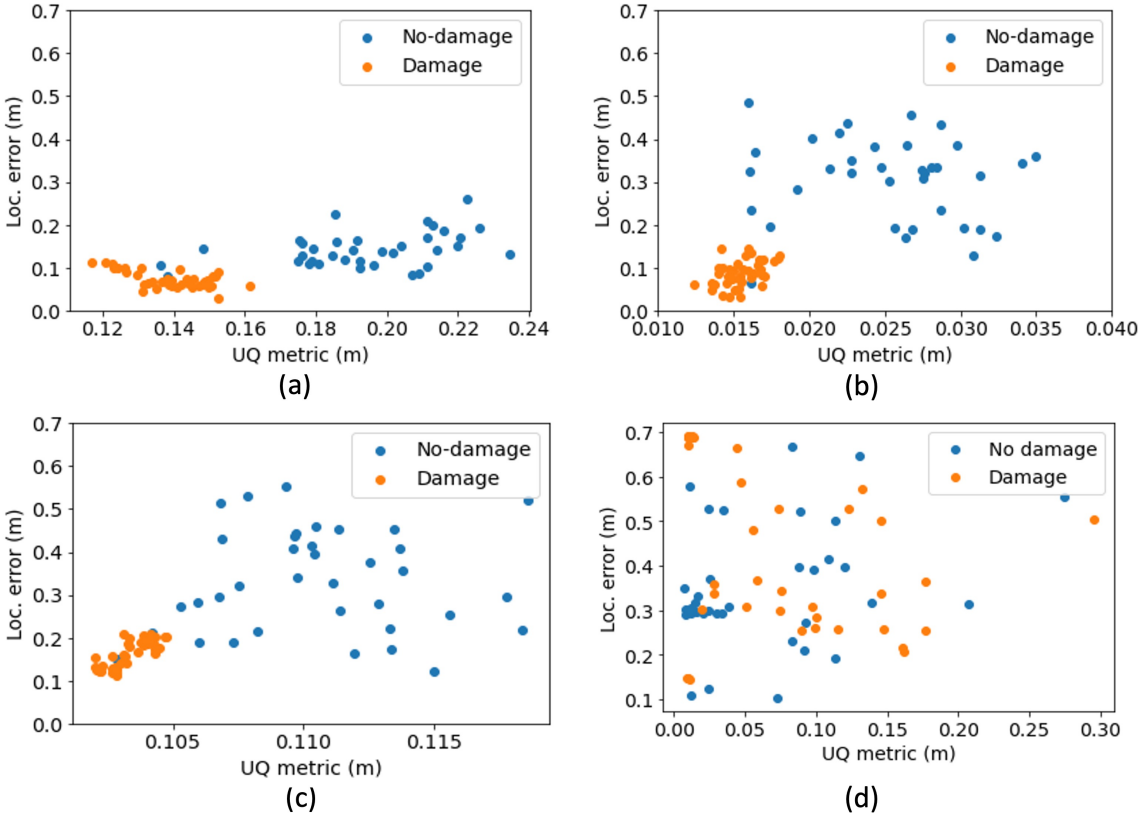

Figure 9: Reliability metric variation with localization error for (a). DNN-Ensemble, (b). DNN-MCDropout, (c) DNN-Gaussian MLE, and (d) delay-and-sum Monte Carlo.

5.2.2. Uncertainty versus localization error

In Fig. 9, the individual values of UQ metric are plotted versus the corresponding localization error values as a scatter plot. Note that when there is no damage, we still compute the error as if the damage is located at the same location as this provides some interesting relationships within the data. However, it is important to note that the localization error should have no physical meaning when no damage is present.

The main point of comparison is the ability of each method to distinguish between no-damage and damage state of the plate solely based on the UQ metric value. To that end, DNN-Ensemble shows the most distinct separation between the damage states. Both DNN-Ensemble and DNN-Gaussian MLE have a damage detection accuracy of

94.7% followed by DNN-MCDropout with an accuracy of 88%. DAS-MC has the lowest detection accuracy out of all due to its sensitivity to external temperature variations. This indicates that out of the deep learning models considered, DNN-Ensemble is a more appropriate choice to distinguish between the damage and no-damage state using the UQ metric as a detection statistic.

All three deep learning methods provide relatively reasonable localization (errors under 0.2m) when applied to data with damage, whereas delay-and-sum Monte-Carlo has a high variance in localization errors (range of 0.2 to 0.7m) as a result of the simulated perturbations in the group wave velocity (c_g) added to measure its uncertainty. Even with the choice of group velocity to match the narrow band spectrum under consideration, the high localization error of 0.7m is a result of using delay-and-sum (DAS) without temperature compensation, which is well known to perform extremely poorly[50]. As mentioned earlier, we still include DAS as a baseline approach due to its widespread use within the guided waves community [42] and other related fields [43].

For each of the DNN methods, the localization error range (errors under 0.2m) is approximately equivalent to the A0 guided wave pulse width ($\approx 100\ \mu$s) translated into space (100 μs times $v_g \approx 1947.9$ m/s is 0.195 m). Hence, the deep neural networks appear to have similar resolution limitations as delay-and-sum imaging (with correct velocity calibration). Note that UQ metric for DNN-Ensemble also has approximately the same range as the localization error when damage is present, indicating a relationship between the UQ metric and localization error.

When no damage is present, the results from DNN-MCDropout and DNN-Gaussian MLE visually show a correlation between the error and the UQ metric (i.e., the points with high localization error also tend to have low uncertainty). However, as previously stated, the localization error has no physical meaning when there is no damage present. Instead, the correlation occurs due to a correlation between uncertainty and the distance from the center of the plate, (i.e., the average value of the labels in the training data). Hence, the DNN-MCDropout and DNN-Gaussian MLE UQ metrics are biased by the training labels. When no damage is present, DNN-Ensemble's UQ metric shows no clear bias, but its localization error is instead biased since the estimate is an average of effectively random guesses, which tends toward predicting the damage at the center of the plate.

5.3. Reliability comparison

An ideal localization method should have a low localization error, physically relevant UQ metric, a large Δ(UQ), and complete separation between the damage and no damage in the UQ metric (indicating that damage can be detected accurately with the UQ metric). These values are all reported together in Table. 4. Overall, DNN-Ensemble satisfies these requirements the best.

The localization error with DNN-Ensemble is the lowest out of all of the methods. While the average UQ metric is higher for DNN-Ensemble than other methods, its overall range corresponds most closely with the method's localization error. Finally, it has the second highest Δ(UQ) value and the highest damage detection accuracy, indicating the best separation in the UQ metric. Every other method has a flaw that prevents the UQ method from relating to the reliability of our localization method.

While DNN-Ensemble's prediction is biased when there is no damage, this is not measurable at the time of the test. The uncertainty bias found in DNN-MCDroupout and DNN-Gaussian MLE is more likely to lead to incorrect decisions. In addition, the UQ metrics from DNN-MCDropout and DNN-MLE do not occur within a physically justifiable range. Finally, The UQ metric from delay-and-sum Monte Carlo shows no separation between the damage and no damage scenarios.

6. Conclusion

In this work, we have motivated the need to assess the reliability of guided wave damage localization systems. There has been prior work in reliability assessment for NDE systems. Yet, the transfer of those methods to SHM is not straightforward. There are unique challenges in developing scalable reliability assessment schemes for SHM damage localization. We connect the fields of reliability assessment and uncertainty quantification and propose a UQ metric for reliability assessment with localization methods.

The damage localization reliability is analyzed with both simulated and experimental data with non-uniform temperature variations. A simple UQ metric is computed to quantify localization reliability corresponding an individual guided wave measurement as opposed to reported average statistics over an entire set of measurements. We define multiple metrics to quantify the damage localization reliability. These metrics can compare the reliability of different methods both with and without the knowledge of test-time damage location.

The DNN-Ensemble model provides better damage localization reliability assessment compared to other deep learning methods. This is demonstrated in terms of sensitivity to temperature variations and the ability to distinguish between damage states. Other deep-learning methods show much lesser sensitivity to temperature variations. This indicates that the DNN-Ensemble method can represent the uncertainty in the model predictions due to the external variations much better than other deep learning methods. Finally, DNN-Ensemble also shows better localization performance than other deep learning methods. Within deep learning models, DNN-Ensemble supported by prediction reliability metrics is a better choice for reliable damage localization. A limitation of this work is the assumption of the simplistic guided wave propagation model. In the future, this work can be extended to composite structures improving the real-world applicability.

7. Acknowledgement

This research is supported by the National Science Foundation under award numbers EECS-1839704 and NSF CISE-1747783.

8. Author contributions

Ishan D. Khurjekar: Conceptualization, Methodology, Analysis, Writing: **Joel B. Harley**: Analysis, Writing, Resources, Funding acquisition, Supervision.

References

[1] Sandeep Kumar Dwivedi, Manish Vishwakarma, and Akhilesh Soni. Advances and researches on non destructive testing: A review. *Materials Today: Proceedings*, 5(2):3690–3698, 2018.

[2] 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.

[3] Mira Mitra and S Gopalakrishnan. Guided wave based structural health monitoring: A review. *Smart Materials and Structures*, 25(5):053001, 2016.

[4] Stefano Mariani, Sebastian Heinlein, and Peter Cawley. Compensation for temperature-dependent phase and velocity of guided wave signals in baseline subtraction for structural health monitoring. *Structural Health Monitoring*, 19(1):26–47, 2020.

[5] Xin Chen, Jennifer E Michaels, Sang Jun Lee, and Thomas E Michaels. Load-differential imaging for detection and localization of fatigue cracks using lamb waves. *NDT & E International*, 51:142–149, 2012.

[6] G. Konstantinidis, P. D. Wilcox, and B. W. Drinkwater. An investigation into the temperature stability of a guided wave structural health monitoring system using permanently attached sensors. *IEEE Sensors Journal*, 7(5):905–912, 2007.

[7] Jochen Moll, Christian Kexel, Serena Pötzsch, Marcel Rennoch, and Axel S Herrmann. Temperature affected guided wave propagation in a composite plate complementing the open guided waves platform. *Scientific data*, 6(1):1–9, 2019.

[8] Claude Fendzi, Marc Rebillat, Nazih Mechbal, Mikhail Guskov, and Gérard Coffignal. A data-driven temperature compensation approach for structural health monitoring using lamb waves. *Structural Health Monitoring*, 15(5):525–540, 2016.

[9] Anthony J Croxford, Paul D Wilcox, George Konstantinidis, and Bruce W Drinkwater. Strategies for overcoming the effect of temperature on guided wave structural health monitoring. In *Health Monitoring of Structural and Biological Systems 2007*, volume 6532, pages 590–599. Proc. of SPIE, 2007.

[10] Joel B Harley and José MF Moura. Scale transform signal processing for optimal ultrasonic temperature compensation. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 59(10):2226–2236, 2012.

[11] Kailash C Kapur and Michael Pecht. *Reliability engineering*, volume 86. John Wiley & Sons, 2014.

[12] George A Georgiou. PoD curves, their derivation, applications and limitations. *Insight-Non-Destructive Testing and Condition Monitoring*, 49(7):409–414, 2007.

[13] Francesco Falcetelli, Nan Yue, Raffaella Di Sante, and Dimitrios Zarouchas. Probability of detection, localization, and sizing: The evolution of reliability metrics in structural health monitoring. *Structural Health Monitoring*, 21(6):2990–3017, 2022.

[14] Inka Mueller, Vittorio Memmolo, Kilian Tschöke, Maria Moix-Bonet, Kathrin Möllenhoff, Mikhail Golub, Ramanan Sridaran Venkat, Yevgeniya Lugovtsova, Artem Eremin, and Jochen Moll. Performance assessment for a guided wave-based shm system applied to a stiffened composite structure. *Sensors*, 22(19):7529, 2022.

[15] Eric B Flynn, Michael D Todd, Paul D Wilcox, Bruce W Drinkwater, and Anthony J Croxford. Maximum-likelihood estimation of damage location in guided-wave structural health monitoring. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 467(2133):2575–2596, 2011.

[16] Jérémy Moriot, Nicolas Quaegebeur, Alain Le Duff, and Patrice Masson. A model-based approach for statistical assessment of detection and localization performance of guided wave–based imaging techniques. *Structural Health Monitoring*, 17(6):1460–1472, 2018.

[17] Roberto Miorelli, Clément Fisher, Andrii Kulakovskyi, Bastien Chapuis, Olivier Mesnil, and Oscar D'Almeida. Defect sizing in guided wave imaging structural health monitoring using convolutional neural networks. *NDT & E International*, 122:102480, September 2021.

[18] Sergio Cantero-Chinchilla, Paul D Wilcox, and Anthony J Croxford. Deep learning in automated ultrasonic NDE – developments, axioms and opportunities. *NDT & E International*, 131:102703, October 2022.

[19] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proc. of International Conference on Machine Learning Workshop on Unsupervised and Transfer Learning*, pages 17–36, 2012.

[20] Ishan D Khurjekar and Joel B Harley. Sim-to-real localization: Environment resilient deep ensemble learning for guided wave damage localization. *The Journal of the Acoustical Society of America*, 151(2):1325–1336, 2022.

[21] Jeremy S Knopp, John C Aldrin, E Lindgren, and Charles Annis. Investigation of a model-assisted approach to probability of detection evaluation. In *AIP conference proceedings*, volume 894, pages 1775–1782. American Institute of Physics, 2007.

[22] Department of Defense, MIL-HDBK. Department of defense handbook: Nondestructive evaluation system reliability assessment, 2009.

[23] Christine M Schubert Kabban, Brandon M Greenwell, Martin P DeSimio, and Mark M Derriso. The probability of detection for structural health monitoring systems: Repeated measures data. *Structural Health Monitoring*, 14(3):252–264, 2015.

[24] John A Swets. Measuring the accuracy of diagnostic systems. *Science*, 240(4857):1285–1293, 1988.

[25] Ramin Ghiasi, Peyman Torkzadeh, and Mohammad Noori. A machine-learning approach for structural damage detection using least square support vector machine based on a new combinational kernel function. *Structural Health Monitoring*, 15(3):302–316, 2016.

[26] Young-Jin Cha and Zilong Wang. Unsupervised novelty detection–based structural damage localization using a density peaks-based fast clustering algorithm. *Structural Health Monitoring*, 17(2):313–324, 2018.

[27] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[28] Viet Tra, Sheraz Ali Khan, and Jong-Myon Kim. Diagnosis of bearing defects under variable speed conditions using energy distribution maps of acoustic emission spectra and convolutional neural networks. *The Journal of the Acoustical Society of America*, 144(4):EL322–EL327, 2018.

[29] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[30] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[31] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. of the International Conference on Machine Learning*, pages 1050–1059, 2016.

[32] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proc. of the Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[33] Richard J Pyle, Robert R Hughes, Amine Ait Si Ali, and Paul D Wilcox. Uncertainty quantification for deep learning in ultrasonic crack characterization. *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, 69(7):2339–2351, July 2022.

[34] Homin Song and Yongchao Yang. Uncertainty quantification in super-resolution guided wave array imaging using a variational bayesian deep learning approach. *NDT & E International*, 133:102753, 2023.

[35] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[36] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30:5580–5590, 2017.

[37] Pawel Packo, Tadeusz Uhl, and Wieslaw J Staszewski. Generalized semi-analytical finite difference method for dispersion curves calculation and numerical dispersion analysis for lamb waves. *The Journal of the Acoustical Society of America*, 136(3):993–1002, 2014.

[38] James S Hall and Jennifer E Michaels. Multipath ultrasonic guided wave imaging in complex structures. *Structural Health Monitoring*, 14 (4):345–358, 2015.

[39] Anthony J Croxford, Jochen Moll, Paul D Wilcox, and Jennifer E Michaels. Efficient temperature compensation strategies for guided wave structural health monitoring. *Ultrasonics*, 50(4-5):517–528, 2010.

[40] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[42] Jennifer E Michaels. Detection, localization and characterization of damage in plates with an in situ array of spatially distributed ultrasonic sensors. *Smart Materials and Structures*, 17(3):035035, 2008.

[43] Giulia Matrone, Alessandro Stuart Savoia, Giosuè Caliano, and Giovanni Magenes. The delay multiply and sum beamforming algorithm in ultrasound b-mode medical imaging. *IEEE transactions on medical imaging*, 34(4):940–949, 2014.

[44] Thomas Clarke, Peter Cawley, Paul David Wilcox, and Anthony John Croxford. Evaluation of the damage detection capability of a sparse-array guided-wave shm system applied to a complex structure under varying thermal conditions. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 56(12):2666–2678, 2009.

[45] Takahiro Hayashi, Chiga Tamayama, and Morimasa Murase. Wave structure analysis of guided waves in a bar with an arbitrary cross-section. *Ultrasonics*, 44(1):17–24, 2006.

[46] Chang Liu, Joel Harley, Nicholas O'Donoughue, Yujie Ying, Martin H Altschul, Mario Bergés, James H Garrett, David W Greve, José MF Moura, Irving J Oppenheim, et al. Robust change detection in highly dynamic guided wave signals with singular value decomposition. In *2012 IEEE International Ultrasonics Symposium*, pages 483–486. IEEE, 2012.

[47] Jennifer E Michaels, Sang Jun Lee, Anthony J Croxford, and Paul D Wilcox. Chirp excitation of ultrasonic guided waves. *Ultrasonics*, 53(1): 265–270, January 2013.

[48] Joel B Harley, Chang Liu, Irving J Oppenheim, and José MF Moura. Managing complexity, uncertainty, and variability in guided wave structural health monitoring. *SICE Journal of Control, Measurement, and System Integration*, 10(5):325–336, 2017.

[49] Lucas Kook, Andrea Götschi, Philipp FM Baumann, Torsten Hothorn, and Beate Sick. Deep interpretable ensembles. *arXiv preprint arXiv:2205.12729*, 2022.

[50] Alexander C. S. Douglass and Joel B. Harley. Dynamic time warping temperature compensation for guided wave structural health monitoring. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 65(5):851–861, 2018. doi: 10.1109/TUFFC.2018.2813278.