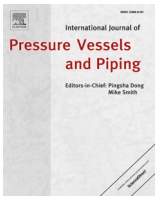
International Journal of Pressure Vessels and Piping 207 (2024) 105108

Contents lists available at ScienceDirect

International Journal of Pressure Vessels and Piping

journal homepage: www.elsevier.com/locate/ijpvp



Calibration, validation, and selection of hydrostatic testing-based remaining useful life prediction models for polyethylene pipes

Dongjin Du, Pranav Karve, Sankaran Mahadevan *

Civil and Environmental Engineering, Vanderbilt University, Nashville, TN, USA

A B S T R A C T

A novel methodology for model selection among competing models for remaining useful life (RUL) prediction is developed in this paper. Due to the long durability of polyethylene (PE) pipes under normal conditions, life data under normal operating conditions is not available for model validation and selection. Physics-based, regression-based, and hybrid models for RUL prediction are available in the literature based on experimental data under accelerated test conditions. These models need to be evaluated for prediction performance under normal conditions, but in the absence of data under normal conditions. A model consistency-based metric for selecting the best model in the absence of data under normal conditions is proposed in this paper. The consistency-based metric evaluates the predictive consistency of the various probabilistic RUL prediction models over the estimated or known probability distribution of the normal operating conditions. A two-step approach for model selection is proposed: first, validation evaluation using empirical data, and second, consistency evaluation under likely operating conditions. The proposed model selection methodology is demonstrated using five candidate models for RUL prediction of PE pipes based on accelerated hydrostatic testing data. Bayesian inference is used to calibrate these models with empirical data, and its benefit over the least squares approach recommended in ISO 9080 is demonstrated. Further, the proposed two-step model selection methodology is compared against traditional model selection methods based on goodness of fit, model complexity and information theoretic metrics. It is seen that the proposed additional consistency criterion is successful in selecting the best model compared to existing methods that are unable to distinguish between the different available models.

1. Introduction

Polyethylene (PE) pipes are widely used in the U.S., with more than 70,000 miles of PE pipes estimated to be in service at present [1]. These pipes are used for many applications such as water distribution, gas distribution, and protecting internet cables. PE pipes degrade due to various physical and chemical phenomena during their service life. Physical degradation mechanisms involve creep, relaxation, secondary crystallization, and molecular rearrangements that adversely affect the mechanical properties of the pipe. Chemical degradation processes involve molecular degradation in the form of chain scission and cross-linking in the polymer material due to external agents (water, ultraviolet radiation, chemical pollutants). This often leads to brittle failure of PE pipes. Both physical and chemical aging can be accelerated by elevated temperature. See Ref. [2] for a detailed discussion on PE pipe aging mechanisms.

Accounting for the degradation processes and predicting the remaining useful life (RUL) is essential for the safe and reliable deployment and timely replacement of PE pipes. The commonly used PE pipe RUL prediction method is based on hydrostatic testing data collected under high temperature and pressure conditions [3]. The test procedure as well as the RUL prediction method (based on extrapolation to normal operating conditions) has been standardized by the International Standards Organization (ISO) [4] and the American Society for Testing and Materials (ASTM) [5]. Testing and modeling related to low crack growth (SCG) and environmental stress cracking (ESC) have also been performed for RUL prediction of plastic pipes [6–8], but these are outside the scope of this paper. In the hydrostatic testing-based method, the form of the RUL prediction model is derived using the Arrhenius equation and the known (log-log) relationship between pipe hoop stress and failure time, and RUL prediction models with different forms and different number of model parameters could be constructed. The RUL model building procedure then involves a) conducting hydrostatic pipe failure tests under high temperature/stress conditions (accelerated failure tests), b) using the accelerated test data to calibrate the model parameters, c) validation of the calibrated model, and d) using the validated model for RUL prediction under normal operating conditions (extrapolation). In this article, we discuss methods to improve calibration and validation of (probabilistic) RUL prediction models. We also propose a novel RUL model selection methodology based on consistency in extrapolation to normal conditions.

The RUL prediction models are typically calibrated using accelerated

* Corresponding author.
E-mail address: sankaran.mahadevan@vanderbilt.edu (S. Mahadevan).

https://doi.org/10.1016/j.ijpvp.2023.105108
Received 22 September 2023; Received in revised form 21 November 2023; Accepted 5 December 2023
Available online 10 December 2023

hydrostatic test data and least-squares regression. This model calibration method provides deterministic model parameter values that minimize the sum of squares of the error between the accelerated test data and corresponding model predictions. It also provides standard deviation of the residual, which is assumed to be a zero-mean Gaussian random variable. Often the residual is neglected to obtain deterministic RUL prediction. However, even if the residual is considered (as recommended in ISO 9080 [4]), the assumption regarding normality and constant variance is difficult to justify for probabilistic RUL prediction. Bayesian inference-based model calibration, on the other hand, does not involve making these restrictive assumptions. In the Bayesian approach, the state of knowledge about the values of unknown parameters of interest is represented using prior and posterior probability distribution functions. The updated knowledge about a parameter (represented by the posterior distribution) is obtained by combining prior knowledge (based on intuition, experience, model prediction, prior data, etc.) and observations (test data). The observations are included in the Bayesian inference by computing the likelihood of observing the data for a given value of the parameter. In this work, we pursue the Bayesian inference approach for calibrating (probabilistic) RUL prediction models and evaluate its benefits over the least-squares approach recommended in ISO 9080.

When multiple calibrated models are available, model validation and selection are critical for ensuring accurate predictions of the quantity of interest. Various engineering standards on testing of PE pipes (ISO [4], ASTM [5]) have correctly recognized the limitation of extrapolating the model to conditions different than those used for laboratory testing and calibration. They have recommended *validation exercises*, which rely on field data to qualitatively establish that extrapolation does not incur large errors. This is different from the quantitative model validation methods defined in classical verification and validation (V&V) literature [9–11]. The latter methods aim to quantify the degree to which a predictive model is an accurate representation of the real world from the perspective of the intended use of the model [11]. For deterministic predictive models, simple metrics that quantify the error between the predicted and measured values (e.g., mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE) etc.) could be used for ascertaining the model's predictive accuracy.

In general, predictive models can be classified into two categories: physics-based models (i.e., based on first principles); and regression-based models (based on fitting empirical data). For these two types of models, accuracy criteria (for both types of models) and additional model complexity criteria (for regression-based models) are typically employed for model evaluation and selection. While regression-based models containing a large number of parameters could improve predictive performance because of the added complexity due to the higher number of model parameters, they also degrade the model performance for conditions different from the training data. Therefore, classical model selection methodologies for regression models consider the trade-off between model complexity (number of model parameters to be estimated) and goodness of fit [12,13]. These methods include the Akaike information criterion (AIC) [14], Bayesian information criterion (BIC) [15], Minimum description length (MDL) [16], etc. For quantitative validation of physics-based, probabilistic predictive models, classical and Bayesian hypothesis testing-based metrics, and area and distance metrics for probabilistic comparison between prediction and observation [17] have been developed, taking into account the uncertainty in both model prediction and empirical observation. These metrics are particularly suitable for physics-based models that result from different physics hypotheses (and not from the number of terms (complexity) as in regression models). Plastic pipe RUL prediction models are *hybrid* models, i.e., they utilize physics hypotheses to arrive at the composition of the model form and then use regression to fit the multiplicative model coefficients to the available test data. In this work, we compare various model validation metrics for such hybrid RUL prediction models built using accelerated hydrostatic test data.

The main challenge to the credibility of probabilistic RUL prediction models for PE pipes is the unavailability of life data under normal operating conditions. As discussed above, the PE pipe RUL models could be validated using data obtained at accelerated hydrostatic test conditions, but these are very different from normal operating conditions. Thus the "intended use" part of the validation definition is not satisfied. In many applications, data under intended use conditions may not be available. Even if such data is available, multiple models could show similar validation performance and the task of selecting the best model (for intended use conditions) may not be trivial.

In this work, we develop a novel, additional model selection criterion of *prediction consistency* when validation data corresponding to a model's intended use condition is not available. We first recognize that there is uncertainty in the intended use (normal operating) conditions for which the predictive model will be used and represent this uncertainty using a probability distribution. For the pipe application considered in this paper, the probability distribution of the intended use condition may be based on data collected for temperatures and pressures for the pipe's normal operating condition. Given this probability distribution, a reference operating condition (e.g., the *expected value* of the operating condition), and a probabilistic predictive model, we develop a four-step approach to evaluate the consistency of model prediction: (a) obtain the predicted RUL probability distribution at the reference operating conditions; (b) obtain the family of predicted RUL probability distributions corresponding to the probability distribution of the normal operating condition; (c) construct a probability distribution of the difference (distance) between the reference predicted RUL distribution and each member of the family of predicted RUL distributions; and (d) use moments of this distance distribution to quantify the consistency of the RUL prediction model under normal operating condition. Model consistency is thus defined in this work as a measure of the consistency of the probabilistic model predictions for the assumed or known variability of intended use conditions.

Thus, we propose a two-step strategy for model selection when empirical data are not available for the actual use condition: (i) validation evaluation using test data, and (ii) consistency evaluation for the likely distribution of the actual use condition. The model has to be deemed satisfactory in both steps. If multiple competing models are available, only the models that show superior validation performance on the test condition data after the first step are considered further for consistency evaluation. Then in the second step, the most consistent model (among the models that pass the first step) is selected as the best prediction model. In this manner, the proposed model selection procedure considers model performance in both test condition and intended use condition, and also includes uncertainty in both conditions as well as in model prediction. The novel contributions of this work are as follows:

1. We develop a novel, model consistency criterion for model selection. The proposed two-step approach (validation with available data, plus consistency evaluation for the intended use condition) is well-suited for selecting the best model when validation data for the intended use condition is not available. This is a general methodological contribution that is useful in many application domains.
2. Specific to RUL prediction models for PE pipes, Bayesian inference is used for the first time in this paper to calibrate such models. We compare the Bayesian approach against the traditional least squares-based approach recommended in ISO 9080 and show that the Bayesian inference-based model calibration gives better validation performance, when accelerated hydrostatic test data is used for model validation.
3. We compare the performance of various methods for validating probabilistic RUL prediction models for PE pipes. We show that the model reliability metric [13] is the most suitable metric to identify the best probabilistic RUL prediction model.

In Section 2, we develop the proposed RUL model calibration,

validation, and selection methodology. In Section 3, we illustrate the proposed methodology for calibrating, validating, and selecting RUL prediction models for PE pipes, using publicly available accelerated hydrostatic test data. In Section 4, we provide concluding remarks and discuss future extensions of this work.

2. Methodology

In this Section, we first describe the available RUL prediction models for PE pipes in the literature. We then discuss the model calibration and validation methods in details. Finally, we discuss the proposed model selection methodology.

2.1. PE pipe RUL models based on hydrostatic testing data

Multiple RUL prediction model forms have been proposed for PE pipes in the literature. For predicting chemical degradation, most of the models use the rate process method (RPM) based on the Arrhenius equation. The Arrhenius equation implies that the logarithm of the rate of chemical degradation (chemical reaction) is a linear function of the reciprocal of the temperature. For mechanical degradation, various theories of mechanical failure have been used and relationships between RUL and hoop stress in the pipe have been derived. RUL prediction models thus consist of additive terms that account for chemical degradation (terms that depend only on temperature) and mechanical degradation (terms that depend only on hoop stress), and terms that account for the coupled action of the two phenomena (terms that depend both on temperature and hoop stress). All models considered in this paper predict the logarithm of RUL in hours. These models have either three or four additive terms and the same number of model parameters. A summary of the five candidate models considered in this work is given below:

1. RPM model

This model was developed by Coleman [18] using the rate process method (RPM) and the observed linear relationship between the logarithm of the hoop stress and the logarithm of the failure time. The three-parameter model is derived by combining the two relationships as

$$\ln(t) = A + \frac{B}{T} + \frac{C\,ln(P)}{T}, \tag{1}$$

where A, B, and C are model parameters to be estimated from the data, t is the remaining useful life (in hours), P is the hoop stress (in MPa), and T is the temperature (in degrees K). Note that many other RUL prediction models have been proposed based on the RPM [18], however the model in Eq. (1) is widely known as the RPM model.

2. Norman Brown's first model (NB1 model)

This model was developed by Brown [19], and it assumes that semi-crystalline (PE) polymers contain crystallized areas, or so-called force centers. The force centers are bonded to each other by tie molecules. When mechanical stress is applied to the material, the chemical attraction (bond) force between the force centers and tie molecules is overcome, and the tie molecules are pulled out from the force center. A crack is initiated when a tie molecule is pulled out from the force center. The NB1 model calculates the rate of the pull-out process, and the RUL has a linear relationship with the pull-out rate. The model has the following form:

$$\ln(t) = A + \frac{B}{T} + \frac{CP^2}{T}. \tag{2}$$

where A, B, and C are model parameters to be estimated from the data, and P, T and t have the same meaning and units as those in Equation (1).

3. Norman Brown's second model (NB2 model)

This model was also developed by Brown [20], and it considers how the average molecule length influences the slow crack growth rate by postulating that the crack growth is caused by the disentanglement of the polymer material fibrils. The disentanglement rate depends on the number of tie molecules. The number of tie molecules is proportional to the size of the amorphous region (the region that contains tie molecules), and the area fraction of the amorphous region has a linear relationship with the average molecule weight. We refer to this model as the NB2 model, and it is given by:

$$\ln(t) = A + \frac{B}{T} + C\,ln(P). \tag{3}$$

where A, B, and C are model parameters to be estimated from the data, and P, T and t have the same meaning and units as those in Equation (1).

4. Bragaw's model (BG model)

The fourth model was proposed by Bragaw [21], by comparing the performance of multiple model forms using least squares regression fit for PE pipe hydrostatic failure test data. Three candidate model forms were postulated, based on the RPM approach and by employing different terms consisting of logarithm of the temperature, logarithm of the hoop stress, or simply the hoop stress. The predictions from the three models were evaluated using the correlation coefficient (R) test and lack of fit (F) test. The most suitable model form was identified as:

$$\ln(t) = A + \frac{B}{T} + \frac{CP}{T}. \tag{4}$$

where A, B, and C are model parameters to be estimated from the data, and P, T and t have the same meaning and units as those in Equation (1).

5. ISO model

ISO 9080 [4] recommends a four-parameter model, which is a combination of the RPM method and the NB2 model. This model considers Arrhenius equation for degenerative chemical reaction (similar to RPM) and disentanglement of fibrils for slow crack growth (similar to NB2). The resulting four-parameter model is given by:

$$\ln(t) = A + \frac{B}{T} + Cln(P) + \frac{C\,ln(P)}{T} \tag{5}$$

where A, B, and C are model parameters to be estimated from the data, and P, T and t have the same meaning and units as those in Equation (1).

The next step is to estimate the parameters of these models using available hydrostatic test data. We discuss the two model calibration methods investigated in this article.

2.2. Model calibration

Consider models of the form

$$y = g(\boldsymbol{x};\boldsymbol{\theta}), \tag{6}$$

Where y is the model output (the quantity to be predicted, e.g., $y = \ln(t)$ in equation (5)), $\boldsymbol{x}$ is the vector of model inputs (e.g., $\boldsymbol{x} = [P,T]^{\mathrm{T}}$ in equation (5)), and $\boldsymbol{\theta}$ is the vector of model parameters (e.g., $\boldsymbol{\theta} = [A\ B\ C\ D]^{\mathrm{T}}$ in equation (5)). Model calibration can be defined as finding a unique joint probability distribution of model parameters ($\boldsymbol{\theta}$) that provides the best description of the system behavior and can be achieved by comparing model predictions against actual data (D) obtained from the system [22]. This data is in the form of observed/measured values of model outputs and the corresponding inputs, i.e., $D = \{y_i^D, \boldsymbol{x}_i^D\}$, $i = 1, 2, \ldots N$.

2.2.1. Least squares (LS) model calibration

Consider the model in Equation (6). When comparing model prediction to observation, we can write:

$$y^D = g(\boldsymbol{x}^D, \boldsymbol{\theta}) + \varepsilon, \tag{7}$$

where ε is the residual between the model estimate $g(\boldsymbol{x}^D, \boldsymbol{\theta})$ and the observed value y^D. To estimate the model parameters $\boldsymbol{\theta}$, we first define the sum of squares of errors (SSE) as:

$$SSE(\boldsymbol{\theta}) = \varepsilon^2 = \sum_{i=1}^{N} \left(y_i^D - g(\boldsymbol{x}_i^D, \boldsymbol{\theta})\right)^2. \tag{8}$$

The least squares regression seeks to estimate the value of $\boldsymbol{\theta}$ that minimizes $SSE(\boldsymbol{\theta})$, such that [22]

$$\nabla_\theta SSE(\boldsymbol{\theta}) = 0. \tag{9}$$

The optimal $\boldsymbol{\theta} = \boldsymbol{\theta}_{LS}$ can be computed by solving equation (9). The residual (ε) is assumed to be a Gaussian random variable with zero mean and standard deviation equal to the square root of $SSE(\boldsymbol{\theta}_{LS})$.

2.2.2. Bayesian inference (BI) for model calibration

In Bayesian inference-based model calibration, the knowledge about the values of model parameters ($\boldsymbol{\theta}$) is represented using prior and posterior probability distribution functions. The updated knowledge (i.e., the posterior distribution) of model parameters is obtained by combining prior knowledge (based on intuition, experience, model prediction, prior data, etc.) and observed data. Specifically, the likelihood of observing the data for a given value of the parameter is computed, and the Bayesian inference for estimation of the model parameters $\boldsymbol{\theta}$ in Eq. (7) is given by:

$$P(\boldsymbol{\theta}|D) \propto L(\boldsymbol{\theta}) P(\boldsymbol{\theta}), \tag{10}$$

where the $P(\boldsymbol{\theta}|D)$ is the posterior distribution, $P(\boldsymbol{\theta})$ is the prior distribution, $L(\boldsymbol{\theta})$ is the likelihood function. The prior distribution is a distribution of $\boldsymbol{\theta}$ based on prior information, and the likelihood function can be written generally as:

$$L(\boldsymbol{\theta}) \propto \Pi_{i=1}^{N} P\left(g(\boldsymbol{x}_i^D, \boldsymbol{\theta}) = y_i^D \middle| \boldsymbol{\theta}\right), \tag{11}$$

where $g(\boldsymbol{x}_i^D, \boldsymbol{\theta})$ is the model prediction with parameter $\boldsymbol{\theta}$ and input data $\boldsymbol{x}_i^D$. The next step after estimating the model parameters is to validate the model using additional data, independent from the calibration dataset [23]. Model validation methods are discussed next.

2.3. Model validation

Several model validation methods and metrics for deterministic and stochastic models have been reported in the literature [11,24,25]. For deterministic models, the model validation methods focus on quantifying the error between the model prediction and the measured (validation) data by computing metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE). These metrics are based on the distance between the model prediction and validation data; larger distance means lower accuracy and worse performance.

Probabilistic model validation methods include classical hypothesis testing, Bayesian hypothesis testing, area metric method, and the model reliability method (distance metric). Classical hypothesis tests include Kolmogorov test, Hellinger distance and f-divergence. In general, these tests assume a null hypothesis (accept the model) and alternate hypothesis (reject the model). Standard procedure of statistical hypothesis testing is used to determine whether the null hypothesis can be rejected or not. The Kolmogorov test measures the difference in predictive and measured cumulative distribution functions (CDFs), and Hellinger distance is the L2 norm of the difference between two distributions. The f-divergence is a generalized metric to measure the distance between two probability distributions. This function is based on the ratios between the probability density function value of two distributions, and it is given by:

$$D_f(P\|Q) = \int_\Omega q(x) f\left(\frac{p(x)}{q(x)}\right) dx, \tag{12}$$

where P and Q are the two distributions; p and q refer to probability density functions (PDFs), and Ω is the union of the supports of the two PDFs. The function f is chosen such that $f(1) = 0$, which ensures that the f-divergence is 0 if the two distributions are identical. A widely used choice for f, $f = -log(x)$, gives the Kullback-Leibler divergence (KLD) as follows:

$$D_{KL}(P\|Q) = -\int_\Omega q(x) log\left(\frac{p(x)}{q(x)}\right) dx \tag{13}$$

It is important to note that if the functions $q(x)$ & $p(x)$ have unequal support, then the KLD value will be infinity. So, the KLD cannot be used for distributions with different support.

Bayesian hypothesis testing calculates the marginal likelihood ratio, known as Bayes factor [11], for the validation data. The Bayes factor is a ratio of the likelihood of observing the (validation) data with the null hypothesis (i.e., the probabilistic prediction of the model being evaluated) vs. the likelihood of observing the data with the alternate hypothesis (i.e., prediction without using the model being evaluated, e.g., range predicted by an expert), and higher likelihood ratio indicates a better model. The area-metric based model validation metric [26] takes the integral of the absolute difference between the PDFs of two distributions, which makes it sensitive to the mean and standard deviation of distributions as well as the entire distribution of the variables [27].

For probabilistic prediction models, the model output is a probability distribution, and the validation data is a set of point data. Bayes factor may be used to perform model selection in this setting, where Bayes factor is calculated as the ratio of data likelihood corresponding to two candidate models. Both KLD and Bayes factor can only reject or accept a model by comparing two models but cannot give the confidence level of the decision to reject or accept. The model reliability metric [11] measures the probability that the difference between the model prediction and validation data is within a prescribed tolerance limit. If the validation data point is y^D and the tolerance limit is σ, then based on the probabilistic model prediction y, the model reliability can be computed as:

$$R = P\left((1-\sigma) * y^D < y < (1+\sigma) * y^D\right). \tag{14}$$

See Ref. [11] for a detailed discussion and comparison of the model validation methods mentioned in this section. All these methods focus on how the prediction of a particular model agrees with the observation. In the next section, we discuss how this information and an additional consideration may be used to *select* a particular model as *the best* model for a given engineering application.

2.4. Model selection

Model selection is concerned with choosing the best predictive model from a set of candidate models. The simplest form of model selection methodology relies on comparing validation metric values (Section 2.3) corresponding to the candidate models and selecting the model that exhibits the best validation metric value. This is validation metric-based model selection. However, although the model validation metrics consider the model prediction error, they do not consider model complexity, which is an important consideration for regression models based on empirical data. A model with higher complexity (i.e., with more parameters) may have a lower validation error, but it may require more calibration data and computational effort and may not perform well for model inputs different from the model training data. To find a balance between model complexity and model accuracy, several model

selection methods that consider model complexity have been proposed in the literature [14–16]. Three of them are discussed in Section 2.4.1.

2.4.1. Complexity-based model selection method

1. Akaike information criterion (AIC):

Since a model rarely represents a real-world process exactly, the model calibration process almost always results in some information loss. The AIC metric measures the amount of information loss for each model. The AIC metric also has a model complexity-related penalty term, which is directly linked to the number of parameters in the model. The AIC metric is given by:

$$AIC = nlog(SSE)+2p+\frac{2p(p+1)}{n-p-1} \tag{15}$$

where *SSE* is the sum of squared error, p is the number of parameters, and n is the size of the dataset. Note that the AIC does not give an absolute measure of model performance but can be used as a relative measure in comparing one model to another.

2. Bayesian information criterion (BIC)

Similar to AIC, the BIC metric uses an additive penalty to account for model complexity. The BIC metric is computed as:

$$BIC = n\ log(SSE)+p\ \log(n) \tag{16}$$

In BIC, the model complexity penalty term is in terms of the logarithm of the number of samples. Due to this term, BIC is more sensitive to number of the parameters and less sensitive to number of the samples compared to AIC.

3. Minimum description length (MDL)

The MDL (minimum description length) casts the (regression) model selection problem as the problem of selecting *the shortest-length description of the training data*. MDL considers the goodness of fit of the model as well as model complexity. Here, we employ the mixture form of the description length to define MDL [28]. MDL is defined for two cases, depending on data availability, as

$$\begin{aligned} MDL &= n\ log\left(\frac{SSE}{n-p}\right)+p\ log\ (F)+log\ (n)\ \left(if\ R^2 > \frac{p}{n}\right), (case\ I)\\ MDL &= n\ \log\left(\frac{\boldsymbol{y}^T\boldsymbol{y}}{n}\right) + \log(n)\ (case\ II), \end{aligned} \tag{17}$$

where R^2 is the R-square value between the model prediction and validation data, p is number of parameters in the model, F is the F-score of the model, n is the size of the dataset, and $\boldsymbol{y}$ is the model output vector. Sometimes the number of (observation) data points is too small compared to the number of parameters (i.e., p/n is too large), and the information contained in the dataset is unlikely to be able to support a correct estimate of the likelihood. In that case, the second metric (Case II in Eq. (17)) is designed to solve this problem. The AIC, BIC and Case I of MDL use the sum of squared errors (SSE) between the model predictions and the experimental values. Selecting one of the two cases in computing the MDL metric also requires experimental data. Without validation data, the three metrics cannot be computed, and model selection cannot be performed. Therefore, in the next Section, we develop a new model selection framework, aimed at alleviating this problem.

2.4.2. Consistency-based model selection method

As discussed above, if data corresponding to the normal operating conditions is not available, then model validation metrics-based model selection methods cannot be used. Even if data is available, several competing models may have similar values for the model validation and selection metrics, thus requiring an additional criterion to distinguish between them. Here, we propose that model selection can be aided by considering model consistency as the additional criterion. We define that a model is consistent if a small change in the input results in only a small change in the (probabilistic) model prediction (output). For model selection, the model consistency could be quantified for the probability distribution of normal operating conditions.

Given a probabilistic prediction model and the probability distribution of normal operating conditions (denoted by $h(\boldsymbol{x})$), the model consistency quantification requires (a) a reference operating condition, and (b) a metric of distance between the probabilistic model predictions at the reference operating condition and other probable operating conditions (obtained from the probability distribution of the normal operating conditions). The reference condition, denoted by $\boldsymbol{x}_r$, could be the mode or the mean value of the normal operating condition distribution. The predicted probability distribution of the output quantity of interest corresponding to the reference condition is obtained using the probabilistic prediction model and is denoted by $f(y_r)$, as:

$$f(y_r) = g(\boldsymbol{x}_r, \boldsymbol{\theta}), \tag{18}$$

where $g(\boldsymbol{x}_r, \boldsymbol{\theta})$ is the prediction of the calibrated probabilistic model, and $\boldsymbol{\theta}$ is the vector of (calibrated) model parameters. For each sample of the normal operating condition $(\boldsymbol{x}_j)$, obtained by sampling the distribution $h(\boldsymbol{x})$, the probability distribution of the model output can be obtained as:

$$f(y_j) = g(\boldsymbol{x}_j, \boldsymbol{\theta}). \tag{19}$$

We now need a metric to quantify the distance between two probability distributions $f(y_r)$ and $f(y_i)$. Here, we use the Jensen-Shannon divergence (JSD) [29] as the metric of distance between two probability distributions. The JSD for two distributions (P and Q) is given by:

$$JSD = JSD(P||Q) = \frac{1}{2}D(P||M) + \frac{1}{2}D(Q||M); M = \frac{1}{2}(P+Q), \tag{20}$$

where $D(P\|Q)$ is the KLD defined in Equation (13). Note that unlike KLD, when distributions P and Q have different ranges, JSD does not become infinitely large. This is an important benefit of using JSD, when the probability distributions of the model output for some sample realizations of the normal operating condition may have different ranges than the probability distribution of the model output for the reference condition. To measure the similarity of the predictive distribution $f(y_j)$ with the reference predictive distribution $f(y_r)$, the JSD value between these distributions is calculated as:

$$JSD_j = JSD\left(f(y_j)||f(y_r)\right). \tag{21}$$

Using Equation (21), JSD_j values corresponding to each sample of the normal operating condition $(\boldsymbol{x}_j)$ can be obtained. A lower value of JSD_j implies stronger similarity between the probabilistic model predictions for $\boldsymbol{x}_j$ and $\boldsymbol{x}_r$. This process is depicted in Fig. 1.

The model consistency could be quantified by computing the moments of the JSD distribution represented by the samples of JSD_j. The JSD distribution effectively quantifies the spread of (probabilistic) model predictions over the probability distribution of the normal operating condition, and hence provides a meaningful model consistency metric. A lower mean value for the JSD distribution indicates more similarity between the model outputs for the reference distribution and samples of normal operating conditions, which implies more consistent model predictions. A lower variance of the

JSD distribution indicates that a larger number of JSD values are closer to the mean value. Thus, the model that has the lowest mean value and standard deviation of the JSD distribution will be judged as the most consistent model.

It is important to note that this procedure gives the model consistency assessment for a given reference value and for a given distribution of operating conditions. Model consistency-based rankings of different models may change if the reference condition or the distribution of operating conditions of interest are changed. Note that this model selection procedure correctly considers the probability distribution of the operating conditions of interest when selecting the most suitable model. When validation data corresponding to the operating conditions is not

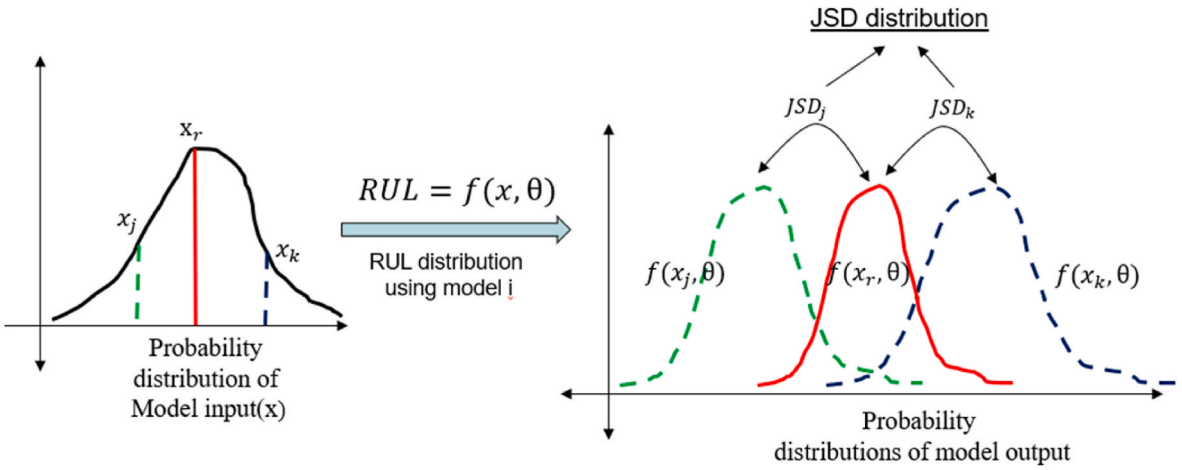


Fig. 1. Computation of the distribution of JSD.

available, this is a reasonable way to compare model performance and select the most suitable model. Also note that in the proposed two-step model selection process, the first step selects the models with the highest validation scores. Then in the second step, the consistency criterion is evaluated only for these chosen models; This ensures that the selected model is *both accurate and consistent*. The proposed two step procedure (first validation then consistency check) thus helps select the model that has a high validation score and the best predictive consistency.

3. Numerical implementation and results

To illustrate the proposed model calibration, validation, and selection methodology, we use two publicly available datasets.

3.1. Example 1

In the first example, we use a PE pipe accelerated failure test dataset (referred to as Dataset I) from the Chevron Phillips Chemical Company [30], a well-known polymer manufacturer. That dataset was developed to investigate the relationship between failure stress and failure time of the PE pipes used to protect electric wires/cables.

3.1.1. Accelerated testing data for model calibration

The pipes are subjected to a certain hoop stress and testing temperature, and failure is defined as continuous pressure loss for the gas inside the pipe, which indicates a crack on the pipe wall. Eighteen data points from the accelerated testing dataset are divided into two parts: 12 data points for model calibration and 6 data points for validation (see Fig. 2).

In the numerical example, this dataset is employed to calibrate five different probabilistic RUL prediction models using the two methods (least-squares regression and Bayesian inference) described in Section 2.1. Several model validation and model selection metrics are used to compare the predictive performance of five models and to demonstrate the utility of the proposed model selection methodology.

3.1.2. Model calibration

3.1.2.1. Least squares regression. The least square linear regression method described in Section 2.2.1 is first used to calibrate the five models given in Section 2.1, using the NumPy Python library. Table 1 gives the parameter estimates.

3.1.2.2. Bayesian calibration. The prior distribution of model parameters represents the prior knowledge based on intuition, experience, model prediction, prior data, etc. Here, we use an approximate methodology to obtain the lower and upper bounds of independent marginal

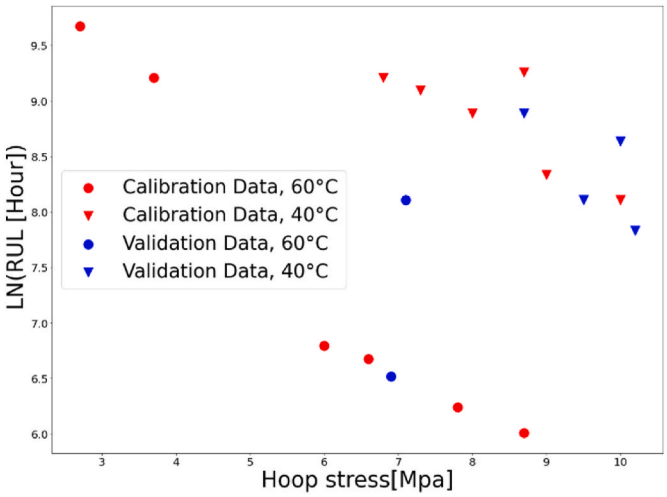


Fig. 2. Accelerated hydrostatic test PE pipe failure data.

Table 1
Parameter estimates using least squares.

Method name	*A*	*B*	*C*	*D*	Std. dev. of ε
RPM	−37	16620	−1149	–	0.73
NB1	−35	14796	−138	–	0.86
NB2	−33	14590	−6	–	0.73
BG	−37	16602	−194	–	0.78
ISO	−5	6002	−20	4028	1.10

uniform prior distributions of the parameters. The methodology involves solving numerous linear systems of equations using hydrostatic test data to obtain initial (upper and lower) bounds. The upper/lower bounds obtained by this approximate method are increased/decreased by 10 % (or more, if desired) to account for any errors in the approximate methodology. Note that we use an unbiased (uniform) distribution as the prior distribution and bounds of these distributions merely represent the possible highest and lowest values of model parameters. The modified upper and lower bounds of uniform prior distributions of different model parameters are given in Table 2. For a calibration dataset with more than one data point, the data is assumed to be statistically independent. The overall likelihood of N data points is thus computed by multiplying their individual likelihoods given the candidate values of model parameters.

We use Markov Chain Monte Carlo (MCMC) sampling (Metropolis

Table 2
Prior distribution bounds.

Method name	*A*		*B*		*C*		*D*	
	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
RPM	−58	−30	13567	25039	−2084	−402	–	–
NB1	−47	−26	11018	18285	−13	−3	–	–
NB2	−48	−26	12217	21805	−7	−1	–	–
BG	−51	−28	12072	20706	−221	−64	–	–
ISO	−40	30	−5000	20000	−20	−3	500	5500

algorithm) [31] to obtain samples from the posterior distribution. In MCMC, the burn-in technique is commonly used [32], in which samples at the beginning of the Markov chain are discarded and the remaining samples are used to approximate the posterior distribution [33]. Discarding the initial part of the chain can negate the influence of the start point's location and make the posterior distribution more stable. The number of discarded samples, or the burn-in ratio, needs to be decided to obtain a good approximation of the posterior distribution. In this article, we use the Kullback–Leibler divergence (KLD, Equation (13)) to decide the burn-in ratio. Specifically, KLD is used to quantify the difference between two (approximate) posterior distributions with different burn-in ratios (initial λ% samples). The posterior obtained using the last Γ% of the original Markov chain is taken as the reference distribution for computing KLD. The first λ% samples are removed and KLD between the remaining part of the original Markov chain and the reference distribution is computed. This is repeated for different λ values to find the relationship between λ and the KLD value. It is found that $\lambda = 70$ provides a good match with the reference distribution, hence this value of λ is chosen as the burn-in ratio. After rejecting (burning-in) the initial 70 % samples from the Markov chain, the final Markov chain is generated. The final Markov chain contains Monte Carlo samples drawn from the posterior distribution of all parameters. Based on the samples, the distribution of parameters is generated. These posterior distributions are depicted in Fig. 3.

Note that in the prior distribution, the model parameters are assumed to be independent, while in the posterior distribution, the model parameters are correlated as expected. The correlations of the model parameters, based on the posterior distribution, are shown in Fig. 4. For four three-parameter models, the highest correlation coefficients among the model parameters are close to 1. This suggests that those parameters are highly correlated with each other. For the ISO model, the correlation is different: the absolute value of correlation coefficients between parameters *A* and *C* as well as parameters *A* and *D* are less than 0.5. In general, the model parameters in three-parameter models have higher correlation among themselves than the parameters in ISO model. This difference may be because the ISO model has four parameters; as we estimate more parameters from the same amount of data, the correlation gets more diffused.

Next, we compare the prior and posterior distributions of the RUL obtained using the Bayesian model calibration described in Section 2.2. For comparing probabilistic RUL predictions, we assume room temperature (293 K) and atmospheric pressure (0.1 MPa) as the operating conditions. Under these normal operating conditions, the RUL with the five probabilistic models is shown in Fig. 5.

The mean and median values of the prior and posterior distributions are close to each other. In contrast, the posterior distribution's variance is smaller than the prior distribution, indicating that significant uncertainty reduction is achieved due to model calibration. The posterior RUL distribution for the RPM model is similar to that of the FT model, and the posterior RUL estimates by the NB1 and BG models show similar behavior.

We also provide comparison of Bayesian inference (BI) and least squares (LS) in Fig. 6; in particular, we compare ln (RUL) as predicted by the five models calibrated using the two methods, at 293 K and 0.1 MPa. Models calibrated using these two methods predict different mean values, and models calibrated using the LS method have smaller variance. However, graphical comparison of posterior PDFs is not adequate to assess the prediction quality resulting from the two methods. In the next Section we report on the validation metrics for the models trained using these two metrics.

3.1.3. Model validation

Four metrics (AIC, BIC, MDL, and model reliability) are computed for the five models, and two calibration methods result in ten different assessments of performance. The tolerance for computing the model reliability metric is set to be 10 % of the experimental value. The validation metrics computed for the five models (each calibrated using LS and BI) using the validation dataset are given in Table 3.

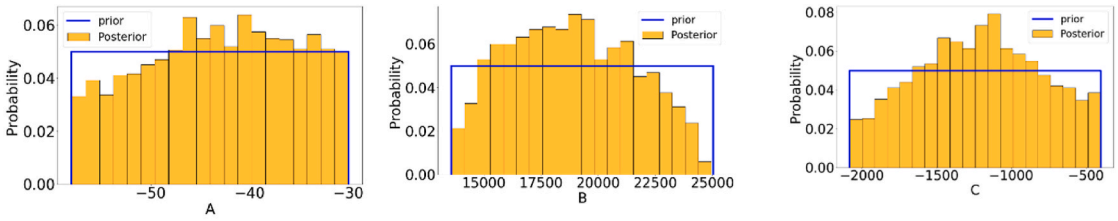
The AIC and BIC values of the ISO model are much higher than those of the other four models (note: lower the better), because this model is penalized for higher complexity (i.e., one additional parameter) compared to the other models. Hence, we discard this model during model selection. The remaining four three-parameter models have very similar values of AIC, BIC and MDL; further, they have the same number of parameters (i.e., same model complexity). Thus, model selection using the information theory-based metrics AIC, BIC, and MDL values is not successful.

As shown in Table 3, the model reliability of the four models calibrated using BI is higher than the models calibrated using LS, which indicates that relaxing the assumptions about normality and constant variance of the probabilistic predictive model (the residual used in LS) helps improving the model's predictive ability. But the reliability metric values of the four models under accelerated test condition are similar to each other; hence the best model cannot be selected solely based on model reliability metric values. As validation data under normal operating conditions is unavailable, model reliability assessment under normal operating conditions is not possible, and thus cannot be used in model selection. Thus, using the validation metric to select a model is also not successful in this case. Hence, consistency-based model selection metric is explored next, to find out whether it can help select the best model.

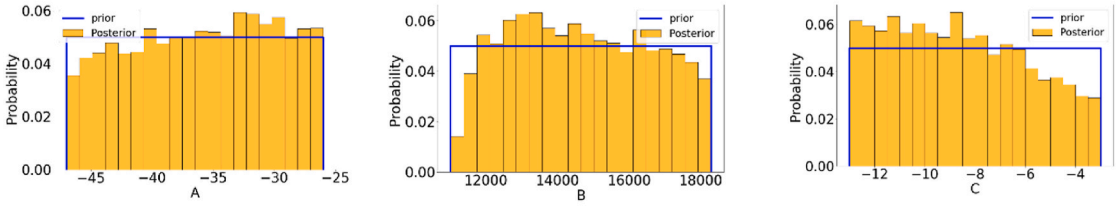
3.1.4. Consistency-based model selection

In this section we demonstrate the proposed model selection methodology that utilizes accuracy in accelerated test conditions as well as consistency in intended use conditions. In Section 3.3, we showed that the AIC, BIC MDL metrics for all four candidate methods have similar values, therefore it is not possible to select the best model using these metrics. We also found that, under accelerated test conditions, the highest model reliability is 0.63 (for the NB1(BI) model). We set the accuracy requirement to be 90 % of highest reliability (accuracy) among the candidate models (0.57). This results in rejection of two models (NB1 (LS) and BG (LS)) as their accuracy is lower than the. Note that the 90 % reliability threshold is a tunable screening parameter, and it merely ensures that poorly performing models are removed from consistency considerations. Any other suitable model reliability threshold could be chosen. The consistency-based model evaluation is performed for the selected models that show good predictive ability in accelerated test conditions.

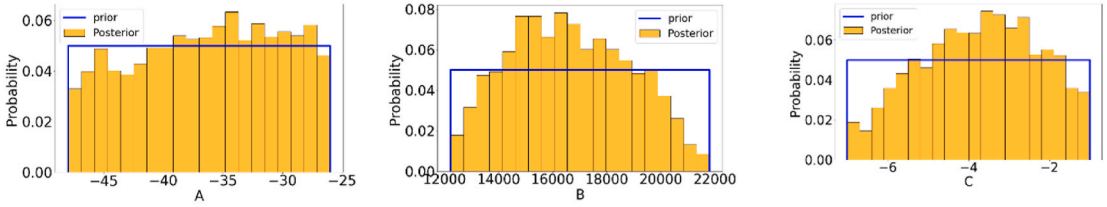
The first step of the proposed consistency-based method for model



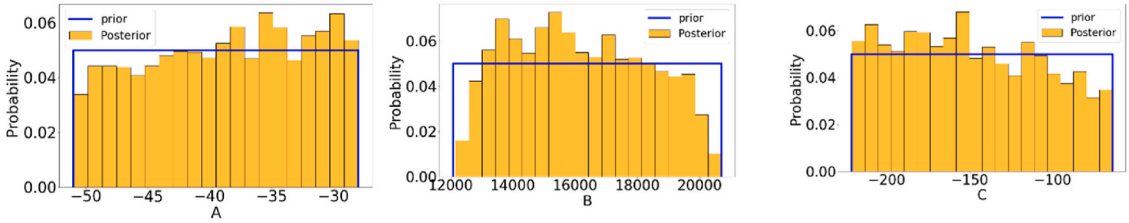
(a) Prior and posterior distributions of the RPM model parameters



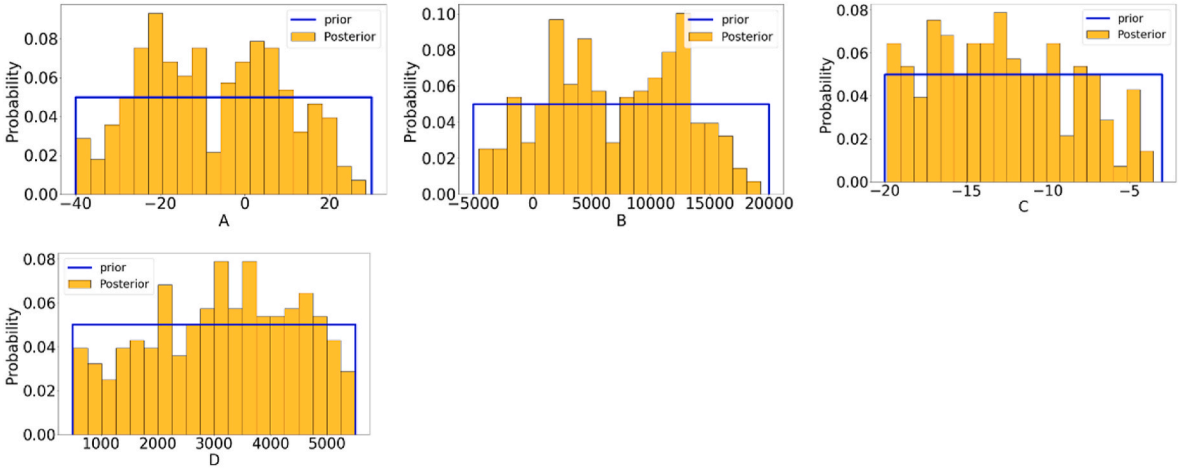
(b) Prior and posterior distributions of the NB1 model parameters



(c) Prior and posterior distributions of the NB2 model parameters



(d) Prior and posterior distributions of the BG model parameters



(e) Prior and posterior distributions of the ISO model parameters

Fig. 3. Prior and posterior distributions of model parameters.

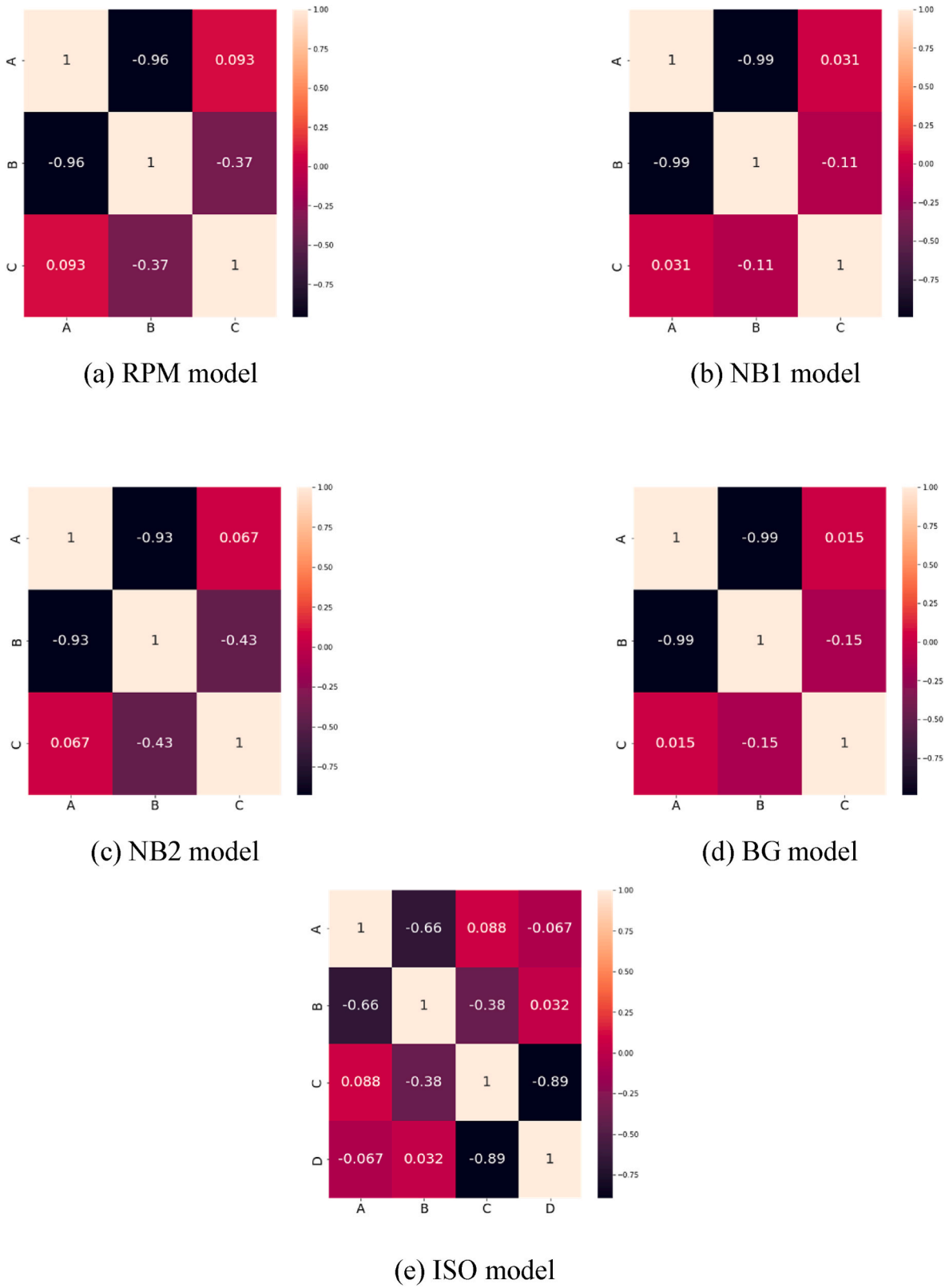


Fig. 4. Correlation coefficient matrices for calibrated model parameters (posterior distributions).

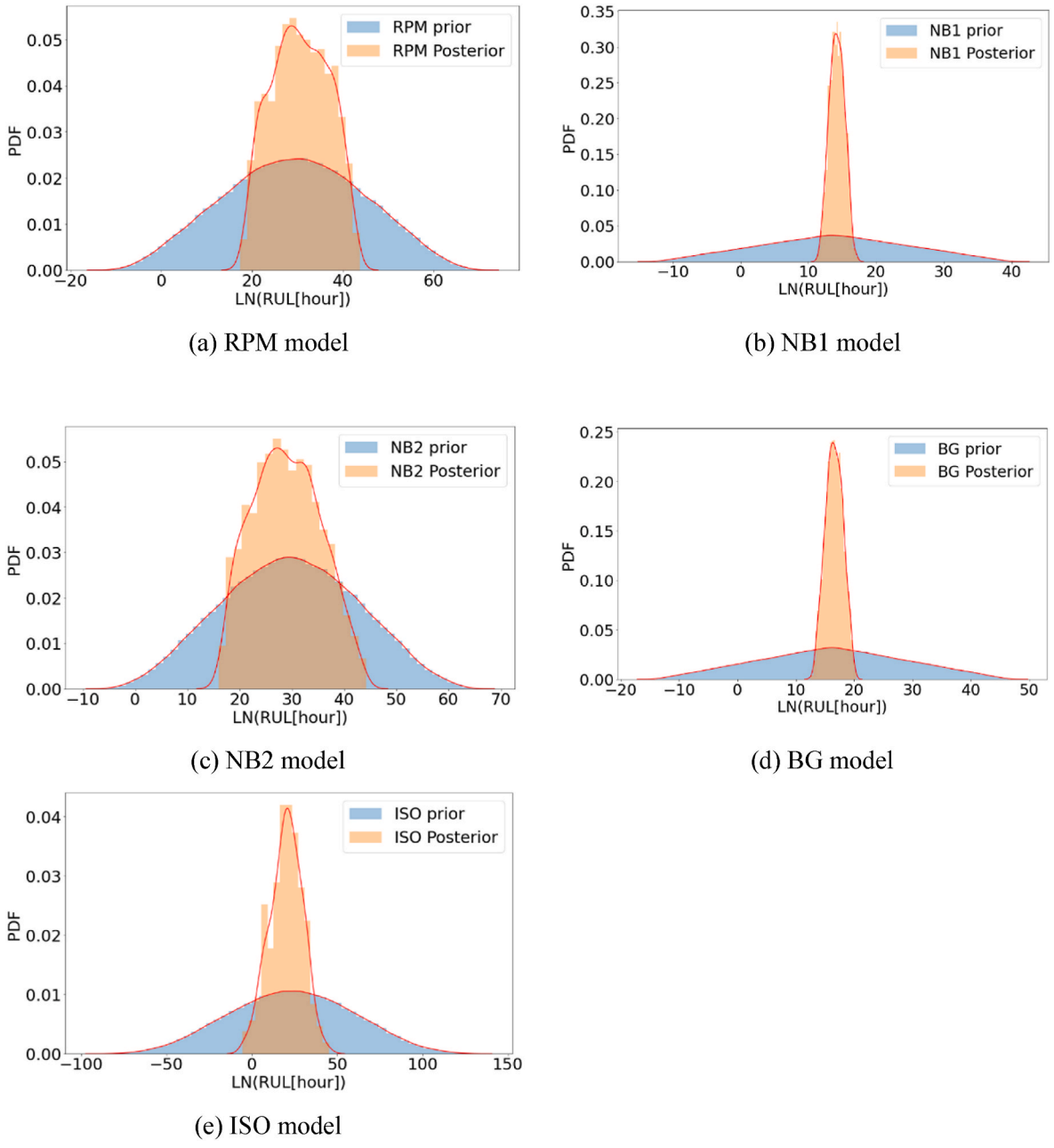


Fig. 5. Prior and posterior distributions of RUL for T = 293K, P = 0.1 MPa.

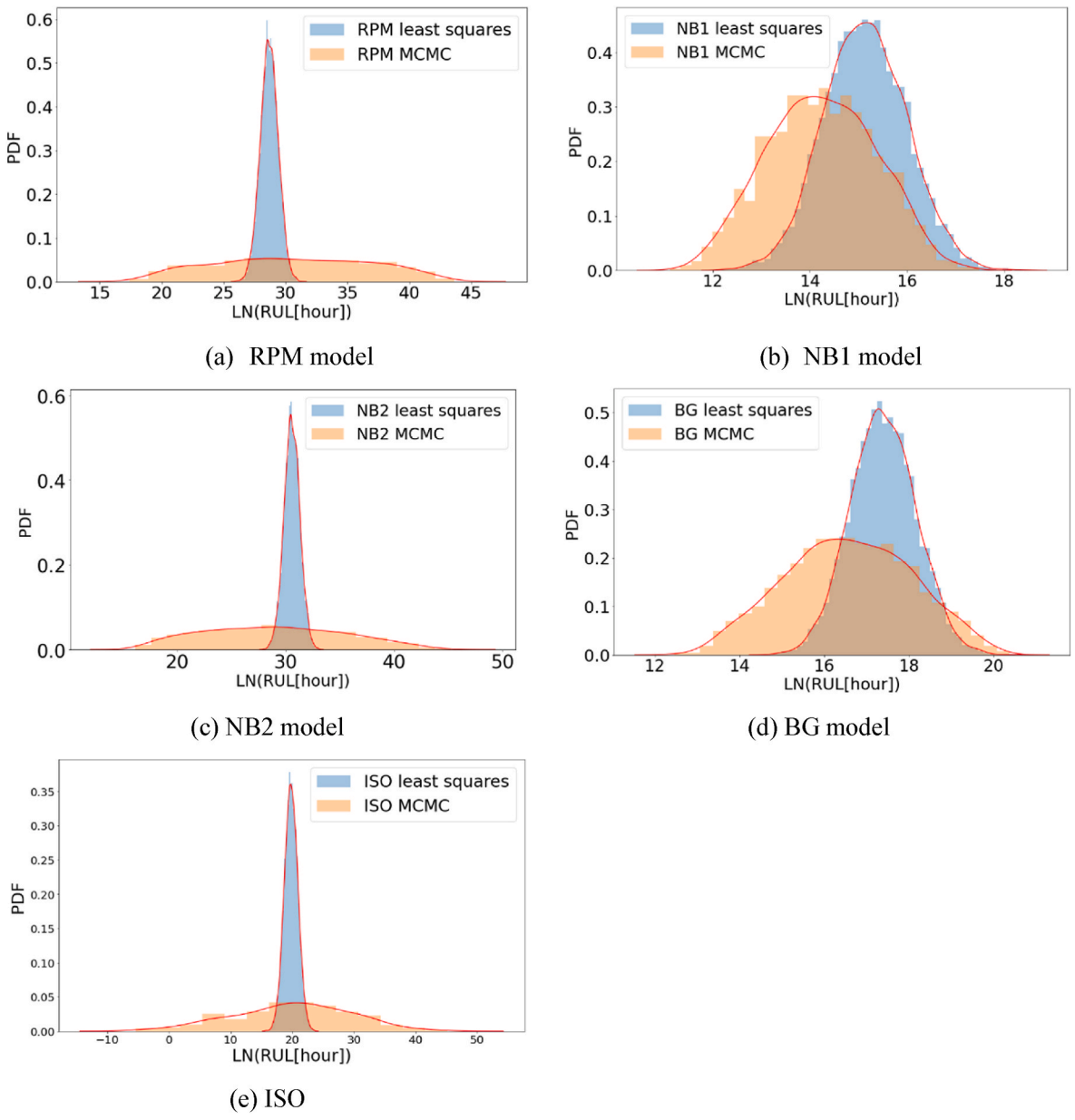


Fig. 6. Comparison of models calibrated using LS and BI for T = 293 K, P = 0.1 MPa.

Table 3
Model validation metrics for models calibrated using LS and BI.

Model	AIC	BIC	MDL	Model reliability
RPM (LS)	19.31	6.68	5.85	0.57
NB1 (LS)	18.73	6.11	5.95	0.46
NB2 (LS)	19.27	6.65	5.86	0.58
BG (LS)	19.20	6.58	5.90	0.52
ISO (LS)	50	9.16	5.90	0.41
RPM (BI)	19.3	6.67	5.90	0.59
NB1 (BI)	18.94	6.31	5.91	0.63
NB2 (BI)	19.28	6.65	5.91	0.61
BG (BI)	19.04	6.41	5.90	0.62
ISO (BI)	49.10	8.27	5.90	0.53

selection is to choose the reference input and the input distribution. In this case, the model input is the normal operating condition, while the output is RUL. Two different operating conditions are considered in the JSD distribution process. The first operating condition (OC1) has a reference temperature of 293 K (20 °C) and a reference hoop stress of 10 MPa, which mimics the conditions for a main pipeline in an underground pipe network. The temperature is set to be uniformly distributed from 291K to 295K. The hoop stress is lognormally distributed with a mean value of 10 MPa and 20 % coefficient of variation (COV). The samples drawn from this distribution of the operating condition are shown in Fig. 7.

Using this distribution of normal operating condition, the JSD distribution is obtained using the procedure described in Section 2.4.1. The distributions of the JSD values of the six models that passed the accuracy requirement are given in Fig. 8(a). In Fig. 8(a) RPM (BI) is the most

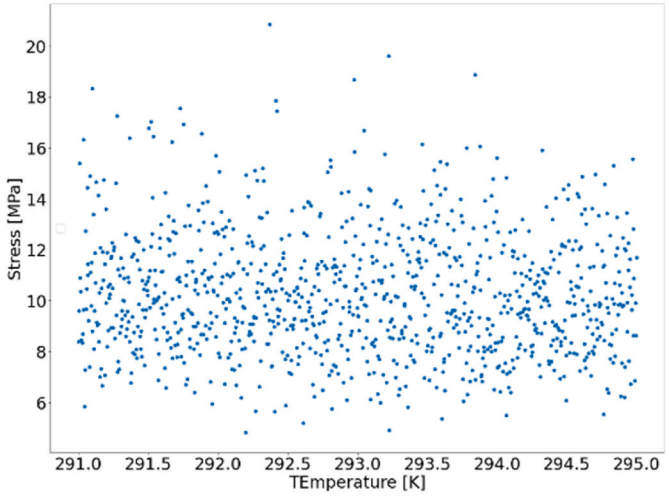


Fig. 7. Samples drawn from the probability distribution of high-stress and low-temperature operating condition (OC1).

consistent model. Its predictive consistency over the intended use condition OC1 is illustrated in Fig. 8(b), where the families of predictive distributions for RPM(BI) and NB2(LS) models are depicted. It can be seen in Fig. 8(b) that the family of RPM(BI) predictive distributions provides more consistent predictions compared to that of NB2(LS).

Another operating condition (OC2) with lower hoop stress is considered next. The temperature is set to be uniformly distributed from 291K to 295K, and the hoop stress is modeled using lognormal distribution with mean value of 0.1 MPa and 20 % COV. The reference temperature is 293 K, and the reference pressure is taken as 0.1 MPa. These conditions mimic operating conditions for a pipe that is not used for liquid transfer but for protecting wires, etc. Also, this is the lowest possible working stress for a pipe. The samples drawn from the distribution of the operating condition OC2 are shown in Fig. 9. The JSD distributions under this operating for candidate models that passed the accuracy requirement are depicted in Fig. 10(a). In Fig. 10(a) BG (BI) is the most consistent model. Its predictive consistency over the intended use condition OC2 is illustrated in Fig. 10(b), where the families of predictive distributions for BG(BI) and NB2(LS) models are depicted. It can be seen in Fig. 10(b) that the family of BG(BI) predictive distributions provides more consistent predictions compared to that of NB2(LS).

The first two moments of the JSD distributions for the two operating conditions (OC1 and OC2) are given in Table 4. For OC1, the RPM (BI) model has the lowest mean value and standard deviation of JSD, thus indicating that the RPM (BI) model is more consistent than any other method. For OC2, the results are different. The BG(BI) model has the lowest mean value and standard deviation of JSD; hence for this operating condition, the BG (BI) model gives the most consistent prediction of RUL. Figs. 8 and 10 show that RUL distribution families corresponding to lower JSD values exhibit consistent predictive distributions and vice versa. For example, for normal operating condition (OC2, Fig. 10(b)), the mean RUL value for the high JSD case (NB2(LS)) varies between roughly e^{28} (10^{12}) and e^{33} (10^{15}) hours. This is a very large range for the mean RUL prediction. Whereas for the low JSD case (BG (BI)), the mean RUL value is always about e^{17} (10^{7}) hours. Also note that the most consistent methods are different at the same temperature with different hoop stress. This difference suggests that the best model under the consistency criterion can be different with different operating conditions.

Two important properties of the proposed model selection method are demonstrated by the numerical example. First, this method does not select a single model from the candidate models for all model inputs. A different model is indicated as the most consistent model for different normal operating conditions; therefore the model selection should be repeated when the intended use conditions change. Secondly, in the numerical example, validation data is not used in the consistency-based model selection. This is the main advantage of the consistency-based model selection: unlike the validation metric-based and the information-theoretic methods, the consistency-based method does not

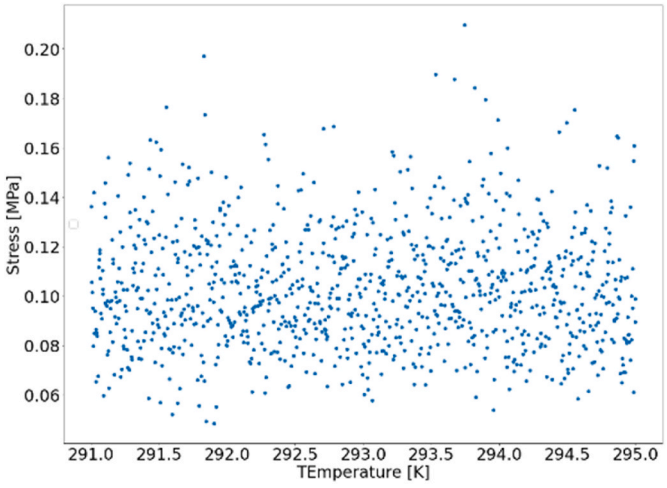
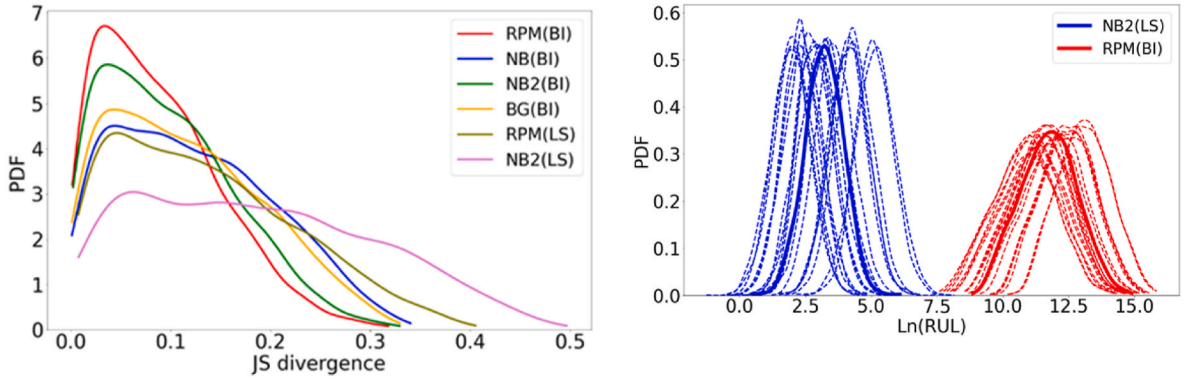
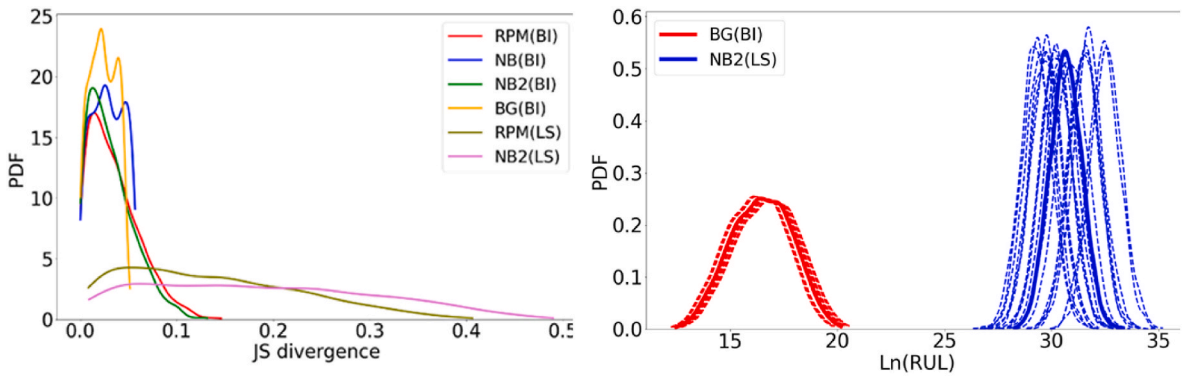


Fig. 9. Samples drawn from the probability distribution of low-stress and low-temperature operating condition (OC2).



(a) JSD distribution for OC1 (b) Families of RUL distributions

Fig. 8. JSD distributions and families of RUL distributions for OC1.



(a) JSD distributions for OC2 (b) Families of RUL distributions

Fig. 10. JSD distributions and families of RUL distributions for OC2.

Table 4
JSD distribution parameters for different models and operating conditions.

Model	Mean	Std. dev	Mean	Std.dev
OC1 (P = 10 MPa)			OC2 (P = 0.1 MPa)	
RPM (BI)	**0.092**	**0.061**	0.036	0.025
NB1 (BI)	0.124	0.074	0.028	0.016
NB2 (BI)	0.098	0.064	0.030	0.021
BG (BI)	0.111	0.071	**0.023**	**0.013**
RPM (LS)	0.144	0.088	0.141	0.087
NB2 (LS)	0.189	0.116	0.187	0.116

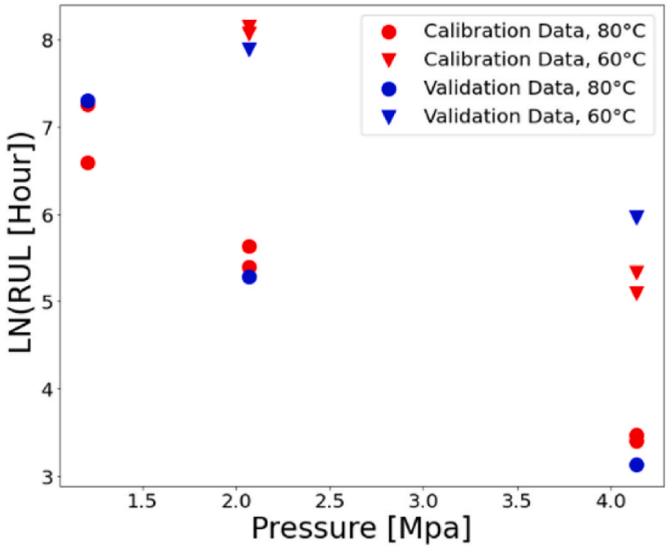


Fig. 11. PE pipe failure data from accelerated hydrostatic tests (Dataset II).

require validation data; thus, it is useful when no validation data is available (as in this case where RUL data for PE pipes under normal operating conditions is not available).

3.2. Example 2

In this example, we use a dataset from the Plastic Pipe Institute (referred to as Dataset II) for predicting the RUL for a typical natural gas delivery pipe using accelerated failure test data [34].

3.2.1. Accelerated test data

The dataset contains 15 accelerated aging test results for two temperature and three pressure (hoop stress) values, as shown in Fig. 11. We use 9 data points as Calibration dataset, and 6 datapoints as validation dataset.

3.2.2. Model calibration

All models described in Section 2.1 are calibrated with this dataset and two methods (least-squares regression and Bayesian inference). The model parameters estimated using least-squares regression are shown in Table 5.

For Bayesian inference, the lower and upper bounds of the prior distributions are computed using the method described in Section 3.1.2. The prior distributions and the posterior distributions obtained using Bayesian inference are shown in Fig. 12.

The prior and posterior distributions of the LN(RUL), obtained using the model parameters given in Fig. 12, are shown in Fig. 13.

3.2.3. Model validation

After model calibration, we use the test dataset and four metrics mentioned in Section 2.1 to compare the accuracy and complexity of the different models. The tolerance for computing the model reliability metric is set to be 10 % of the experimental value. The validation metrics computed for five models and two calibration methods are given in Table 6.

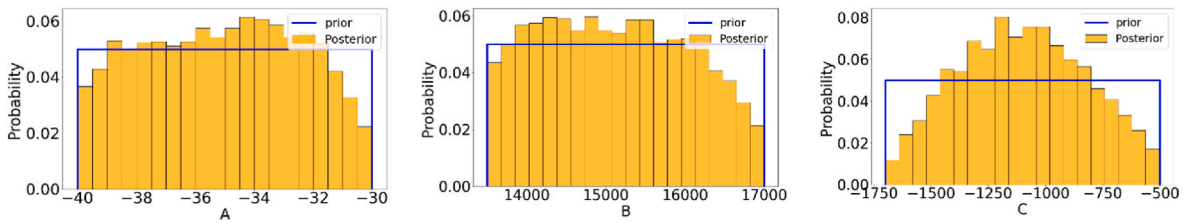
From Table 6, we notice that the ISO model exhibits very high AIC, BIC and MDL scores, hence we discard the ISO model from the model selection procedure. Both NB1 and BG model (for both calibration methods) have a relatively high BIC and MDL value, and the NB1 model (for both calibration methods) has reliability lower than 0.6. These models are also discarded. The remaining two models, RPM and NB2, have similar performance for all validation metrics; hence we need to perform consistency-based model selection among these two models.

3.2.4. Consistency-based model selection

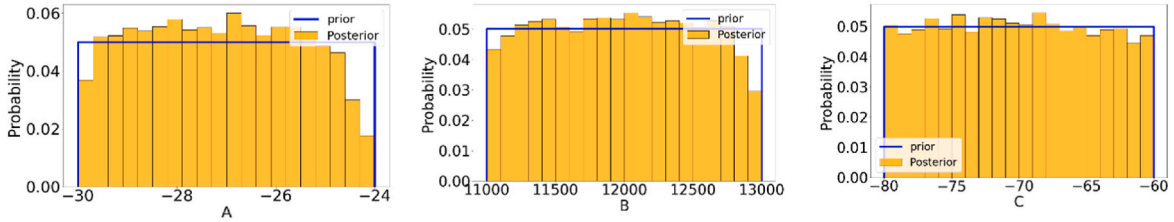
For consistency-based model selection, we use the same reference input and input distribution as in Section 3.1. The predictive consistency of the models under two different operating conditions (OC1 and OC2) is

Table 5
Parameter estimates using least squares.

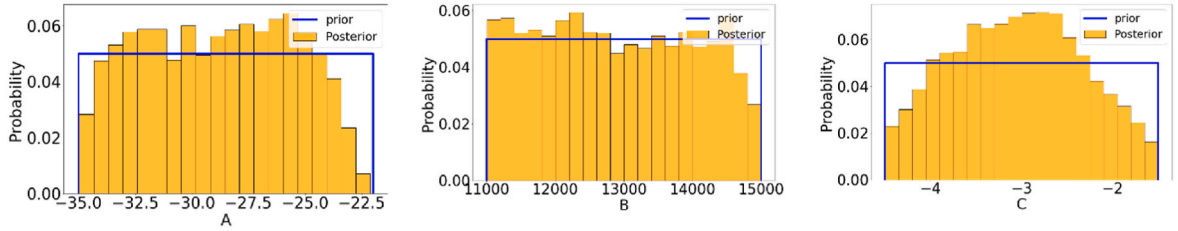
Method name	A	B	C	D	Std. dev. of ε
RPM	−34	14670	−1104	–	0.40
NB1	−29	12602	−73	–	0.45
NB2	−31	13541	−3	–	0.44
BG	−31	13897	−429	–	0.33
ISO	−53	21448	20	−7886	0.43



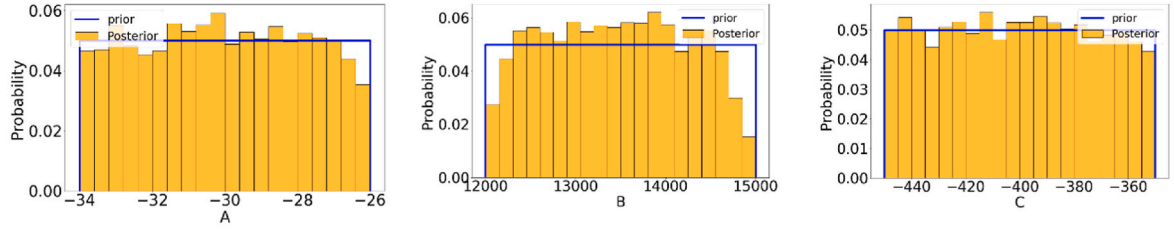
(a) Prior and posterior distributions of the RPM model parameters



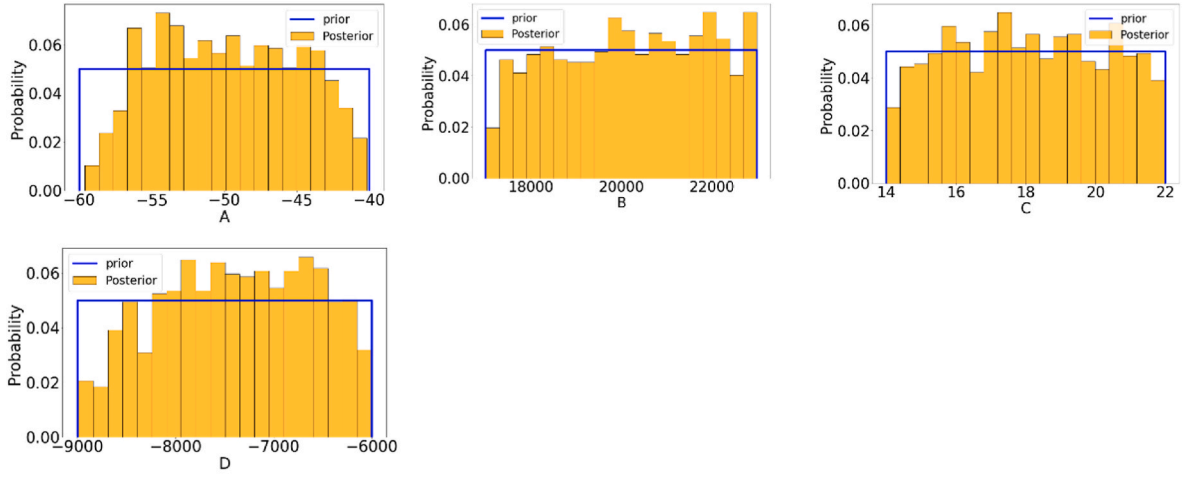
(b) Prior and posterior distributions of the NB1 model parameters



(c) Prior and posterior distributions of the NB2 model parameters



(d) Prior and posterior distributions of the BG model parameters



(e) Prior and posterior distributions of the ISO model parameters

Fig. 12. Prior and posterior distributions of model parameters.

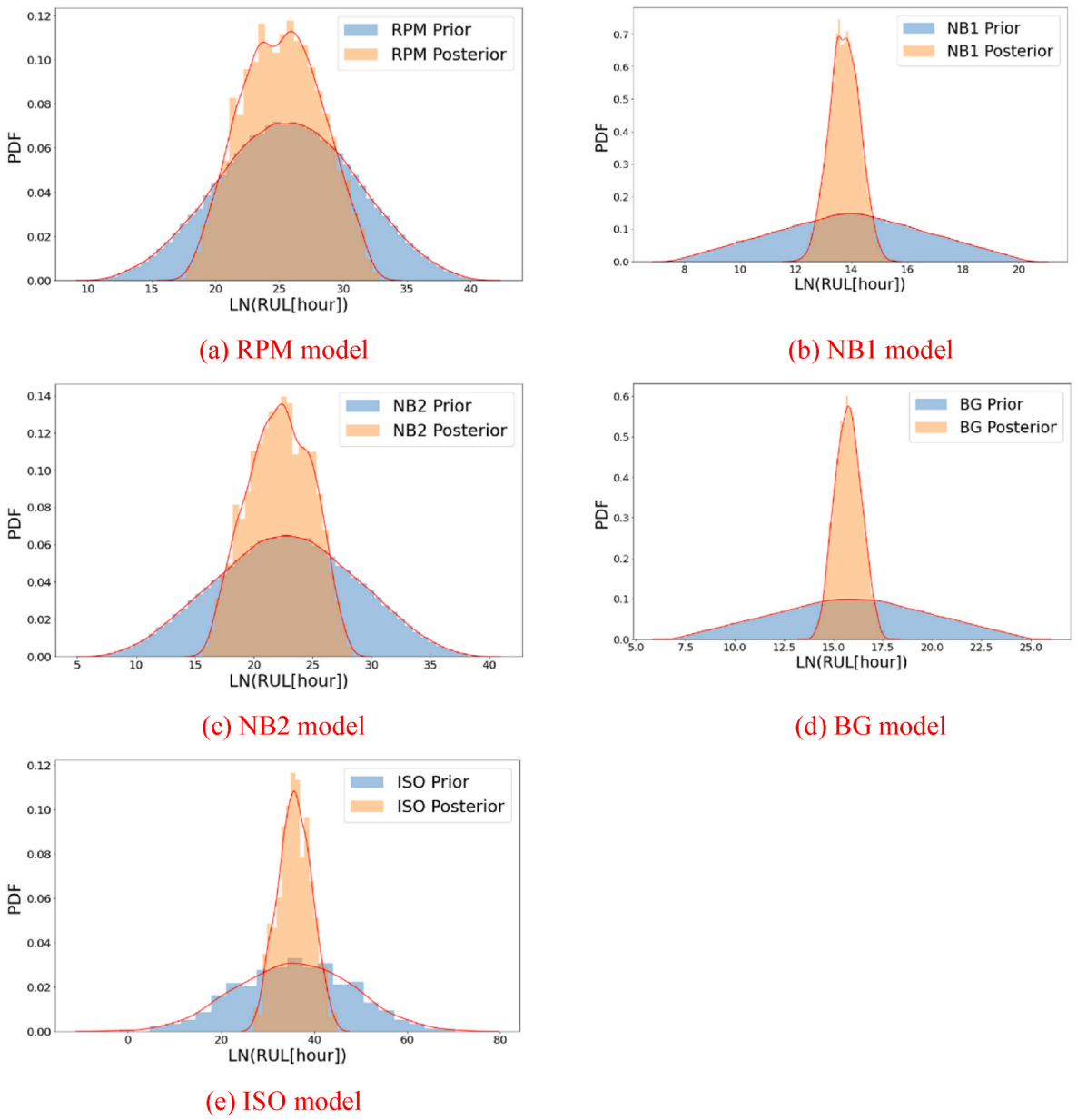


Fig. 13. Prior and posterior distributions of RUL for T = 293K, P = 0.1 MPa.

Table 6
Model validation metrics for models calibrated using LS and BI (Dataset II).

Model	AIC	BIC	MDL	Model reliability
RPM (LS)	28.94	3.76	5.52	0.76
NB1 (LS)	30.65	5.48	9.82	0.52
NB2 (LS)	28.76	3.59	5.08	0.74
BG (LS)	30.07	4.90	8.37	0.67
ISO (LS)	47.92	7.09	8.90	0.63
RPM (BI)	28.48	3.31	4.40	0.66
NB1 (BI)	30.52	5.35	9.50	0.53
NB2 (BI)	28.91	3.74	5.48	0.65
BG (BI)	29.95	4.78	8.08	0.64
ISO (BI)	48.05	7.21	9.26	0.55

given by the JSD distributions, which are shown in Figs. 14 and 15.

The numerical model consistency results (i.e., JSD distribution parameters) are shown in Table 7.

For OC1, the RPM(BI) model has the lowest JSD mean and standard deviation, hence it is a more consistent predictor of RUL than the other models. The RPM(BI) model should thus be chosen as the RUL prediction model in this operating condition. However, for OC2, the lowest JSD mean and standard deviation corresponds to the NB2(BI) model, hence it is the most consistent model among all candidate models and should be chosen for RUL prediction in this operating condition.

The model calibration and selection process for Dataset II (Example 2) corroborates the key finding from Example 1, that the best model is different for OC1 and OC2. This further reinforces the conclusion from Dataset I that the most consistent model depends on the expected use

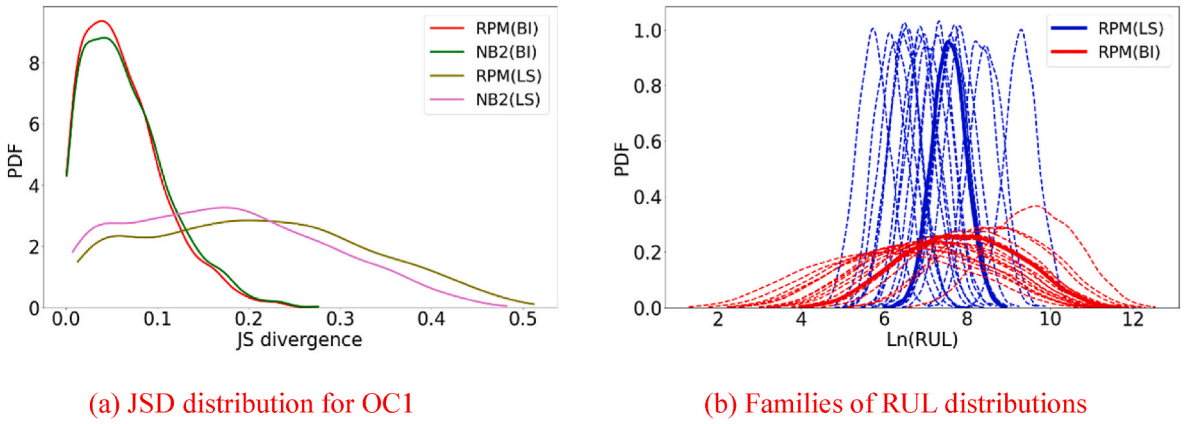


Fig. 14. JSD distributions and families of RUL distributions for OC1.

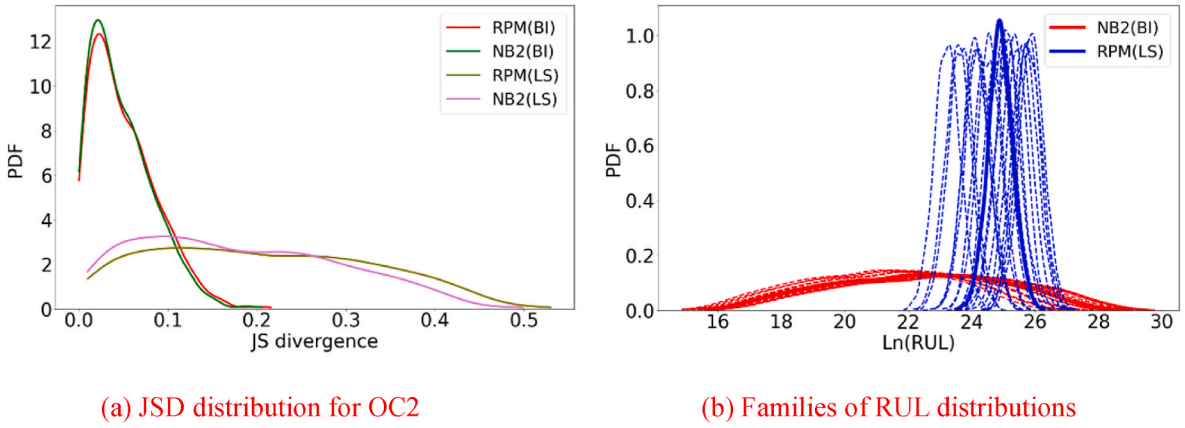


Fig. 15. JSD distributions and families of RUL distributions for OC2.

Table 7
JSD distribution parameters for different models and operating conditions.

Model	Mean	Std. dev	Mean	Std. dev
	OC1 (P = 10 MPa)		OC2 (P = 0.1 MPa)	
RPM (BI)	**0.062**	**0.044**	0.050	0.037
NB2 (BI)	0.066	0.047	**0.048**	**0.023**
RPM (LS)	0.203	0.121	0.204	0.122
NB2 (LS)	0.179	0.112	0.180	0.112

conditions. We have thus shown that the proposed consistency-based model selection can be used for different PE pipeline systems with different applications (Datasets I and II). In Section 3.1, we considered PE pipes for cable and wire protection, whereas in Section 3.2 we considered gas/liquid transfer pipes.

4. Conclusion

This article presented a framework for probabilistic model calibration, validation and selection for the RUL prediction of PE pipes, using accelerated hydrostatic test data. To the best of our knowledge, this is the first time that Bayesian inference and the model reliability metric have been used to calibrate and validate RUL models for PE pipes. The utility of Bayesian inference for calibrating RUL prediction models is established by comparing against the performance of models calibrated using least squares. The ability of the model reliability metric to identify the best RUL prediction model using validation data from accelerated failure tests is demonstrated. Finally, a model consistency-based model selection metric is developed and used to select the best RUL prediction model for PE pipes. The proposed consistency metric is useful when validation data for RUL under normal operating condition is not available. Model consistency is evaluated using the Jenson-Shannon divergence between model predictions, given the probability distribution of normal operating conditions. The utility of the proposed two-step model selection approach (first validation, then consistency check) is demonstrated using five PE pipe RUL prediction models and publicly available accelerated PE pipe failure test data.

Note that the most consistent model is different for different operating conditions. Since the real-world pipeline system may have to function adequately under multiple conditions, the operator must repeat the proposed model selection procedure over a range of possible operating conditions. An aggregated model consistency score may then be evaluated for different models and then the appropriate model could be selected.

The proposed methodology cannot guarantee model accuracy under the operating condition. Step II (validation) of the method does consider model accuracy, but w.r.t. the validation dataset (under accelerated test conditions). Step III (consistency-based model selection) of the proposed method only considers the model consistency of the candidate models, but not model accuracy. If a model has high accuracy for accelerated test conditions but low accuracy for the operation condition, this model may still get selected using the proposed method. Unless real-world test data is available, model performance evaluation under operating condition is a difficult task. However, the proposed methodology offers an initial model selection procedure when there is no data under operating condition. The chosen model could then be updated based on available operating data.

Compared to previously developed model selection methods, the

proposed approach combining Bayesian calibration, model reliability metric, and the consistency criterion is found to be effective in RUL model selection. Unavailability of calibration/validation data under normal operating conditions is a common problem in many engineering applications. The proposed methodology provides a viable recourse for model selection when empirical data under normal operating conditions is not available.

CRediT authorship contribution statement

Dongjin Du: Conceptualization, Investigation, Methodology, Software, Visualization, Writing - original draft. **Pranav Karve:** Conceptualization, Methodology, Validation, Writing - review & editing. **Sankaran Mahadevan:** Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

This research was funded by the U.S. Department of Transportation (PHMSA Project No. 693JK32050006CAAP, Project Monitor: Zhongquan Zhu, PI: Prof. Jinying Zhu at the University of Nebraska-Lincoln). The support is gratefully acknowledged.

References

[1] Plastic Pipe Data Collection Initiative Status Report August 2019, *Plastic Pipe Database Committee*, 2019.
[2] J.R. White, Polymer ageing: physics, chemistry, or engineering? Time to reflect, Compt. Rendus Chem. 9 (11–12) (2005) 1396–1408, https://doi.org/10.1016/j.crci.2006.07.008.
[3] G. Pinter, R.W. Lang, Effect of stabilization on creep crack growth in high-density polyethylene, Appl. Polym. Sci. 90 (2003) 3191–3207, https://doi.org/10.1002/app.12944.
[4] International Organization for Standardization, ISO 9080: Plastics Piping and Ducting Systems — Determination of the Long-Term Hydrostatic Strength of Thermoplastics Materials in Pipe Form by Extrapolation, 2012. Retrieved from:iso.org/standard/43860.html.
[5] American Society for Testing and Materials, Basis for Thermoplastic Pipe Materials or Hoop Stress Design Basis for Thermoplastic Pipe Products, ASTM International, West Conshohocken, PA, 2021, https://doi.org/10.1520/D2837-21. ASTM D2837-21: Standard Test Method for Obtaining Hydrostatic Design.
[6] S. Zha, H.Q. Lan, H. Huang, Review on lifetime predictions of polyethylene pipes: limitations and trends, Int. J. Pres. Ves. Pip. 198 (102663) (2022), https://doi.org/10.1016/j.ijpvp.2022.104663.
[7] B.H. Choi, Z. Zhou, A. Chudnovsky, S.S. Stivala, K. Sehanobish, C.P. Bosnyak, Fracture initiation associated with chemical degradation: observation and modeling, Int. J. Solid Struct. 42 (2) (2005) 681–695, https://doi.org/10.1016/j.ijsolstr.2004.06.028.
[8] M.R. Contino, L. Andena, Environmental stress cracking of high-density polyethylene under plane stress conditions, Eng. Fract. Mech. 241 (2021), 107422, https://doi.org/10.1016/j.engfracmech.2020.107422.
[9] American Institute of Aeronautics and Astronautics, AIAA-G-077-1998: AIAA Guide for the Verification and Validation of Computational Fluid Dynamics Simulations, 1998, https://doi.org/10.2514/4.472855.001. Reston, vol. A.
[10] American Society of Mechanical Engineers, ASME Standard V&V 10-2006: Guide for Verification and Validation in Computational Solid Mechanics, 2006. New York, NY. Retrieved from, https://www.asme.org/codes-standards/find-codes-standards/v-v-10-standard-verification-validation-computational-solid-mechanics.
[11] Y. Ling, S. Mahadevan, Quantitative model validation techniques: new insights, Reliab. Eng. Syst. Saf. 111 (2013) 217–231, https://doi.org/10.1016/j.ress.2012.11.011.
[12] I.J. Myung, The importance of complexity in model selection, J. Math. Psychol. 44 (1) (2000) 190–204, https://doi.org/10.1006/jmps.1999.1283.
[13] V. Hombal, S. Mahadevan, Model selection among physics-based models, J. Mech. Des. 135 (2013), 021003, https://doi.org/10.1115/1.4023155.
[14] H. Akaike, Information Theory and an Extension of the Maximum Likelihood Principle, 2nd International Symposium on Information Theory, Akademiai Kiado, 1973.
[15] G.E. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[16] J. Rissanen, Modeling by shortest data description, Automatica 14 (5) (1977) 465–471, https://doi.org/10.1016/0005-1098(78)90005-5.
[17] R. Rebba, S. Mahadevan, Computational methods for model reliability assessment, Reliab. Eng. Syst. Saf. 93 (8) (2018) 1197–1207, https://doi.org/10.1016/j.ress.2007.08.001.
[18] B.D. Coleman, Application of the theory of absolute reaction rates to the creep failure of polymeric filaments, J. Polym. Sci. 20 (1956) 447–455, https://doi.org/10.1002/pol.1956.120209604.
[19] N. Brown, Fundamental mechanism of slow crack growth in semi-crystalline polymers under a constant load, Mater. Sci. Appl. 10 (2019) 721–731, https://doi.org/10.4236/msa.2019.1011052.
[20] N. Brown, Y.L. Huang, The effect of molecular weight on slow crack growth in linear polyethylene homopolymers, J. Mater. Sci. 23 (1988) 3648–3655, https://doi.org/10.1007/BF00540508.
[21] C. Bragaw, The Forecast of Polyethylene Pipe and Fitting Burst Life Using Rate Process Theory, Plastic Pipes Conference Association, York, 1982.
[22] A. Haldar, S. Mahadevan, Probability, Reliability, and Statistical Methods in Engineering Design, John Wiley & Sons, Inc, 1999.
[23] L.A. Kerr, D.R. Goethel, Stock Identification Methods, second ed., Academic Press, 2014, pp. 501–533, https://doi.org/10.1016/B978-0-12-397003-9.00021-7.
[24] S. Riedmaier, B. Danquah, B. Schick, F. Diermeyer, Unified framework and survey for model verification, validation and uncertainty quantification, Arch. Comput. Methods Eng. (2020) 1–34, https://doi.org/10.1016/j.ress.2012.11.011.
[25] P. Gardner, C. Lord, R.J. Barthorpe, An evaluation of validation metrics for probabilistic model outputs, in: In Verification and Validation, 2018, 4079 (V001T06A001), 1-34.
[26] S. Ferson, W.L. Oberkampf, L. Ginzburg, Model validation and predictive capability for the thermal challenge problem, Comput. Methods Appl. Mech. Eng. 197 (29–32) (2008) 2408–2430, https://doi.org/10.1016/j.cma.2007.07.030.
[27] D. Wu, Z. Wu, Y.W. Lu, C. Lei, Model validation and calibration based on component functions of model output, Reliab. Eng. Syst. Saf. 140 (2015) 59–70, https://doi.org/10.1016/j.ress.2015.03.024.
[28] M. Hansen, B. Yu, Model selection and the principle of minimum description length, J. Am. Stat. Assoc. 96 (454) (2001) 746–774, https://doi.org/10.1198/016214501753168398.
[29] D.M. Endres, &J.E. Schindelin, A new metric for probability distributions, IEEE Trans. Inf. Theor. 49 (7) (2003) 1858–1860, https://doi.org/10.1109/TIT.2003.813506.
[30] K. Krishnaswamy, R. Sukhadia M, A. Lamborn, M. J, IS PENT A TRUE INDICATOR OF PE PIPE SLOW CRACK GROWTH RESISTANCE ? Chevron Phillips Chemical Company, 2017.
[31] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equations of state calculations by fast computing machines, J. Chem. Phys. 21 (6) (1953) 1087–1092, https://doi.org/10.1063/1.1699114.
[32] K. Kengo, Stability of Markov Chain Monte Carlo Methods, first ed., Springer Tokyo, 2023.
[33] J.K. Kruschke, Doing Bayesian Data Analysis, second ed., Elsevier Inc, 2015.
[34] Rate Process Method for Projecting Performance of Polyethylene Piping Components, Plastic Pipe Institute, 2008.