TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
2020, Vol. 2674(9) 867–877
© National Academy of Sciences:
Transportation Research Board 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0361198120932567
journals.sagepub.com/home/trr
SAGE

Developing an Automated Technique to Calibrate the AASHTOWare Pavement ME Design Software

Shuvo Islam[1], Avishek Bose[2], Christopher A. Jones[1], Mustaque Hossain[1], and Cristopher I. Vahl[3]

Abstract

Many state highway agencies are in the process of implementing the AASHTOWare Pavement ME Design (PMED) software for routine pavement design. However, a recurring implementation challenge has been the need to locally calibrate the software to reflect an agency's design and construction practices, materials, and climate. This study introduced a framework to automate the calibration processes of the PMED performance models. This automated technique can search PMED output files and identify relevant damages/distresses for a project on a particular date. After obtaining this damage/distress information, the technique conducts model verification with the global calibration factors. Transfer function coefficients are then automatically derived following an optimization technique and numerical measures of goodness-of-fit. An equivalence statistical testing approach is conducted to ensure predicted performance results are in agreement with the measured data. The automated technique allows users to select one of three sampling approaches: split sampling, jackknifing, or bootstrapping. Based on the sampling approach chosen, the automated technique provides the calibration coefficients or suitable ranges for the coefficients and shows the results graphically. Model bias, standard error, sum squared error, and *p*-value from the paired *t*-test are also reported to assess efficacy of the calibration process.

The National Cooperative Highway Research Program (NCHRP) developed a user-friendly procedure for executing mechanistic-empirical (M-E) pavement design while accounting for local environmental conditions, local highway materials, and actual highway traffic distribution using axle load spectra (*1*). Another NCHRP study was conducted to investigate the M-E pavement design procedure and underlying assumptions, evaluate engineering reasonableness, and implement the M-E approach (*2*). The outcomes of this effort were included in a software (version 1.1) and a performance prediction model calibration guide. Since its inception, the software underwent several improvements (versions) and was later adopted by the American Association of State Highway and Transportation Officials (AASHTO) and incorporated into the AASHTOWare Pavement ME Design (PMED) software. AASHTOWare Pavement ME version 2.5 is the latest (as of 2019) version of this software.

Transfer functions in the PMED software were calibrated using Long-Term Pavement Program (LTPP) sections as a representative database of in-service pavement sections in North America (*3*). The nationally calibrated prediction models may not be accurate for routine pavement design for a particular state or region, since design and construction practices, pavement materials, and climatic conditions vary throughout the country. Thus, prediction models in the PMED software must be calibrated for a specific state or region. The local calibration guide for the PMED software, developed under the NCHRP project 1-40B, outlined an 11-step procedure for local calibration (*4*). These steps include selecting hierarchical levels of input parameters, developing an experimental template, estimating the sample size, selecting roadway segments, extracting and evaluating distress data, verifying and calibrating PMED models, reducing standard error, and deciding on the adequacy of calibration parameters.

Many highway agencies have conducted verification and local calibration of the PMED performance models.

[1]Department of Civil Engineering, Kansas State University, Manhattan, KS
[2]Department of Computer Science, Kansas State University, Manhattan, KS
[3]Department of Statistics, Kansas State University, Manhattan, KS

Corresponding Author:
Shuvo Islam, sislam@ksu.edu

However, software improvements or updates in the performance models necessitate recalibration of the PMED transfer functions (*5*). Furthermore, since new performance data and test results are constantly becoming available, performance models in PMED must be continually validated to see if recalibration is necessary (*6*).

Therefore, this study developed a Python-based script to automate the computational process of the PMED model calibration system. In addition to the traditional split sampling approach, the jackknifing and bootstrapping techniques were also incorporated into the automated technique to verify PMED models. Furthermore, an equivalence testing approach was studied to ensure PMED-predicted performance results compare favorably with actual performance data.

Need Statement

Following the release of MEPDG in 2004, several states tried to implement this software for routine pavement design. From its inception, the software underwent several improvements that necessitated frequent software recalibration (*5*). Changes in agency construction and design policy may also require recalibration of the performance models. Many highway agencies also reported long-term concerns in relation to continuing resource allocation for PMED model recalibration (*7*).

Development of an automatic calibration process for hydrological models has been rigorously investigated, and several automated global search algorithms were identified that have been developed for rainfall-runoff models designed to locate the global optimum on a response surface with numerous local optima (*8*). Madsen et al. explained that, in automated calibration, parameters are adjusted automatically according to a specified search scheme and numerical measures of goodness-of-fit (*9*). The specified search scheme refers to the minimization of a certain objective function of a dataset using a suitable deterministic algorithm or optimization technique. Fernández further elaborated that the automated calibration process involves minimizing the objective function of a dataset by adjusting calibration parameters using a deterministic algorithm under a set of constraints (*10*). Development challenges, however, include data availability, availability of a robust and physically meaningful calibrated model, and computational feasibility.

Objective

The key objective of this study was to develop an automated technique to help highway agencies conduct periodic in-house calibration of the PMED software. The necessary steps to accomplish this objective included forming a programming routine to extract required information from the software outputs, defining a clear objective function and optimization technique for parameter adjustments, incorporating several sampling techniques to validate the calibrated PMED models, and conducting suitable statistical testing so that the predicted performance data are in agreement with the measured data.

Methodology

The primary goal of the automated calibration process is to employ an optimization technique to determine calibration parameters that minimize the model bias and standard error of estimate (S_e). This study deployed a systematic method to automatically search for the optimal calibration parameters in prediction model transfer functions.

Objective Function

In the PMED performance model calibration process the objective function was defined as the mathematical equation that minimizes the total sum squared error between measured and predicted distress values and International Roughness Index (IRI), and as given in Equation 1.

$$\text{minimize}\, F(x) = \sum\nolimits_{i=1}^{n} \left[f(x(\varnothing_{\text{Local}})) - y_{\text{obs}}\right]^2 \qquad (1)$$

where

$F(x) =$ Objective function that needs to be minimized;

$f(x(\varnothing_{\text{Local}})) =$ Pavement ME predicted distresses or IRI for i^{th} data point;

$\varnothing_{\text{Local}} =$ Set of locally calibrated coefficients;

$y_{\text{obs}} =$ Measured distresses or IRI for i^{th} data point; and

$n =$ Total number of data points.

Optimization Technique

The automated calibration process uses a deterministic algorithm or optimization technique to obtain a set of locally calibrated coefficients that minimize the objective function. Several methods can solve the optimization problem in Equation 1. These methods generally iterate on x in some manner; an initial value of each parameter for $F(x)$ is chosen, the objective function is computed, and an algorithm is applied to generate a new x that will reduce the objective function (*11*).

This study utilized two optimization techniques, Polak and Ribière conjugate gradient (CG) and limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method, to calibrate the PMED models (*7*). CG is a nonlinear gradient-based optimization method used to obtain local calibration coefficients. This method minimizes objective function $F(x)$ by applying an iterative line search strategy (*10*). The line search strategy finds a descent direction along which $F(x)$ will be reduced and then computes a step size that determines how far x should move in that direction (*12*). The CG technique is computationally inexpensive and converges quickly.

The L-BFGS optimization technique is an iterative method for solving unconstrained nonlinear optimization problems. L-BFGS optimization technique is recommended for bounded problems with large sample sizes, such as 1,000 bootstrap samples. The L-BFGS uses a limited amount of computer memory and is one of the fastest algorithms for parameter estimation in machine learning problems (*13*). In this study, the global coefficients of the PMED models were used as the initial seed values.

Resampling Techniques

The local calibration guide for the PMED software recommended using resampling techniques to improve prediction model accuracy. The local calibration guide listed traditional split sampling and jackknifing resampling approaches for reliable assessment of prediction accuracy. However, the guide recommended using the jackknifing approach for a small sample size of the calibration dataset.

Brink calibrated and validated PMED performance models in Michigan using five types of sampling techniques: no sampling, traditional split sampling, repeated split sampling, jackknifing, and bootstrapping (*14*). He concluded that PMED models calibrated using the bootstrapping sampling technique consistently showed lower standard error and bias. This study incorporated traditional split sampling, jackknifing, and bootstrapping resampling approaches into the automated calibration technique.

Traditional Split Sampling Approach. Traditional split sampling is the most common resampling approach used in PMED model calibration studies. This approach requires that sample projects be randomly split into two subsets (*15*). One set is used to calibrate the performance models, and the other set validates the accuracy of the calibrated models. Researchers have typically used a 70–30 split or 80–20 split of data to calibrate and validate MEPDG transfer functions. The local calibration guide for the PMED software asserts that the traditional split sampling approach can produce misleading indications of model accuracy for small sample sizes (*15*).

Jackknifing Approach. The jackknifing method is an iterative process that uses systematic sampling to adjust calibration coefficients of the PMED performance models. Samples are selected by taking the original data vector and systematically deleting one observation from the dataset. One of the n selected projects is left out, and calibration is performed for the n–1 samples in each iteration. The prediction and standard errors can then be computed from the omitted sample. Thus, if n pavement sections are in the dataset, the jackknife sampling technique will consist of n samples, each with n–1 data points in each sample subset analysis (*16*). The process of omitting, calibrating, and predicting is repeated until the total dataset is used for prediction. As a result, n values of standard error are generated, and jackknife goodness-of-fit statistics can be computed.

The NCHRP 1-40B project recommends the jackknifing method for small sample sizes. According to the report, this approach produces reliable assessments of PMED models in relation to prediction accuracy and goodness-of-fit statistics (*15*).

Bootstrapping Approach. The bootstrapping method is a resampling technique used to compute statistics of a population by sampling a dataset with replacement. For example, for a dataset of N samples or pavement sections, X bootstrap samples of size N are randomly selected with replacement from the original dataset. Each X bootstrap sample omits several sections and creates multiple copies of other sections, allowing a given observation or pavement section to be included in the given sample size more than once (*14*). The number of bootstrap repetitions, X, should be large enough to ensure meaningful statistics can be computed, but the number of samples can only reduce the effects of random sampling errors and not increase the amount of information in the original data.

For AASHTOWare PMED calibration, the bootstrapping method can find a distribution of optimized coefficients of the performance models.

Equivalence Testing

Previous local calibration studies of the AASHTOWare PMED software primarily used a traditional equivalence testing or a paired t-test to determine bias. The local calibration guide of the PMED software recommends using the paired t-test to evaluate significant differences between measured and predicted data. A known limitation of the paired t-test, however, is that when the null hypothesis states the true effect size is zero, the absence of an effect can be rejected but yet not statistically supported (*17*). In other words, the paired t-test can successfully conclude if the measured and predicted distress data differ, but it cannot confirm whether these data sets are the same regardless of the statistical p-value.

Traditional Equivalence Testing. The null and alternate hypotheses for a traditional equivalence testing are shown in Equations 2 and 3, respectively.

$$\text{Null hypothesis,} \quad Ho : \mu_{\text{measured}} = \mu_{\text{predicted}} \qquad (2)$$

$$\text{Alternate hypothesis,} \quad H_1 : \mu_{\text{measured}} \neq \mu_{\text{predicted}} \qquad (3)$$

where

μ_{measured} = Sample mean of measured distress values or IRI data; and

$\mu_{\text{predicted}}$ = Sample mean of PMED-predicted distress values or IRI data.

In a traditional equivalence testing set up, these two means are assumed to be similar if the statistical *p*-value is larger than a threshold value, often 0.05, or at a 95% confidence level (i.e., the null hypothesis cannot be accepted). Walker and Nowacki argued that the burden of proof for statistical equivalence is on the wrong hypothesis (i.e., that of a difference) (*18*). In this method, a significant result establishes a difference, whereas a non-significant result implies only that equivalency or equality cannot be ruled out. Consequently, the risk of incorrectly concluding equivalence can be very high. They also argued that since no margin of equivalence is considered in a paired *t*-test, the concept of equivalence is not well defined.

Two One-Sided t-Test. Although the null hypothesis cannot be supported when the true effect size is zero, large effects can be rejected in a frequent hypothesis testing framework in equivalence testing. A two one-sided *t*-test (TOST) is a simple equivalence approach that specifies an upper (Δ_U) and lower (Δ_L) equivalence bound based on an equivalence margin, δ (*17*). In TOST procedure, equivalence is established at the α significance level if a (1–2α) × 100% confidence interval for the difference in efficacies (new—current) is contained within the interval ($\Delta_U = \delta$, $\Delta_L = -\delta$). The null and alternate hypotheses for the TOST equivalence testing are stated in Equations 4 through 9.

$$\text{Null Hypothesis,} \quad Ho : \mu_{\text{measured}} - \mu_{\text{predicted}} \leqslant -\delta \qquad (4)$$

$$\text{Alternate hypothesis,} \quad Ha : \mu_{\text{measured}} - \mu_{\text{predicted}} > -\delta \qquad (5)$$

and

$$\text{Null Hypothesis,} \quad Ho : \mu_{\text{measured}} - \mu_{\text{predicted}} \geqslant \delta \qquad (6)$$

$$\text{Alternate hypothesis,} \quad Ha : \mu_{\text{measured}} - \mu_{\text{predicted}} < \delta \qquad (7)$$

Determination of equivalence margin δ is a critical step in equivalence testing because δ directly affects the outcome because of a small value making it difficult to establish equivalence (*18*). In the TOST procedure, accuracy of optimized calibration coefficients depends on how well the equivalence margin can be established in relation to relevant evidence and engineering considerations. If the *p*-value is smaller than a threshold (e.g., 0.05), then the difference between two samples is smaller than the thresholds given by the equivalence margin.

Since no established equivalence margin was available for the PMED models, this study used a heuristic method to set an equivalence margin, δ. First, an initial value of δ was assumed to verify PMED models with the global coefficients. This initial value of δ for a model was assumed to be the same as the limiting S_e estimate recommended by AASHTO (*15*). After model calibration, TOST was repeated multiple times to compute the minimum δ value for which the equivalence between measured and predicted distress values can be established. This approach of setting the equivalence margin helped determine the narrowest equivalence region for the measured and predicted distress data. The initial values of δ assumed for PMED model verification are as follows: (1) Asphalt concrete (AC) rutting (0.1 in.); (2) Top-down cracking (600 ft/mile); (3) Transverse cracking (250 ft/mile), and (4) IRI (17 in/mile).

Programming Routine for the Automate Calibration Technique

The Python 3.7 programming language and Python "xlrd" library was used in this study to read and format information from the Excel files (e.g., PMED outputs and distress data). The Python "NumPy" library was used for scientific computing within the programming routine. PMED model parameters were calibrated using the optimization package within the "SciPy" library. The CG and L-BFGS techniques were used for optimization of transfer function parameters.

To calibrate PMED performance models using the automated technique, the user must first successfully run the PMED software. One of the major challenges in automating the PMED software calibration process is to recognize relevant software outputs for a project at a specific time. PMED software generates several Excel, pdf, and text files, so the developed automated calibration technique can search PMED output files and identify mechanistic damages and distresses for a project on a specific date. The user must browse the directory where PMED software outputs are stored.

The automated technique requires inputs, such as number of projects, measured distresses, and corresponding distress collection dates. Currently, inputs are provided in a tab-delimited text file. After running the automated technique, the user is prompted to select a

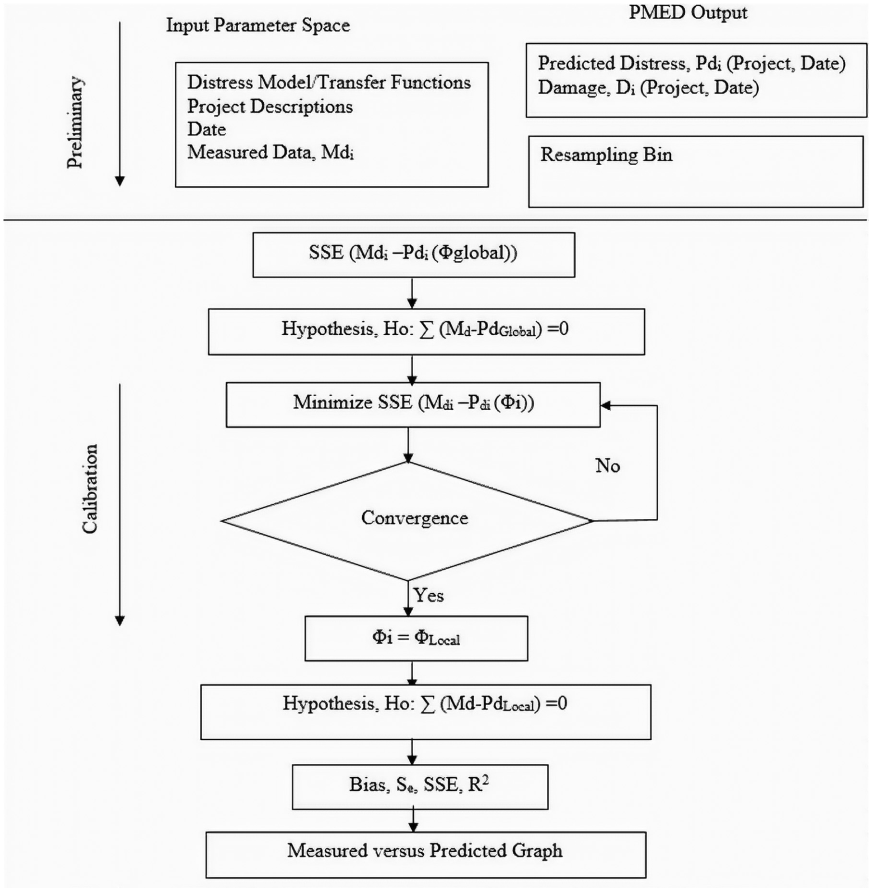
Preliminary
Input Parameter Space
Distress Model/Transfer Functions
Project Descriptions
Date
Measured Data, Md_i
PMED Output
Predicted Distress, Pd_i (Project, Date)
Damage, D_i (Project, Date)
Resampling Bin
SSE (Md_i –Pd_i (Φglobal))
Hypothesis, Ho: $\sum$ (M_d-Pd_{Global}) =0
Calibration
Minimize SSE (M_{di} –P_{di} (Φi))
Convergence
No
Yes
Φi = Φ_{Local}
Hypothesis, Ho: $\sum$ (Md-Pd_{Local}) =0
Bias, S_e, SSE, R^2
Measured versus Predicted Graph

Figure 1. Analysis process for automated calibration technique.
Note: PMED = Pavement ME Design; R^2 = coefficient of determination; S_e = standard error of estimate; SSE = sum squared error.

model for calibration and choose a sampling approach, such as split sampling, jackknifing, or bootstrapping.

After obtaining relevant damage data from the PMED output files, the application predicts distresses with a set of global model calibration coefficients (Φ_{Global}) to verify the prediction model. A paired *t*-test is then conducted to determine the initial bias between actual data and the AASHTOWare PMED software-predicted values at 95% confidence level. TOST is also conducted, and the coefficient of determination (R^2), S_e, and sum squared error (SSE) are recorded. In addition, the ratio of S_e and standard deviation of the measured data (S_y) is calculated. An S_e/S_y ratio greater than one indicates that the variability in residual error is larger than variability in the measured data; and an S_e/S_y ratio less than one indicates that the variability in residual error is smaller than in the measured data. The second scenario is preferred for each distress model calibration (*19*).

The CG or L-BFGS optimization technique is applied to the calibration dataset to obtain a set of model calibration coefficients (Φ_{Local}) that minimizes the SSE between measured and predicted distresses. Bias, R^2, S_e, and SSE are generated for the calibrated model. Equivalence testing is also conducted for the calibrated model, and the application generates graphs for the measured versus predicted distress data. Figure 1 presents the key steps in the automated calibration process.

In the split sampling approach, the automated technique randomly divides the projects into calibration and validation sets. To implement the jackknife approach for *n* number of projects, the procedure first removes one pavement section from the *n* projects, and the model calibration is performed with the remaining (*n–1*) set of projects. This process is repeated *n* times for all pavement sections in the database, allowing a distribution of calibration parameters to be obtained to help users select appropriate calibration coefficients. The automated calibration technique considers an 80–20 split of the data for calibration and validation. The "Scikit-Learn" machine learning library in the Python programming environment was used to incorporate the traditional split sampling and jackknifing approach, and the CG technique was used for parameter optimization for both cases. To implement the bootstrap approach for *N* number of projects, *X* bootstrap samples of size *N* are randomly selected with replacement. Each *B* sample omits several sections and creates multiple copies of other sections (*14*). Model calibration is performed for each *X* sample, and a parameter distribution is obtained.

PMED Inputs

AC overlay models were calibrated using the automated technique developed in this study for the Kansas Department of Transportation (KDOT) highway network. Twenty five AC over AC sections were selected in consultation with KDOT. The project selection criteria include: (i) projects where the existing highway segment construction occurred in 1995 or later to ensure the availability of the required PMED inputs; (ii) projects where less than five overlays had occurred during the life of the existing pavement structure; (iii) projects that contained a minimum of 1-inch hot-mix asphalt (HMA) overlay on top of the existing HMA pavement; and (iv) projects that encompassed all six districts of KDOT. The general features of the selected projects can be found elsewhere (*5*).

Calibration of PMED software required developing of a database for input parameters. In this study, Level 2 truck traffic volumetric factors were developed from data collected at 11 automatic vehicle classification (AVC) stations and traffic load spectra from 10 weigh-in-motion (WIM) stations. Project-specific AC volumetrics, for example, percent air void, void in mineral aggregates, and void filled with asphalt information were extracted from the KDOT construction management system (CMS) database. Project-specific binder grade information and mixture aggregate gradation were used to generate AC mechanical properties. Modern-era retrospective analysis for research and application (MERRA) files were used for climatic inputs. Pre-overlay falling weight deflectometer (FWD) data were available for 16 projects to characterize the existing AC damage. Pre-overlay pavement condition data, that is, rut depth, transverse cracking, top-down cracking, and IRI values were also available for the selected 25 sections.

KDOT Distress Data

KDOT began collecting automated pavement distress data on the entire road network using the laser crack measurement system (LCMS) from 2013. LCMS-generated automated distress data have primarily been used in this study. However, for pavement sections overlaid before 2013, manually collected distress data were used. The LCMS-generated cracking data were interpreted and quantified following the AASHTO Standard PP 67-16 and AASHTO Standard 68-14 (*20*, *21*). Following the recommendation of the NCHRP 1-40 B study, the magnitudes of time-series distress and IRI data were evaluated and inspected for outliers.

Local Calibration

The automated calibration technique was used to calibrate PMED performance models for the selected AC over AC sections. Calibrated models include the AC rutting model, transverse cracking model, top-down cracking model, and the IRI model. Because of brevity, only transverse cracking and permanent deformation models are discussed here.

Transverse Cracking Model

The PMED software computes transverse cracking as the sum of AC thermal cracking and reflection cracking. KDOT collects full-lane-width transverse cracking data for AC pavements but does not distinguish between thermal and reflection cracks. In this study, global factors were used for the AC thermal cracking model and the transverse cracking model was calibrated by adjusting the reflection cracking model coefficients only.

The first step was to verify the model with global coefficients. The automated calibration technique reported bias, SSE, *p*-value from the paired *t*-test, and the *p*-value from TOST. Table 1 lists summary statistics of the transverse cracking model with the global coefficients. Results in the table show that *p*-value from the TOST procedure was less than 0.05 for an equivalence margin 250 ft/mile, indicating that the PMED-predicted and measured transverse cracking were equivalent for a margin of 250 ft/mile. The S_e was 188 ft/mile, which was also within the

Table 1. Summary Statistics for the Transverse Cracking Model

Coefficients	SSE	S_e	S_e/S_y	*p*-value from paired *t*-test	*p*-value from TOST (δ =250 ft/mile)
Global	3,995,594	188	0.95	<0.001 (rejected)	<0.001
Local	1,006,772	99	0.50	0.42 (failed to reject)	<0.001

Note: SSE = sum squared error; TOST = two one-sided *t*-test; S_e = standard error of estimate; S_y = standard deviation of the measured data.

Table 2. Calibrated Model Equivalence Margin for Measured and Predicted Transverse Cracking

Equivalence margin, δ (ft/mile)	*p*-value	Remark
250	<0.001	Equivalent
150	<0.001	Equivalent
100	0.002	Equivalent
55	0.03	Equivalent
50	0.06	Not equivalent

AASHTO-suggested range of 250 ft/mile. However, the *p*-value from the paired *t*-test was less than 0.05, indicating that predicted and measured transverse cracking differ.

The automated calibration technique was used to optimize the reflective cracking model coefficients C_4 and C_5. The calibration parameters C_4 and C_5 can be optimized outside the PMED environment if the damage ratio (D) and percent reflective cracking rate (RCR) values are available. PMED software generates a text file "transverseReflectiveCracking.log," that contains time-series D and RCR values. The developed automated technique has the ability to extract the required time-series D and RCR values for each project at the corresponding date. The automated technique then uses the optimization approach to determine optimized values of C_4 and C_5. Results are provided for the traditional split sampling, jackknifing, and bootstrapping sampling approaches.

Traditional Split Sampling Approach. Table 1 presents the summary statistics of the transverse cracking model with the local coefficients. The null hypotheses for equivalence testing are presented in Equation 9. Results in the table show that *p*-value from the TOST procedure was less than 0.05 for an equivalence margin of 250 ft/mile, indicating that the PMED-predicted transverse cracking and measured transverse cracking were equivalent for a margin of 250 ft/mile. The *p*-value from the paired *t*-test after local calibration is greater than 0.05, which suggests that there is no evidence that predicted and measured transverse cracking are different. A comparison of values with the global and local coefficients in Table 1 shows that transverse cracking model goodness-of-fit statistics improved significantly after local calibration.

After calibration, the TOST procedure was repeated to determine the narrowest equivalence region for measured and predicted transverse cracking. Table 2 presents a summary of the repeated TOST procedure for the data used in the calibration dataset. In all cases, the level of significance (α) was 0.05. Results in Table 2 show that the calibrated transverse cracking model predictions for the AC over AC sections were equivalent to measured transverse cracking for a margin of 55 ft/mile.

For the validation dataset, the *p*-value from the TOST procedure was higher than 0.05 for an equivalence margin of 55 ft/mile, which suggests that predicted and measured transverse cracking were not equivalent for a margin of 55 ft/mile for projects in the validation dataset. Repeated TOST procedure was performed again for the validation dataset, and the narrowest equivalence margin was 85 ft/mile.

The measured versus predicted transverse cracking with global and local coefficients are shown in Figure 2. A comparison of Figure 2, *a* and *b*, show that the measured versus predicted transverse cracking with the local coefficients demonstrated improved data location relative to the line of equality.

Jackknifing Approach. The CG optimization method was used to optimize the C_4 and C_5 coefficients of the transverse cracking model for all AC over AC sections in this study. Figure 3, *a* and *b*, show the distribution of C_4 and C_5 coefficients, respectively, using the jackknife sampling approach. Figure 3*a* shows that the range of C_4 was 251–253 for 16 jackknife samples, and Figure 3*b* shows that the value of coefficient C_5 ranged from –2.45 to –2.58 for 18 jackknife samples.

The TOST procedure was repeated for a combination of C_4 and C_5, with C_4 in the range of 251–253 and C_5 ranging from –2.45 to –2.58. Results show that the calibrated transverse cracking predictions for the AC over AC sections were equivalent to measured transverse cracking for a margin of 55 ft/mile for all considered C_4 and C_5 combinations. This indicates that the PMED transverse cracking predictions (level of significance, α = 0.05) would likely be equivalent to the transverse cracking measured in the field for a margin of 55 ft/mile if the C_4 and C_5 coefficients range from 251 to 253 and –2.45 to –2.58, respectively.

Bootstrapping Approach. The C_4 and C_5 coefficients of the transverse cracking model were optimized using the bootstrapping approach. One thousand bootstrap samples were used optimized using the L-BFGS optimization technique. Figure 4, *a* and *b*, show the distribution of C_4 and C_5 coefficients using the bootstrap approach. Figure 4*a* shows that the range of coefficient C_4 was 266–268 for approximately 600 bootstrap samples. In fact, coefficient C_4 was 260–269 for almost every bootstrap sample. Figure 4*b* shows that coefficient C_5 ranged from –2.75 to –2.15 for all bootstrap samples.

TOST was repeatedly conducted for a combination of C_4 and C_5, with C_4 in the range of 260–269 and C_5 ranging from –2.75 to –2.15. Results showed that PMED transverse cracking predictions (level of significance, α = 0.05) would likely be equivalent to transverse cracking measured in the field for a margin of 53 ft/mile if the

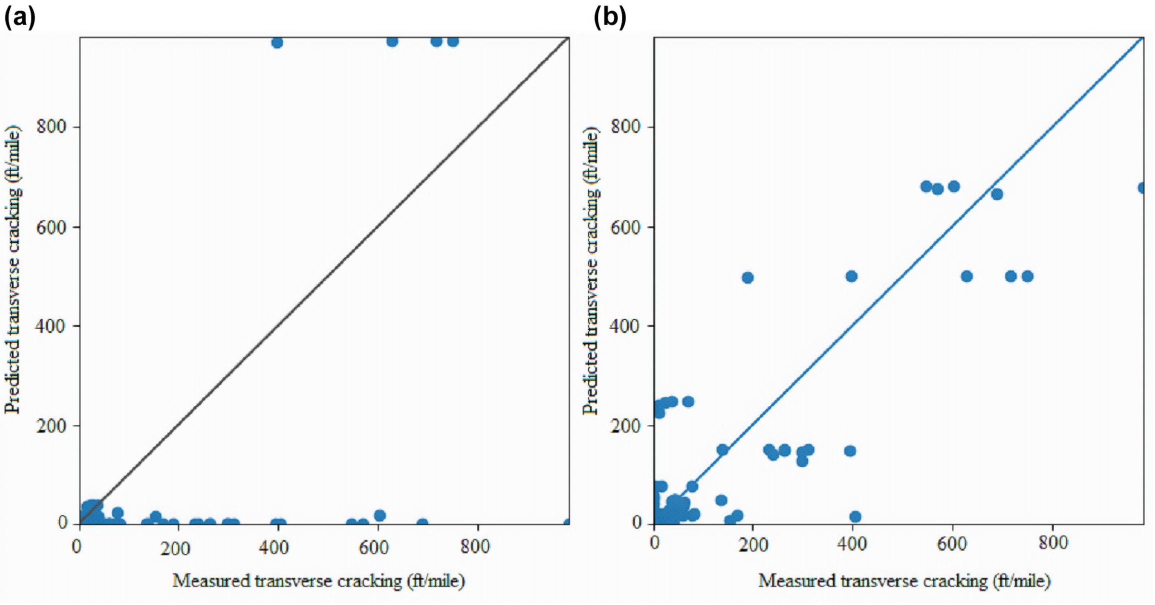

Figure 2. Predicted versus measured transverse cracking with: (*a*) global and (*b*) local coefficients.

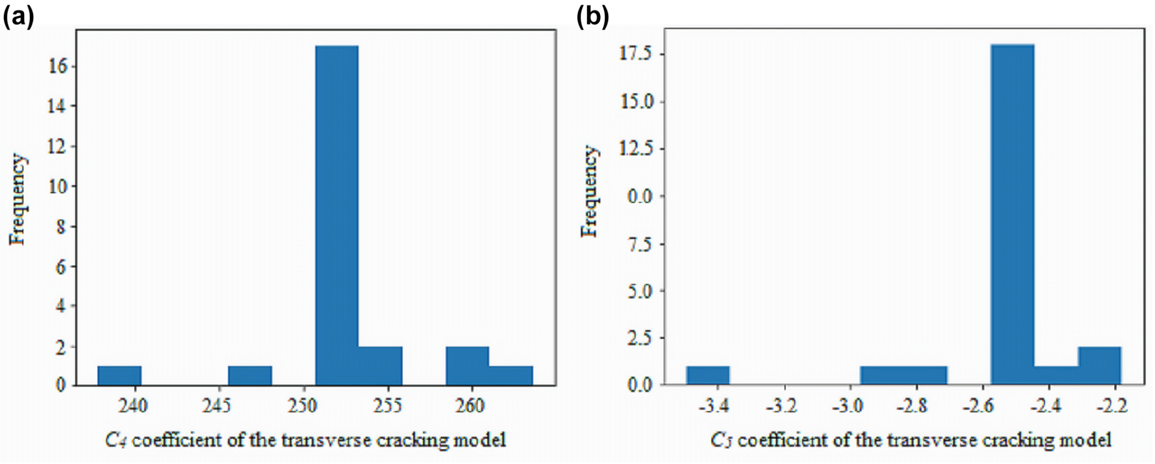

Figure 3. Transverse cracking model parameter: (*a*) C_4 and (*b*) C_5 distribution using the jackknife sampling approach.

C_4 and C_5 coefficients range from 260 to 269 and –2.15 to –2.75, respectively.

Permanent Deformation Model

The developed automated technique was used to calibrate the permanent deformation model for the AC over AC sections. Table 3 lists the summary statistics of the rutting model with the global and local coefficients using the split sampling technique. The null and alternate hypotheses for equivalence testing are presented in Equation 9. Only the β_{1r} parameter of the rutting model was optimized using the automated technique. A comparison of values with the global and local coefficients in Table 3 shows that rutting model goodness-of-fit statistics improved significantly after local calibration. The A S_e/S_y ratio greater than one, even after local calibration, indicates the variability in residual error is larger than the variability in the measured rut depth.

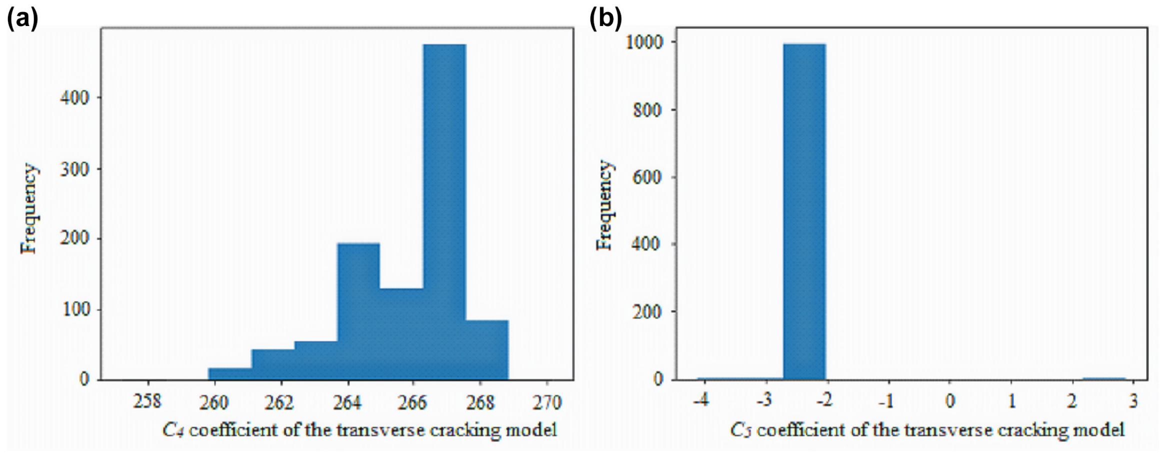

Figure 4. Transverse cracking model parameter: (*a*) C_4 and (*b*) C_5 distribution using the bootstrap sampling approach.

Table 3. Summary Statistics for the Locally Calibrated Rutting Model

Coefficients	SSE	S_e	S_e/S_y	*p*-value from paired *t*-test	*p*-value from TOST (δ = 0.1 in.)
Global	0.33	0.054	1.28	0.027 (rejected)	<0.001
Local	0. 25	0.046	1.07	0.24 (failed to reject)	<0.001

Note: SSE = sum squared error; TOST = two one-sided *t*-test; S_e = standard error of estimate; S_y = standard deviation of the measured data.

After model calibration, the TOST procedure was repeated multiple times to determine the narrowest equivalence region for measured and predicted rut depths. The repeated TOST procedure showed that the calibrated rutting model predictions were equivalent to the measured rut depths for a margin of 30 mils.

The distribution of the β_{1r} coefficient is shown in Figure 5 for the jackknife and bootstrap approaches. Figure 5*a* shows that the β_{1r} coefficient ranged from of 0.290–0.325 except for one instance. The TOST procedure was conducted multiple times for the value of β_{1r} in the range of 0.290–0.325, and the narrowest equivalence margin δ was determined in each case. Results showed that the PMED-predicted rutting and measured rutting were always equivalent for a margin 30 mils. The equivalence margin was lowest (20 mils) for a β_{1r} value of 0.325. Figure 5*b* demonstrated that for approximately 990 bootstrap samples, the β_{1r} coefficient was in the range of 0.24–0.41. Repeated TOST was conducted for the β_{1r} value ranging from 0.24 to 0.41. Results showed that the equivalence margin was lowest (0.012 in.) for a β_{1r} value of 0.36, indicating that the equivalence margin for the PMED-predicted and measured rut depths was 12 mils when the rutting model coefficient β_{1r} value was 0.36.

Conclusion

The primary objective of this study was to develop an automated technique for highway agencies to perform in-house calibration of PMED performance models. Three sampling approaches, traditional split, jackknifing, and bootstrapping, were incorporated into the automated technique. Both jackknifing and bootstrapping approaches provided ranges for the calibration coefficients. The range of adjusted coefficients was wider when the bootstrapping was used, potentially because of the differences in sample size (1,000 bootstrap samples versus 25 jackknife samples). The width of the range is dependent on the variability of the measured distress data across pavement sections. Since the bootstrapping technique omits several sections and creates multiple copies of others for each of the 1,000 samples, a large variation in distress values from project to project would generate a wider range of calibration coefficients.

The automated calibration technique conducts a paired *t*-test and TOST. Although the traditional paired *t*-test can confirm differences between two datasets, it cannot support agreement among them. TOST can establish equivalence between two sets of data for a predetermined margin. This study repeatedly conducted TOST for the calibrated model to identify an equivalence margin.

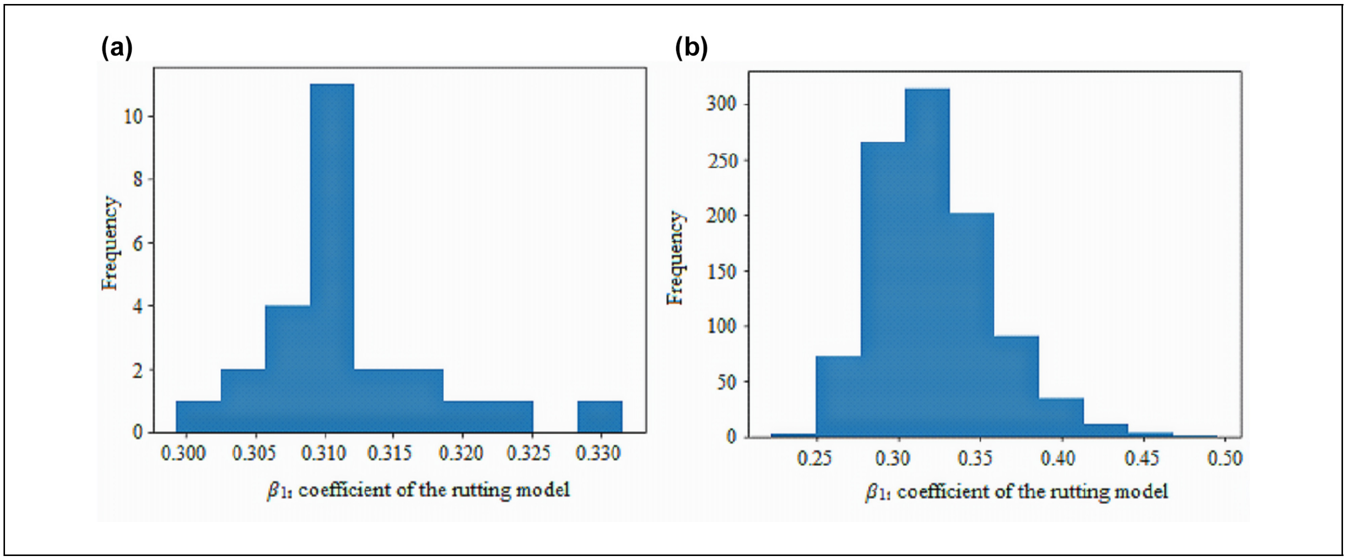

Figure 5. Rutting model β_{1r} parameter distribution using: (*a*) jackknife and (*b*) bootstrap resampling approach.

Repeated TOST resulted in an equivalence margin of 30 mils for the calibrated rutting model. The traditional split sampling and jackknifing approaches for the transverse cracking model for AC over AC sections resulted in the narrowest δ of 55 ft/mile, whereas the bootstrapping technique demonstrated the narrowest δ of 53 ft/mile.

Limitations and Possible Improvements

The following limitations were found in this study and recommendations have been developed to overcome these limitations:

- The automated technique developed in this study cannot calibrate parameters that require multiple simulations of the PMED software. For AC pavements, the β_{2r} and β_{3r} coefficients of the permanent deformation model and the thermal cracking model cannot be calibrated using this application. Wojtkiewicz et al. used the DAKOTA software and the Cygwin package to run PMED models numerous times outside of the AASHTOWare environment to investigate the effect of variability in HMA mixes on pavement performance prediction (*22*). The automated calibrated technique developed in this study could be integrated into the DAKOTA software and Cygwin package to automate PMED model calibrations that require multiple PMED simulation.
- This study conducted TOST to calibrate PMED models with an equivalence margin concept. A heuristic method was used to determine the narrowest equivalence margin for the calibrated PMED models. Further research must be conducted to establish a realistic equivalence region for different distress types and models. PMED models then should be calibrated to achieve the equivalence margin.
- The optimization technique employed in this study searches for a global optimum value for the calibration parameter for which the objective function (SSE) value is the smallest. While looking for the optimum model coefficient, unrealistic values for calibration parameters may be obtained, potentially minimizing the SSE, but may not be meaningful from an engineering point of view. Limited study has been conducted to date to identify lower and upper bounds of the PMED model coefficients beyond realistic predictions. Further research is needed to establish typical bounds for these calibration parameters.
- This study used the CG and L-BFGS optimization techniques to adjust PMED model parameters. Researchers previously have often used GRG and GA optimization techniques for the PMED calibration coefficients (*19*, *23*). Other robust optimization techniques, such as derivative-free methods and multi-objective and multi-constrained search strategies, may be implemented to determine the best possible calibration coefficients for the PMED models.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: S. Islam, M. Hossain, C. Jones, C. Vahl; data collection: S. Islam, A. Bose; analysis and interpretation of results: S. Islam, M. Hossain, C. Jones, C. Vahl; draft manuscript preparation: S. Islam and M. Hossain. All authors

reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Kansas Department of Transportation (KDOT).

References

1. NCHRP Project 1-37A. *Development of the 2002 Guide for the Design of New and Rehabilitated Pavement Structures. Design guide and Supplemental Documentation*. Transportation Research Board of the National Academies, Washington, D.C, 2004.
2. Ceylan, H., S. Kim, O. Kaya, K. Gopalakrishnan, and D. Ma. *Investigation of AASHTOWare Pavement ME Design/ DARWin-ME Performance Prediction Models for Iowa Pavement Analysis and Design*. Trans Project 14-496. Iowa Department of Transportation, Ames, 2015.
3. Robbins, M. M., C. Rodezno, N. Tran, and D. Timm. *Pavement ME Design–A Summary of Local Calibration Efforts for Flexible Pavements*. NCAT Report No.17-07. National Center for Asphalt Technology, Auburn, Alabama, 2017.
4. Von Quintus, H., M. I. Darter, and J. Mallela. *NCHRP Report 1-40B: Local Calibration Guidance for the Recommended Guide for Mechanistic-Empirical Design of New and Rehabilitated Pavement Structures*. Transportation Research Board of the National Academies, Washington, D.C., 2005.
5. Tran, N., M. M. Robbins, C. Rodezno, and D. Timm. *Pavement ME Design–Impact of Local Calibration, Foundation Support, and Design and Reliability Thresholds*. NCAT Report 17-08. National Center for Asphalt Technology, 2017.
6. Islam, S., M. Hossain, M. C. A. Jones, A. Bose, R. Barrett, and N. Velasquez. Implementation of AASHTOWare Pavement ME Design Software for Asphalt Pavements in Kansas. *Transportation Research Record: Journal of the Transportation Research Board*, 2019. 2673: 490–499.
7. Islam, S. *Implementation of AASHTOWare Pavement ME Design Software for Pavement Rehabilitation*. PhD dissertation. Kansas State University, Manhattan, 2019.
8. Duan, Q., S. Sorooshian, and V. Gupta. Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models. *Water Resources Research*, Vol. 28, No. 4, 1992, pp. 1015–1031.
9. Madsen, H. Automatic Calibration of a Conceptual Rainfall–Runoff Model using Multiple Objectives. *Journal of Hydrology*, Vol. 235, No. 3–4, 2000, pp. 276–288.
10. Fernández, B. *Automated Calibration for Numerical Models of Riverflow*. MS thesis. Institute for Modeling Hydraulic and Environmental Systems, University of Stuttgart, Germany, 2016.
11. Adams, B. M., S. Ebeida, M. S. Eldred, G. Geraci, J. D. Jakeman, K.A. Manupin, J. A. Monschke, J. AdamStephens, L. P. Swiler, D. M. Vigil, T. M. Wildey, W. J. Bohnhoff, K. R. Dalbey, J. P. Eddy, J. R. Frye, R. W. Hooper, K. T. Hu, P. D. Hough, M. Khalil, E. M. Ridgway, J. G. Winokur, and A. Rushdi. *DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.0 User's Manual*. Tech. Rep. SAND 2014-4633. Sandia National Laboratories, 2018.
12. Shewchuk, J. R. *An Introduction to the Conjugate Gradient Method without the Agonizing Pain*. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1994.
13. Mokhtari, A., and A. Ribeiro. RES: Regularized stochastic BFGS algorithm. *IEEE Transactions on Signal Processing*, Vol. 62, No. 23, 2014, pp. 6089–6104.
14. Brink, W. C. *Use of Statistical Resampling Techniques for the Local Calibration of the Pavement Performance Prediction Models*. PhD dissertation. Michigan State University, 2015.
15. *Guide for the Local Calibration of the Mechanistic - Empirical Pavement Design Guide*. American Association of State Highway and Transportation Officials, Washington, D.C., 2010.
16. Nisbet, R., J. Elder, and G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press, Canada, 2009.
17. Lakens, D. Equivalence Tests: A Practical Primer for *t*-Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, Vol. 8, No. 4, 2017, pp. 355–362.
18. Walker, E., and A. S. Nowacki. Understanding Equivalence and Noninferiority Testing. *Journal of General Internal Medicine*, Vol. 26, No. 2, 2011, pp. 192–196.
19. Kim, Y. R., F. M. Jadoun, T. Hou, and N. Muthadi. *Local Calibration of the MEPDG for Flexible Pavement Design*. Report No. FHWA\ NC\ 2007-07. NC Department of Transportation Research and Analysis Group, North Carolina Department of Transportation, Raleigh, 2011.
20. AASHTO PP 67-16. *Quantifying Cracks in Asphalt Pavement Surfaces from Collected Images Utilizing Automated Methods*. American Association of State Highway and Transportation Officials, Washington, D.C., 2016.
21. AASHTO PP 68-14. *Collecting Images of Pavement Surfaces for Distress Detection*. American Association of State Highway and Transportation Officials, Washington, D.C., 2016.
22. Wojtkiewicz, S. F., L. Khazanovich, G. Gaurav G., and R. Velasquez. Probabilistic Numerical Simulation of Pavement Performance using MEPDG. *Road Materials and Pavement Design*, Vol. 11, No. 2, 2010, pp. 291–306.
23. Ayed, A., and S. Tighe. Local Calibration for Mechanistic-Empirical Design using Genetic Algorithm. *Proc., 2015 Conference and Exhibition of the Transportation Association of Canada*, Charlottetown PEI, Canada, 2015.