NCAT Report 17-07

PAVEMENT ME DESIGN – A SUMMARY OF LOCAL CALIBRATION EFFORTS FOR FLEXIBLE PAVEMENTS

By

Dr. Mary M. Robbins

Dr. Carolina Rodezno

Dr. Nam Tran, P.E., LEED GA

Dr. David H. Timm, P.E.

August 2017

National Center for Asphalt Technology

NCAT

at AUBURN UNIVERSITY

277 Technology Parkway ■ Auburn, AL 36830

PAVEMENT ME DESIGN – A SUMMARY OF LOCAL CALIBRATION EFFORTS FOR FLEXIBLE PAVEMENTS

Dr. Mary M. Robbins*
Research Engineer
Ohio Research Institute for Transportation and the Environment
(*Work completed while at National Center for Asphalt Technology)

Dr. Carolina Rodezno
Assistant Research Professor
National Center for Asphalt Technology

Dr. Nam Tran, P.E.
Associate Research Professor
National Center for Asphalt Technology

Dr. David H. Timm, P.E.
Brasfield and Gorrie Professor of Civil Engineering
Principal Investigator

Sponsored by
National Asphalt Pavement Association
State Asphalt Pavement Associations

August 2017

ACKNOWLEDGEMENTS

The authors wish to thank the National Asphalt Pavement Association and the State Asphalt Pavement Associations for sponsoring this research as part of the Optimizing Flexible Pavement Design and Material Selection research project and for providing technical review of this document.

DISCLAIMER

The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the sponsoring agencies, the National Center for Asphalt Technology or Auburn University. This report does not constitute a standard, specification, or regulation. Comments contained in this paper related to specific testing equipment and materials should not be considered an endorsement of any commercial product or service; no such endorsement is intended or implied.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Many state highway agencies have considered adopting the Mechanistic-Empirical Pavement Design Guide (MEPDG) and the accompanying AASHTOWare Pavement ME Design software to supplement or replace the empirical American Association of State Highway and Transportation Officials (AASHTO) Pavement Design Guides and the widely used DARWin pavement design software. In flexible pavement design, the software "mechanistically" calculates pavement responses (stresses and strains) based on the inputs and trial design information and uses those responses to compute incremental damage over time. It then utilizes the cumulative damage in transfer functions to "empirically" predict pavement distresses for each trial pavement structure.

The MEPDG was nationally calibrated using Long-term Pavement Program (LTPP) pavement sections as a representative database of pavement test sites across North America. While the resulting performance models are representative of national-level conditions, they do not necessarily represent construction and material specifications, pavement preservation and maintenance practices, and materials specific to each state. It is expected that such parameters would affect field performance. Therefore, local calibration studies should be conducted to address these differences and to adjust, if necessary, local calibration coefficients of the transfer functions used to predict pavement performance in the Pavement ME Design software to better reflect actual performance. Without properly conducted local calibration efforts, the implementation of MEPDG will not improve the pavement design process and may yield errors in predicted thickness of asphalt pavements.

Recognizing the importance of local calibration of flexible pavement performance models, this study was conducted to review the general approach undertaken for state highway agencies, the results of those efforts, and recommendations for implementing the nationally or locally calibrated models.

While it is often referred to as "local calibration," the process may include local verification, calibration, and validation of the MEPDG. As a minimum, the local verification process is needed to determine if state practices, policies, and conditions significantly affect design results. In this process, the distresses predicted by the Pavement ME Design software using the nationally calibrated coefficients are compared with measured distresses for selected pavement sections. If the difference between the predicted and measured distresses is acceptable to the agency, the Pavement ME Design can be adopted using the default models and coefficients; otherwise, it should be calibrated to local materials and conditions.

If local calibration is warranted, it is important that it addresses both the potential bias and precision of each transfer function in the Pavement ME Design software. Figure A illustrates how the bias and precision terms can be improved during the local calibration process. In practical terms, bias is the difference between the predicted distress at the 50% reliability level and the mean measured distress. Precision dictates how far the predicted values at a specified design reliability level would be from the corresponding predicted values at the 50% reliability

prediction. The locally calibrated models are then validated using an independent set of data. The models are considered successfully validated to local conditions if the bias and precision statistics of the models are similar to those obtained from model calibration when applied to a new dataset.

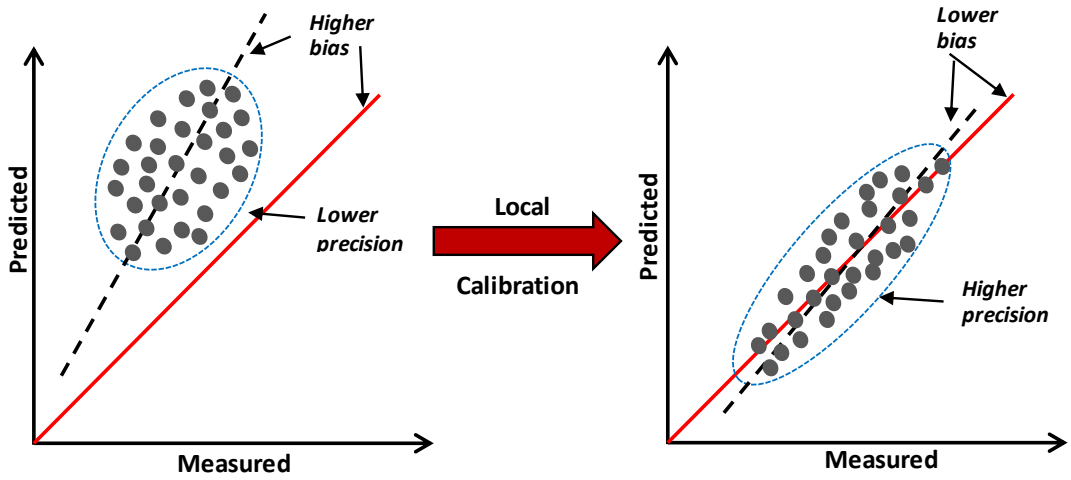

Figure A: Improvement of Bias and Precision through Local Calibration

While the AASHTO calibration guide details a step-by-step procedure for conducting local calibration, it was found that the actual procedures utilized vary from agency to agency, and in some ways deviated from the recommended procedures. This is partially due to the timing of the publication relative to the initiation of such efforts and the release of new versions of the software. This presents challenges for state agencies, as local calibration is a cumbersome and intensive process and the software and embedded distress models are evolving faster than local calibration can be completed. Therefore, it is recommended that ongoing and future calibration efforts are completed in accordance with the AASHTO calibration guide and current performance models and their coefficients are verified.

In reviewing calibration efforts for asphalt concrete (AC) pavements conducted across the country, it was found that calibration was typically attempted by looking at the predicted and measured distress for a set of roadway segments and reducing the error between measured and predicted values by optimizing the local calibration coefficients. However, other approaches were taken. Although a minimum number of roadway segments necessary to conduct the local calibration for each distress model is provided in the AASHTO calibration guide, the step for estimating sample size for assessing the distress models was not always reported. For those efforts that did report a sample size, some were smaller than the minimum amount recommended.

The AASHTO calibration guide recommends conducting statistical analyses to determine goodness of fit, spread of the data, as well as the presence of bias in the model predictions. Three hypothesis tests are recommended: 1) to assess the slope, 2) to assess the intercept of the measured versus predicted plot, and 3) a paired t-test to determine if the measured and predictions populations are statistically different. In a few cases local calibration efforts

included all three hypothesis tests for each performance model. However, some studies evaluated only one or two of the statistical tests and others relied only on qualitative comparisons of measured versus predicted distresses. When qualitative analyses were conducted, it was due to inadequate data. It is recommended that the calibration guide be followed to establish a dataset with adequate data necessary to conduct quantitative statistical analysis. Statistical parameters help to determine if local calibration has reduced bias and improved precision and if implementation is warranted. This will also help in identifying any weaknesses that may exist in the model that must be considered during the design process.

The table below denotes the number of verification, calibration, and validation efforts conducted for each performance model. In evaluating the nationally calibrated models, not all of the studies evaluated the presence of bias; however, for those that did, results varied by model and by study. Bias was most frequently reported for the nationally calibrated total rutting model. Although under-prediction was reported for some states, the majority reported that the default MEPDG model over-predicted total rutting. Consequently, this model was also the most calibrated, with all twelve verification efforts resulting in local calibration attempts. The longitudinal cracking model was found to have the poorest precision. Given the significant spread reported for the current default longitudinal cracking model, this model should not be used for design. It is anticipated that a new model will be developed under the ongoing NCHRP 1-52 project.

The results of the local calibration efforts are also summarized in the table. For asphalt pavements, the rutting model was the most commonly calibrated model. The transverse and longitudinal cracking models were calibrated the least. General improvements in predictions were realized with local calibration, however, the degree to which those improvements were made varied by state.

The intent of MEPDG and the AASHTOWare® Pavement ME software is to improve pavement design. Local calibration and validation of the performance models are essential to the implementation of this design framework. However, the software continues to evolve with future refinements of transfer models, such as the longitudinal cracking model, expected. Calibration efforts will also need to be completed as the use of unconventional materials becomes more commonplace. Therefore, it is expected that calibration efforts will be, in many ways, ongoing. Local calibration can be a time consuming and labor-intensive process; therefore, an agency may need to consider the implications of conducting local calibration efforts while the embedded models and software continue to be refined.

Table A Number of Verification/Calibration Studies and Summary of Calibration Results

Model	Verification	Calibration	Validation	Results of Calibration
Fatigue Cracking	[10] AZ, CO, IA, MO, NE states, NC, OR, UT, WA, WI	[6] AZ, CO, NE states, NC, OR, WA	[1] NC	• All seven efforts resulted in improvements in predictions. • Two studies (AZ, CO) resulted in sizeable increases in R^2 compared to R^2 in verification effort. Both studies had R^2 values much greater than the development of the global model (R^2 = 27.5%) but were only moderately high (50% and 62.7%). • Reduction or elimination of bias was reported in four (AZ, CO, NC, OR) of the seven studies. • One study (NE states) reported only the Sum of the Squared Error (SSE), which was reduced with calibration. • Two efforts (WA, WI) were qualitative analyses. Both resulted in predictions closely matching measured data.
Total Rutting	[12] AZ, CO, IA, MO, NE states, NC, OH, OR, TN, UT, WA, WI	[12] AZ, CO, IA, MO, NE states, NC, OH, OR, TN, UT, WA, WI	[2] IA, NC	• Generally, improvements in predictions were reported with calibrated models. • Four efforts (AZ, MO, NC, UT) resulted in an increase in R^2. Two efforts (CO, OH) saw decreases in R^2. • Overall, R^2 remained low for the efforts that reported it, ranging from 14.4% to 63%, with only one greater than the R^2 (57.7%) reported in the development of the default model. • Eight studies (AZ, IA, MO, NC, OH, OR, TN, UT) resulted in improvements in standard error of the estimate, S_e, while

				one study (CO) resulted in an increase in S_e. • Even though most saw improvements in standard error, S_e remained greater than 0.107, the S_e for the development of the default model, in four studies (AZ, CO, NC, OR). • Bias was eliminated or reduced in at least eight studies (AZ, CO, IA, MO, NC, OR, UT, WI). One effort (OH) showed bias remained despite calibration. • Three efforts (NE states, TN, WA) did not report on bias, but all four resulted in improvements in predictions.
Transverse Cracking	[10] AZ, CO, IA, MO, NE states, OH, OR, UT, WA, WI	[5] AZ, CO, MO, OR, WI	[0]	• Two studies (CO, MO) resulted in improvements in R^2 with both values (43.1% and 91%) greater than the R^2 reported in the development of the default model at a Level 1 analysis (34.4%). • Two calibration attempts (AZ, OR) were unsuccessful in improving transverse cracking predictions, and therefore, were not recommended for use. • Predictions were reasonable for one study (CO) with the elimination of bias. Two studies (MO, WI) resulted in good predictions with slight bias.
IRI	[10] AZ, CO, IA, MO, NE states, OH, TN, UT, WA, WI	[6] AZ, CO, MO, NE states, OH, WI	[1] IA	• Generally, improvements in IRI predictions were realized with locally calibrated models, especially for WI. • Three efforts (AZ, CO, OH) resulted in an improvement in R^2, ranging from 64.4% to

				82.2%, all of which were greater than the R^2 for the development of the default model (56%). • Only SSE was reported for one study (NE states), which indicated an improvement in predictions with the calibrated model • Bias was removed through three efforts (AZ, CO, WI), while the bias that remained in two efforts (MO, OH) was considered reasonable.
Longitudinal Cracking	[4] IA, NE states, OR, WA	[4] IA, NE states, OR, WA	[1] IA	• For two studies (IA, OR) the predictions and bias were reportedly improved, however, the S_e remained large in both calibrated models. • Despite improved predictions through calibration in one study (IA), the model was recommended for use only for experimental or informational purposes. • One study (NE states) only reported SSE, which was reduced with calibration, resulting in improved predictions. • One qualitative analysis (WA) was conducted and resulted in reasonable predictions.

1 INTRODUCTION

Many state highway agencies have considered adopting the *Mechanistic-Empirical Pavement Design Guide* (MEPDG) and the accompanying AASHTOWare Pavement ME Design software to supplement or replace the empirical American Association of State Highway and Transportation Officials (AASHTO) Pavement Design Guides and the widely used DARWin pavement design software (*1, 2*). A survey of state agencies conducted by Pierce and McGovern in 2013 indicated that 43 agencies were evaluating the MEPDG and 15 agencies planned to implement the new design procedure in the next two years (*3*). The implementation plans of these agencies have the following elements (*3*, *4*):

- *Scope of implementation.* The scope identifies the types of pavement designs (e.g., new construction or rehabilitation for asphalt and concrete pavements) that will be conducted using the Pavement ME Design software once implemented.
- *Design inputs.* This element describes the inputs required for the Pavement ME Design software and identifies the source of available information. This element also includes a plan for additional testing or analysis needed to provide critical inputs that are currently not available.
- *Local verification, calibration, and validation.* This element includes a plan for verifying the distresses predicted by the Pavement ME Design software with the distresses measured in the field for a number of representative pavement sections. If the predicted and measured distresses are reasonably comparable, the new design procedure can be adopted; otherwise, local calibration and validation are needed as recommended in the *AASHTO Guide for the Local Calibration of the MEPDG* (*5*).
- *Design thresholds and reliability levels.* This element provides proposed levels of distress, International Roughness Index (IRI), and reliability for acceptable pavement designs.
- *Software and design manuals.* These documents are prepared to meet the needs and design policies of the individual state agencies.
- *Training.* A training program may be developed in-house or through universities, consultants, or national programs to train the staff in pavement design.
- *Concurrent designs.* Before full implementation, concurrent designs are often conducted to compare the results of the Pavement ME Design with those of previous design procedures. This effort helps the staff become more familiar with the software and improve their confidence in the Pavement ME Design results.

To prepare for implementation, state agencies have conducted studies focusing on (1) building libraries for important input parameters; (2) conducting local verification, calibration, validation, and selecting design thresholds and reliability levels; and (3) preparing design manuals and materials for training staff.

Resources are available to help state highway agencies train staff and establish plans for building materials libraries. Through pooled fund studies and national implementation efforts led by the Federal Highway Administration (FHWA), state highway agencies can acquire testing equipment and set up an experimental plan to develop important material libraries by

themselves or through a research organization(s) in their states and access training materials online (*6, 7*).

The remaining implementation activities—conducting local verification, calibration, and validation and selecting design thresholds and reliability levels—often require more time and financial resources. While the AASHTO MEPDG calibration guide is available to help state agencies plan and conduct local calibration, it requires competent statistical and engineering knowledge to understand the concepts and to properly conduct local calibration. State agencies are also required to collect materials and pavement performance data for the local calibration. Thus, even though the local calibration process is needed for implementation, highway agencies are still reluctant to invest in this activity (*3*). Without properly conducting local calibration, the implementation of MEPDG will not improve the pavement design process. It has been reported that use of the globally calibrated Pavement ME Design software may yield thicker, overdesigned asphalt pavements (*8*, *9*).

Recognizing the importance of local calibration, this report discusses the general approach to local calibration undertaken for state agencies for asphalt concrete (AC) pavements, followed by results of the local calibration efforts and where possible, the recommendations for implementing the globally or locally calibrated models.

2 LOCAL VERIFICATION, CALIBRATION, AND VALIDATION

The MEPDG was developed to design new and rehabilitated pavement structures based on mechanistic-empirical principles. Basic steps of the Pavement ME Design are illustrated in Figure 1. Based on the inputs and trial design information, the Pavement ME Design software "mechanistically" calculates pavement responses (stresses and strains) and uses those responses to compute incremental damage over time. The program then utilizes the cumulative damage to "empirically" predict pavement distresses for each trial pavement structure. The mechanistic analysis utilizes the Enhanced Integrated Climatic Model (EICM), structural response models, and time-dependent material property models. The empirical analysis uses the transfer (regression) models to relate the cumulative damage to observed pavement distresses. While the mechanistic models are assumed to be accurate and to correctly simulate field conditions, inaccuracies still exist and affect transfer function computations and final distress predictions (*5*). The local verification, calibration, and validation process is often related to the transfer functions, but it essentially addresses the errors of both the mechanistic and empirical models.

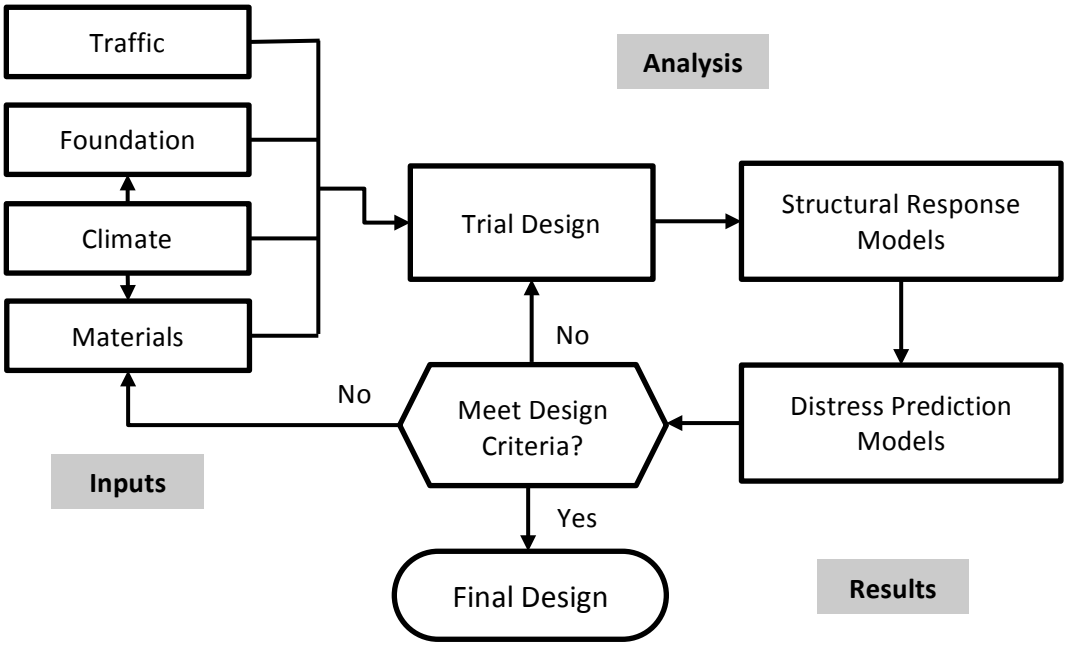

Figure 1 Basic Steps of Pavement ME Design

Under the NCHRP 1-37A and 1-40 projects, the MEPDG was "globally" calibrated using a representative database of pavement test sites across North America. Most of these test sites have been monitored through the Long Term Pavement Performance (LTPP) program. They were used because of the consistency in the monitored data over time and the diversity of test sections spread throughout North America. However, construction and material specifications, pavement preservation and maintenance practices, and materials and climatic conditions vary widely across North America. These differences can significantly affect distress and performance. However, they are not currently considered directly in the Pavement ME Design software but indirectly considered through local calibration in which the calibration coefficients of transfer functions in the ME Design software can be adjusted (*5*).

While it is often referred to as "local calibration", the process may include local verification, calibration, and validation of the MEPDG. As a minimum, the local verification process is needed to determine if state practices, policies, and conditions significantly affect design results. In this local verification process, the distresses predicted by the Pavement ME Design software using the globally calibrated coefficients are compared with the distresses measured in the field for selected pavement sections. If the difference between the predicted and measured distresses is not significant, the Pavement ME Design can be adopted; otherwise, it should then be calibrated to local conditions since these conditions were not considered in the global calibration process.

If local calibration is warranted, it is important that it addresses both the potential bias and precision of each transfer function in the Pavement ME Design software. Figure 2 illustrates how the bias and precision terms can be improved during the local calibration process. In practical terms, bias is the difference between the 50% reliability prediction and the measured

mean. Precision dictates how far the predicted values at a specified design reliability level would be from the corresponding predicted values at the 50% reliability prediction. The locally calibrated models are then validated using an independent set of data. The models are considered successfully validated to local conditions if the bias and precision statistics of the models are similar to those obtained from model calibration when applied to the validation dataset.

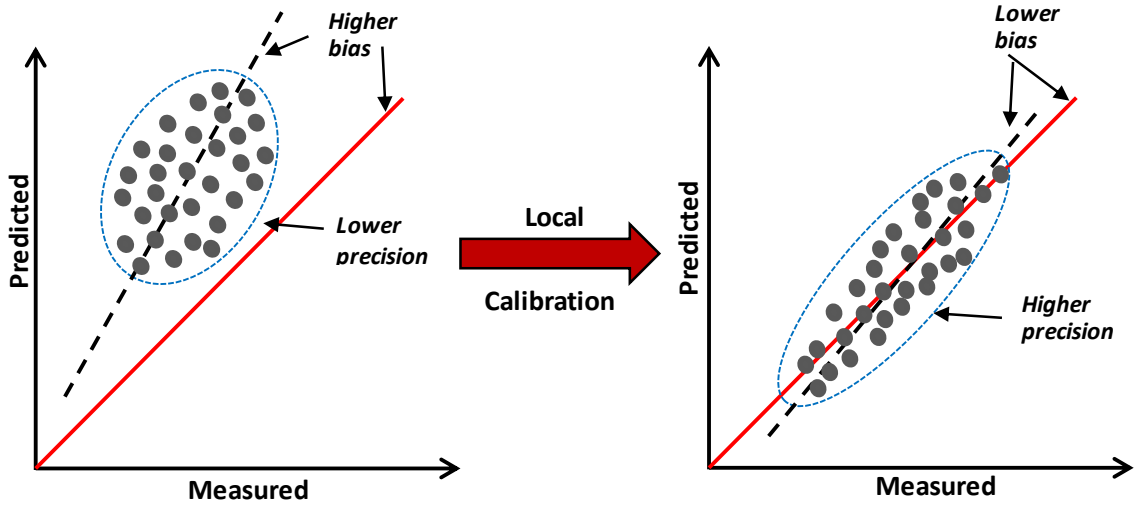

Figure 2 Improvement of Bias and Precision through Local Calibration

A detailed step-by-step procedure for local calibration is described in the *AASHTO Guide for the Local Calibration of the MEPDG* (*5*). The key steps of this procedure are as follows:

1. *Select hierarchical input level for each input parameter.* This is likely a policy-based decision that can be influenced by several factors, including the agency's field and laboratory testing capabilities, material and construction specifications, and traffic collection procedures and equipment. Agencies can refer to the MEPDG Manual of Practice (*4*) for recommendations on selecting the hierarchical input level for each input parameter.
2. *Develop experimental design.* An experimental plan or matrix is set up in this step to help select pavement segments that represent the pavement distresses observed in the state and local factors that may affect the observed distresses, such as the agency's design and construction practices and materials, as well as traffic and climatic conditions.
3. *Estimate sample size for assessing distress models.* This step is to estimate the number of pavement segments, including replicates, which should be included in the local calibration process to provide statistically meaningful results. The minimum number of pavement segments recommended for each distress model is as follows:
    - Total rutting: 20 roadway segments
    - Load-related cracking: 30 segments
    - Non-load related cracking: 26 segments
    - Reflection cracking (asphalt surface only): 26 segments

4. *Select roadway segments.* Appropriate roadway segments and replicates are identified in this step to satisfy the experimental plan developed in Step 2. The pavement segments selected are recommended to have at least three condition surveys conducted in the past 10 years.
5. *Extract and evaluate data*. The inputs available for each roadway segment are compiled and verified in this step. Data not compatible with the format required for the Pavement ME Design software will be converted accordingly. Missing data will be identified for further testing to be conducted in Step 6.
6. *Conduct field and forensic investigations of test sections.* This step encompasses field sampling and testing of the selected pavement segments to obtain missing data as identified in Step 5. The level of testing should be selected appropriately so that the data generated are compatible with the hierarchical input level selected in Step 1. Forensic investigations are necessary to confirm assumptions in the MEDPG, at the discretion of the agency. Investigations suggested include test cores, and trenching to identify location, initiation, and propagation of distresses in the pavement structure.
7. *Assess local bias.* The Pavement ME Design software with global calibration factors is conducted to design pavements using the inputs available from the selected pavement segments at 50% reliability. For each distress model, the predicted distresses are plotted and compared with the measured distresses for which linear regression is performed. Diagnostic statistics, bias, and the standard error of the estimate (S_e), are determined. Bias is determined by performing linear regression using the measured and MEPDG predicted distress and comparing it to the line of equality. Three hypotheses, listed below, are tested to determine if bias is present. If bias exists the prediction model should be recalibrated (see Step 8). If the difference is not significant, the standard error of the estimate is assessed (see Step 9).
    - Assess if the measured and predicted distress/IRI represents the same population of distress/IRI using a paired t-test.
    - Assess if the linear regression model developed has an intercept of zero.
    - Assess if the linear regression model has a slope of one.
8. *Eliminate local bias.* If significant bias exists (as determined in Step 7), the cause should be determined. Inputs that may cause prediction bias include traffic, climate, and material characteristics (*3*). If possible, the bias should be removed by adjusting the calibration coefficients listed in Table 1. Figure 3 illustrates basic steps for determining local calibration coefficients. Then, the same analysis conducted in Step 7 is performed using the adjusted calibration factors.
9. *Assess standard error of the estimate.* In this step, the S_e values determined in Step 7 or 8 based on the predicted and measured distresses (local S_e) are compared with the S_e values of the globally calibrated distress models provided in the Pavement ME Design software (global S_e). Models whose local S_e values are greater than the global S_e values should be recalibrated in an attempt to lower the standard error (see Step 10). For the other models, the local S_e values can be used for pavement design. The S_e values found to be reasonable based on the global calibration process are provided in Table 2 for reference.

10. *Reduce standard error of the estimate*. Table 1 lists the calibration coefficients that can be adjusted to reduce the standard error of the estimate for each distress model. If the S_e cannot be reduced, the agency can decide whether it should accept the higher local S_e or lower global S_e values for pavement design. This decision should take into account the difference in sample size used in the global and local calibration processes.
11. *Interpret the results and decide on the adequacy of calibration parameters.* The agency should review the results and check if the expected pavement design life is "reasonable" for the performance criteria and reliability levels used by the agency.

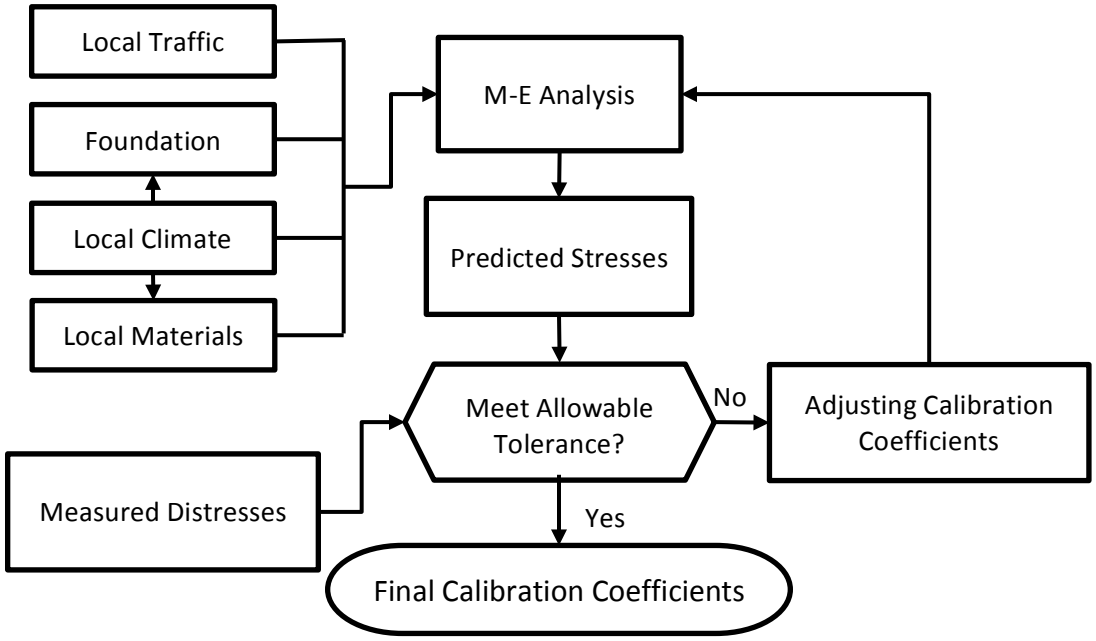

Figure 3 Local Calibration of Pavement ME Design

Table 1 Coefficients to be Adjusted for Eliminating Bias and Reducing Standard Error (*3, 5*)

Distress	Eliminate Bias	Reduce Standard Error
Total rut depth	k_{r1} = -3.35412 β_{r1} = 1 β_{s1} = 1	k_{r2} = 1.5606 k_{r3} = 0.4791 β_{r2} = 1 β_{r3} = 1
Alligator cracking*	k_{f1} = 0.007566 C_2 = 1	k_{f2} = -3.9492 k_{f3} = -1.281 C_1 = 1
Longitudinal cracking*	k_{f1} = 0.007566 C_2 = 3.5	k_{f2} = -3.9492 k_{f3} = -1.281 C_1 = 7
Transverse cracking	β_{t3} = 1 k_{t3} = 1.5	β_{t3} = 1 k_{t3} = 1.5
IRI	C_4= 0.015	C_1 = 40 C_2 = 0.4 C_3 = 0.008

* Coefficients are consistent with the Manual of Practice; see Appendix A.4 and A.5 for coefficients listed in the Pavement ME Design software

Table 2 Standard Error of the Estimate (*3, 5*)

Performance Prediction Model	Standard Error (S_e)
Total Rutting (in)	0.10
Alligator Cracking (% lane area)	7
Longitudinal Cracking (ft/mi)	600
Transverse Cracking (ft/mi)	250
Reflection Cracking (ft/mi)	600
IRI (in/mi)	17

3 SUMMARY OF METHODOLOGIES FOLLOWED FOR LOCAL CALIBRATION

The *Guide for the Local Calibration of the MEPDG* provides direction on the procedure for local calibration (*5*). Despite this AASHTO publication, many calibration efforts did not follow this procedure or the terminology used in the guide, in part due to the timing of the publication (2010) relative to the timing of calibration efforts in each state and the time dedicated to such efforts. Table 3 summarizes the verification efforts and calibration attempts for each of the twelve local or regional calibration studies included in this report, as well as the timing of each study. Some studies were published prior to and some after 2010, the year the *Guide for the Local Calibration of the MEPDG* was released. However, these studies often take several years to complete, beginning at least one to two years in advance of the actual date of publication.

Table 3 Summary of Verification/Calibration Efforts by State (*10-25, 27*)

State/Region	Sponsoring Agency	Year of Publication	Verification (V)/Calibration (C) Efforts									
			Fatigue Cracking		Rutting		Transverse Cracking		IRI		Longitudinal Cracking	
			V	C	V	C	V	C	V	C	V	C
AZ	AZ DOT*/ FHWA**	2014	✓	✓	✓	✓	✓	✓	✓	✓		
CO	CO DOT	2013	✓	✓	✓	✓	✓	✓	✓	✓		
IA	IA DOT	2013	✓		✓	✓	✓		✓		✓	✓
MO	MO DOT	2009	✓		✓	✓	✓	✓	✓	✓		
Northeastern States	NY State DOT	2011	✓	✓	✓	✓	✓		✓	✓	✓	✓
NC	NC DOT	2011	✓	✓	✓	✓						
OH	OH DOT	2009			✓	✓	✓		✓	✓		
OR	OR DOT/ FHWA	2013	✓	✓	✓	✓	✓	✓			✓	✓
TN	TN DOT	2013			✓	✓			✓			
UT	UT DOT	2009/2013	✓		✓	✓	✓		✓			
WA	WA State DOT	2009	✓	✓	✓	✓	✓		✓		✓	✓
WI	WI DOT/ FHWA	2009/2014	✓		✓	✓	✓	✓	✓	✓		

*DOT: Department of Transportation

**FHWA: Federal Highway Administration

In this section, the definitions of verification, calibration, and validations used in the AASHTO calibration guide are first presented. Next, the methodology used by each state is summarized. The results of the efforts by performance model are then presented in Section 4.

Verification: "Verification of a model examines whether the operational model correctly represents the conceptual model that has been formulated." It should also be noted that field data are not needed in the verification process, as it is "primarily intended to confirm the internal consistency or reasonableness of the model. The issue of how well the model predicts reality is addressed during calibration and validation" (*5*).

Calibration: "A systematic process to eliminate any bias and minimize the residual errors between observed or measured results from the real world (e.g., the measured mean rut depth in a pavement section) and predicted results from the model (e.g., predicted mean rut depth from a permanent deformation model). This is accomplished by modifying empirical calibration parameters or transfer functions in the model to minimize the differences between the predicted and observed results. These calibration parameters are necessary to compensate for model simplification and limitations in simulating actual pavement and material behavior" (*5*).

Validation: "A systematic process that re-examines the recalibrated model to determine if the desired accuracy exists between the calibrated model and an independent set of observed data. The calibrated model required inputs such as the pavement structure, traffic loading, and environmental data. The simulation model must predict results (e.g., rutting, fatigue cracking) that are reasonably close to those observed in the field. Separate and independent data sets should be used for calibration and validation. Assuming that the calibrated models are successfully validated, the models can then be recalibrated using the two combined data sets without the need for additional validation to provide a better estimate of the residual error" (*5*).

The process of calibration generally consists of three steps: (1) verification or evaluation of the existing global model to determine how well the model represents actual distresses and to evaluate the accuracy and bias; (2) calibration of the model coefficients to improve the model and reduce bias, typically using the same dataset as used in the verification step; and (3) validation of the newly calibrated model using a separate dataset. The AASHTO calibration guide specifically states that the verification procedure does not need to utilize field data to assess if the model is reliable and consistent (*5*). It is suggested that this should be addressed in the calibration and validation steps; however, it becomes rather confusing when reporting two sets of results (results for the statistical comparison with measured data for performance predicted using the nationally calibrated model and those results for the performance predicted by the locally calibrated model) in the calibration procedure. To distinguish between the various results reported for each calibration effort, the more commonly used terminology is utilized in this report such that: verification refers to the application of the globally calibrated model for the available data used in design and compared with actual field performance data to assess bias and accuracy; results reported under the calibration step are the results from the local calibration of the model coefficients and compared with the field performance data; validation refers to the application of the newly calibrated model to a new dataset (and field performance data), separate from the dataset used to calibrate the model. The following subsections summarize the methodology used in the calibration efforts documented in this study with more detailed summaries provided in Appendix B.

3.1 Methodology used in Efforts for Arizona (*10*)

Verification and local calibration efforts were completed by Applied Research Associates, Inc. (ARA) in a 2014 study sponsored by Arizona Department of Transportation (DOT) and FHWA. Researchers utilized DARWin ME (version of the software was not stated) for the study. The pavement sections in their study included new pavements (AC over granular layer, thin AC over jointed plain concrete pavement (JPCP)) and rehabilitated pavements (AC over AC and AC over JPCP) that covered northern, central, and southern regions with low and high elevations. The asphalt materials used in the study included conventional and Superpave mixtures with thicknesses above and below 8 inches. Material properties were characterized at different levels. For example, HMA creep compliance was at Level 1 while effective binder content was at Level 3. The base and subgrade materials were typically granular and coarse-grained, respectively.

3.2 Methodology used in Efforts for Colorado (*11*)

A 2013 study conducted by ARA utilized Version 1.0 of the MEPDG to complete verification and local calibration efforts for Colorado DOT. A variety of new and overlay asphalt mix sections were used with neat and modified binders. HMA layer thicknesses varied, but most of them were less than 8 inches. The climatic zones range from hot to very cool locations.

The asphalt material properties were characterized at Levels 2 or 3 depending on the information available. For example, HMA dynamic modulus used Level 2, but other volumetric properties used Level 3. MEPDG global models were calibrated using nonlinear model optimization tools (SAS statistical software). The criteria used for determining model adequacy for Colorado conditions are presented in Table 4.

Table 4 Criteria for Determining Models Adequacy for Colorado Conditions (*11*)

Criterion	Test Statistics	R^2 Range/Model SEE	Rating
Goodness of Fit	R^2, percent (for all models)	81-100	Very Good
		64-81	Good
		49-64	Fair
		<49	Poor
	Global HMA Alligator Cracking model SEE	<5 percent	Good
		5-10 percent	Fair
		>10 percent	Poor
	Global HMA Total Rutting model SEE	<0.1 in	Good
		0.1-0.2 in	Fair
		>0.2 in	Poor
	Global HMA IRI model SEE	<19 in/mi	Good
		19-38 in/mi	Fair
		>38 in/mi	Poor
Bias	Hypothesis testing-Slope of Linear measured vs. Predicted Distress/IRI model (β_1=slope) H_0:β_1=0	p-value	Reject if p-value is <0.05
	Paired t-test between measured and predicted distress/IRI	p-value	Reject if p-value is <0.05

3.3 Methodology used in Efforts for Iowa

An initial verification study was conducted in 2009 for Iowa DOT by researchers at the Center for Transportation Research and Education at Iowa State University. The study aimed to evaluate the HMA performance models embedded in the MEPDG software Version 1.0 (*12*). In this study, a Level 3 analysis was conducted, and predicted performance was compared with measured performance data for rutting and IRI. As a result, bias was reported for both predicted rutting and IRI.

Local calibration was conducted in 2013 in an Iowa DOT-sponsored study by researchers at the Institute for Transportation at Iowa State University. The study included a verification effort to evaluate and assess the bias associated with global performance models in Version 1.1 of the MEPDG software (*13*). Accuracy of the global performance models was evaluated by plotting the measured performance measures against the predicted performance measures and observing the deviation from the line of equality. Additionally, the average bias and standard error were determined and used to evaluate the nationally calibrated and locally calibrated models.

Local calibration was attempted for alligator (fatigue) cracking, rutting, thermal (transverse), IRI, and longitudinal cracking (*13*). For the calibration effort, a total of 35 representative HMA sections were chosen, one of which was an Iowa LTPP section. As part of the calibration effort, a sensitivity analysis was first conducted to understand the effect of each calibration coefficient on performance predictions and to more easily identify coefficients that should be optimized. New calibration coefficients were then determined through linear and non-linear optimization. Non-linear optimization was utilized for local calibration of the cracking and IRI performance models. To reduce the large number of computations associated with the trial-and-error procedure, linear optimization was chosen for fatigue, rutting, and thermal fracture models. Accuracy and bias of the recalibrated models were evaluated in the same manner as was done in assessing the global models. A portion of the dataset was reserved for use in validation of the recalibrated model coefficients (*13*).

3.4 Methodology used in Efforts for Missouri (*14*)

Verification and local calibration efforts were completed for Missouri DOT in a 2009 study conducted by ARA. Version 1.0 of the MEPDG was utilized for the study. The pavement sections used in the study included new or reconstructed HMA, HMA over HMA, and HMA over PCC with different thicknesses. Material properties were characterized at different levels depending on the information available. For example, dynamic modulus was at Level 2 and volumetric properties were at Level 1.

When possible, a statistical approach was taken for evaluating the nationally calibrated models and for local calibration. Model prediction capabilities were evaluated by comparing measured values with values predicted using the global calibration (default) coefficients. The coefficient of determination, R^2, and the standard error of the estimate, S_e, were used to compare the measured and predicted values. Using the same statistics for evaluation, R^2 and S_e, global coefficients were calibrated for Missouri materials and conditions. Bias was evaluated by first identifying the linear regression between measured values and distress predicted by the MEPDG. Once identified, the three hypothesis tests recommended in the Manual of Practice were carried out using a level of significance of 0.05, such that the rejection of any hypothesis indicates the model is biased. Local calibration resulted in changes to the thermal cracking, rutting, and IRI models.

A non-statistical approach was used when the measured distress/IRI was zero or close to zero for the sections under evaluation. Comparisons between predicted and measured distress/IRI

were conducted by categorizing them into groups. The evaluation consisted of determining how often measured and predicted distress/IRI remained in the same group. This is an indication of reasonable and accurate predictions without bias.

A second local calibration effort is currently underway with special emphasis on thin HMA overlays using reclaimed materials and other binder modifications.

3.5 Methodology used in Efforts for Northeastern States (*15*)

In a 2011 study conducted at the University of Texas at Arlington and sponsored by New York State Department of Transportation, calibration was attempted for the Northeastern states. For this effort, seventeen LTPP pavement sections in the northeastern (NE) region of the United States were selected to best represent conditions in New York State. LTPP sites from the followings states were chosen: Connecticut, Maine, Massachusetts, New Jersey, Pennsylvania, and Vermont. Using Version 1.1 of the MEPDG, verification was performed by executing the MEPDG models with the default, nationally calibrated coefficients, and by comparing the predicted distresses with the measured distresses for each model. Five models were evaluated: permanent deformation (rutting), bottom-up fatigue (alligator) cracking, top-down fatigue (longitudinal) cracking, smoothness (IRI) model, and transverse (thermal) cracking. These models, except for the thermal cracking model, were then calibrated.

3.6 Methodology used in Efforts for North Carolina

A 2008 study was conducted using MEPDG Version 1.0 to determine if the national calibration coefficients could capture the rutting and alligator cracking on North Carolina asphalt pavements (*16*). More recently, verification was conducted as part of a calibration effort completed in 2011 (*17*). Both studies were sponsored by North Carolina DOT and completed by researchers at North Carolina State University. The 2011 study used Version 1.1 of the MEPDG and completed local calibration for the permanent deformation (rutting), and alligator cracking model. Additionally, material-specific calibration was also conducted for the twelve most commonly used asphalt mixtures in North Carolina. The material-specific global field calibration coefficients, k_{r1}, k_{r2}, k_{r3}, in the rutting model, (as shown in equation A.1), and the fatigue model coefficients, k_{f1}, k_{f2}, k_{f3}, (as shown in equation A.12) were determined for each of the twelve asphalt mixtures. These material-specific calibration coefficients were used in the recalibration procedure for the local calibration coefficients in both the fatigue cracking and rutting models.

A total of twenty-two pavement sections were used for the calibration of the fatigue cracking and rutting models, while twenty-four pavement sections were used for validation of the recalibrated models (*17*). The level of inputs used in the software ranged from Level 2 for asphalt mixtures to Level 3 for the unbound materials. Two approaches were taken in performing local calibration. The first approach considered a large factorial of calibration coefficients (β_{r2} and β_{r3} for the rutting model and β_{f2} and β_{f3} for the alligator cracking model) and consisted of executing the software numerous times for each model whilst optimizing the remaining coefficients for each model. The second approach utilized a genetic algorithm optimization technique for each model to optimize the coefficients simultaneously. Model

adequacy for the global performance models and locally calibrated models (in both the calibration and validation stage) were evaluated with the coefficient of determination, standard error of the estimate, ratio of standard error of the estimate to the standard deviation of the measured performance, and the p-value for null hypothesis in which the average bias is zero at the 95% confidence level (*17*).

3.7 Methodology used in Efforts for Ohio (*18*)

A local calibration study was conducted for Ohio DOT in 2009 by researchers at ARA. The study included a verification effort to determine if the global models in Version 1.0 of the MEPDG software were sufficient in predicting performance for selected pavements in Ohio. Four performance models were evaluated: alligator cracking, rutting, transverse cracking, and IRI. Measured performance from thirteen LTPP new or reconstructed pavement sections in Ohio were used for the verification and calibration efforts. In simulating these sections in the MEPDG software, the hierarchical input levels varied. For example, dynamic modulus was entered at Level 2, while unit weight and volumetric properties of the asphalt mixtures were entered at Level 1. As part of the verification effort, the global performance models were evaluated for prediction capability, accuracy, and bias using the coefficient of determination, standard error of the estimate, and the three hypothesis tests recommended in the AASHTO calibration guide, respectively. The locally calibrated models were evaluated in a similar manner.

3.8 Methodology used in Efforts for Oregon (*19*)

Verification and local calibration studies using Darwin M-E (Version 1.1) were conducted for Oregon DOT in a 2013 study completed at by the Institute for Transportation at Iowa State University. The calibration was based on a Level 3 analysis. Pavement work conducted by Oregon DOT mainly involves rehabilitation of existing pavements; hence, calibration was conducted for rehabilitation of existing structures. Pavement sections were selected based on their location (Coastal, Valley, or Eastern), type (HMA over aggregate base, HMA inlay or overlay over aggregate base, HMA inlay or overlay over cement treated base, or HMA overlay of CRCP), traffic level (low or high), and pavement performance (very good/excellent, as expected, or inadequate).

3.9 Methodology used in Efforts for Tennessee (*20*)

In a 2013 study conducted at the University of Tennessee at Knoxville and sponsored by Tennessee DOT efforts were aimed at developing local calibration factors for Tennessee (*20*). In this study, an initial verification of the rutting and roughness models for new pavement design were evaluated and where applicable, local calibration was performed using Version 1.100 of the MEPDG software. Focus of the local calibration was placed on existing pavements that had received an overlay, as pavement rehabilitation was a large portion the pavement activities conducted in Tennessee. The nineteen pavement sections used for verification and eighteen sections used for local calibration were mostly Interstate pavements and consisted of AC pavements without an overlay, AC pavements with an AC overlay, and Portland cement concrete (PCC) pavements with an AC overlay. In the initial verification process, two hierarchical input levels were considered: “Level 1.5,” which looked at Level 1 for material

properties of AC layers and Level 2 inputs for the base and subgrade, and "Level 2.5," which consisted of Level 3 inputs for AC layers and Level 2 for base and subgrade properties. The roughness model was evaluated by considering roughness in terms of PSI, which is consistent with Tennessee DOT's method for characterization of roughness.

In evaluating and calibrating the rutting model, three different categories of pavements were considered: asphalt pavements and asphalt pavement overlaid with AC; concrete pavements overlaid with AC for low volume traffic (0-1,000 Average Annual Daily Truck Traffic (AADTT)); and concrete pavements overlaid with AC for heavy traffic (1,000-2,500 AADTT). In evaluating the PCC pavements with AC overlays, only rutting in the surface (AC) layer was considered. For AC pavements with AC overlays, the sections were treated as new asphalt pavements because asphalt overlays were not included in the national calibration of the rutting model. Therefore, new asphalt pavements and asphalt pavements with AC overlays were grouped together and total rutting (as opposed to rutting in the surface layer) was evaluated for that dataset. Validation of the locally calibrated rutting model was conducted for AC pavement sections overlaid with AC.

3.10 Methodology used in Efforts for Utah (*21, 22*)

Utah DOT sponsored a verification and local calibration study that was completed by ARA in 2009 using an early version, Version 0.8, of the MEPDG. The pavement sections in the study included new HMA and HMA over HMA with different thicknesses, but most of them were between 4-8 inches. Most of the material properties were characterized as Level 3 with the exception of the subgrade that used a Level 1 (backcalculated using deflection data). Local calibration was conducted using linear and non-linear regression procedures (SAS statistical software). Optimization was performed to select local calibration coefficients to maximize R^2 and minimize S_e, both goodness of fit and bias were checked, and a limited sensitivity analysis was performed.

The rutting models for all the layers were recalibrated in 2013 utilizing the test sections used in the 2009 local calibration (except for those that had been overlaid) and four more years of rutting data (2009-2012). The recalibration analysis was conducted in the same manner as in the 2009 local calibration. The main difference in coefficients is for the subgrade where the 2013 recalibration will yield lower subgrade rutting (*22*).

3.11 Methodology used in Efforts for Washington (*23*)

A study for Washington State DOT (WSDOT) was conducted through a joint effort with engineers at WSDOT and Applied Pavement Technology, Inc. in 2009 (*23*). Using Version 1.0 of the MEPDG, verification and local calibration efforts were completed for conditions specific to Washington State. It was reported that the calibration process follows a combination of a split-sample approach and a jackknife testing approach per recommendation in the draft report for NCHRP Project 1-40A (Recommended Practice for Local Calibration of the ME Pavement Design Guide). For the calibration procedure, data from the Washington State Pavement Management System (WSPMS) was used. The calibration efforts focused on fatigue damage, longitudinal

cracking, alligator cracking, and rutting models. First, an elasticity analysis was conducted to assess the sensitivity of each distress model to each of its calibration coefficients; the higher absolute elasticity value, the greater impact of the factor on the model. A set of sensitive calibration factors were then selected, and the design software was conducted by varying the calibration factors for two representative pavement sections, one each in eastern and western Washington. The predicted distresses were then compared with the measured distresses for the two sections. A set of acceptable calibration factors with the least root mean square errors (RMSE) was selected.

The calibrated models were then tested against each of the validation sections, which included five sections from a previous study, six representative sections from WSPMS for several iterations, and two sections in Washington State that had been used in the national calibration effort of the MEPDG. The calibration factors were slightly changed between the iterations to reduce the RMSE between the MEDPG prediction and WSPMS measurements. Final calibration factors with the least RMSE were selected and reported, but no statistics were presented. When asked to comment on WSDOT's local calibration efforts, a senior pavement engineer from WSDOT said, "The major distress on WSDOT's asphalt pavements is top-down cracking, which is not properly modeled in the MEPDG. In other words, the MEPDG cannot be successfully calibrated for WSDOT asphalt pavements unless the model is properly redeveloped."

3.12 Methodology used in Efforts for Wisconsin

A study on the Pavement ME Design local calibration was completed for Wisconsin DOT by ARA in 2009 (*24*), and the results were updated and published in a draft user manual completed in 2014 (*25*). The 2009 study was conducted using the LTPP sections in Wisconsin. Most of the inputs required for the verification and calibration efforts were from the LTPP database with some inputs using national defaults in the software. A design was conducted for each LTPP section to predict pavement distresses and IRI. The Pavement ME Design predictions were then compared with the measured distresses in the LTPP database to develop recommendations for Wisconsin DOT (*24*). Results of the model verification summarized in this report are from the 2009 report (*24*), and the updated model coefficients reported in this report are from the draft user manual (*25*), which does not include the model calibration statistics.

4 RESULTS OF VERIFICATION, CALIBRATION, AND VALIDATION EFFORTS

The previous section summarized local calibration studies conducted for numerous state highway agencies. This section presents the results of those studies divided according to major distress type. The distresses include fatigue cracking, rutting, transverse (thermal) cracking, and IRI. Summaries of the results of each individual local calibration effort can be found in Appendix C.

4.1 Summary of Verification and Calibration Results

The results of the previously discussed verification efforts have been summarized in the following subsections for each performance model: fatigue (alligator) cracking, total rutting,

transverse (thermal) cracking, IRI, and longitudinal (top-down) cracking. National calibration coefficients considered are noted, as well as any reported results of statistical comparisons with field measured data. The local calibration results are also tabulated, noting the new local calibration coefficients as well as the results of statistical comparisons with field measured data, where applicable.

4.2 Fatigue Cracking

Tables 5 and 6 summarize the results of verification and/or local calibration of the fatigue cracking model, respectively. Also listed in the tables are the calibration coefficients and statistical parameters reported for the development of the nationally calibrated model reported in the Manual of Practice. It should be noted, as shown in Appendix A.4, coefficients k_{f2} and k_{f3} for the global model are listed as negative values in the Manual of Practice but are shown as positive values in the Pavement ME Design software due to a slight change in the form of the equation.

4.2.1 Fatigue Cracking – Verification Results

Of the states investigated for this report, ten conducted verification exercises, such that predictions from the nationally calibrated fatigue model were compared with field-measured fatigue cracking (Table 5). Four states: Missouri, Utah, Washington and Wisconsin were unable to assess the nationally calibrated fatigue cracking model with statistical measures. The nationally calibrated model was found to predict the observed fatigue cracking reasonably well for Utah, while slight over and under-predictions were reported for Missouri. As a result, local calibration was not recommended for either state. Verification results for Washington showed the model tended to under-predict alligator cracking and therefore, local calibration was recommended. For Wisconsin, verification results indicated good predictions by the model, as most of the measured and predicted alligator cracking fell within the same category; thus, no local calibration was warranted.

In the verification process, the coefficient of determination, R^2, was only reported for two states, for which the highest R^2 was only 17.5%, well below the 27.5% reported in development of the global model. Another state simply reported it as “poor” and the remaining states did not report it at all. Although R^2 was not reported for Iowa’s and Wisconsin’s verification results, the nationally calibrated model resulted in good estimates of the measured fatigue cracking and did not require local calibration. Inadequate estimates with the nationally calibrated model were reported for Arizona, Colorado, and Washington. Verification efforts for those three states, as well as North Carolina, the Northeastern states, and Oregon, all showed the global model generally resulted in under-predictions of observed cracking.

Table 5 Summary of Verification Efforts for Fatigue (Alligator) Cracking Model (*4, 10, 11, 13, 15, 17, 19, 23, 24*)

Calibration Coefficient	Global Model (Manual of Practice)	AZ	CO	IA	NE States	NC	OR	WA	WI
k_{f1}	0.007566	0.007566	0.0076	NR	NR	0.0076	NR	NR	0.007566
k_{f2}	-3.9492	-3.9492	-3.9492			-3.9492			-3.9492
k_{f3}	-1.281	-1.281	-1.281			-1.281			-1.281
β_{f1}	1	1	1	1	NR	1	NR	1	1
β_{f2}	1	1	1	1		1		1	1
β_{f3}	1	1	1	1		1		1	1
C_1	1	1	1	1	1	1	1	1	1
C_2	1	1	1	1	1	1	1	1	1
C_4	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000	6,000
Statistical Parameters - Verification									
R^2, %	27.5	8.2	17.5	NR	NR	"poor"	NR	NR	NR
S_e	5.01	14.3	0.175	1.22		19.498	3.384		
N	405	363	50	327		124	NR		
p-value (paired t-test)	NR	NR	0.0059	NR		0.000	NR		
p-value (slope)	NR	NR	<0.0001	NR		NR	NR		

NR: Not Reported

4.2.2 Fatigue Cracking – Calibration Results

Local calibration procedures were conducted for six state agencies (AZ, CO, NC, OR, and WA) and regionally for the Northeastern states. Generally, improvements in the model estimates were found with local calibration. This is especially evident for those efforts for which statistical measures were reported, as listed in Table 6. No statistics were reported in the Washington State study; however, it was reported that local calibration resulted in predictions that matched well with the WSPMS measured data. For the Northeastern states study, only the sum of squared error (SSE) was reported. It was found that the SSE was reduced and improved predictions were found with regional calibration. For Arizona and Colorado, the only two studies that reported R^2, a large improvement was found in the coefficient of determination over the verification effort and the R^2 reported for the development of the nationally calibrated model. The locally calibrated coefficients for Colorado resulted in a model that accounts for nearly 63% of the variability in the data, whereas the nationally calibrated model only accounted for 17.5% of the variability in Colorado's dataset. Although an improvement in fatigue cracking predictions and a reduction in bias were reported with the results of the local calibration effort in North Carolina (shown in Table 6 as "NC Cal"), the coefficient of

determination was still considered “poor”, indicating that the locally calibrated model was a poor predictor of the observed fatigue cracking. This was also the case when researchers applied the locally calibrated model to a new dataset to validate the model, shown in Table 6 as “NC Val”. Although the standard error was reduced with locally calibrated coefficients and material-specific calibration factors, it was still twice as large as the standard error reported in the development of the model, indicating substantial amount of scatter remained in the fatigue cracking predictions.

Transfer function coefficients (C_1 and C_2) were the most common coefficients to be calibrated, with five of the seven calibration efforts altering these coefficients. These coefficients are both fixed at 1.0 in the nationally calibrated model. As a result of the local calibration efforts summarized in this report, calibrated values for C_1 ranged between -0.06883 and 1.071, and between 0.225 and 4.5 for C_2. Local calibration coefficients (β_{f1} and β_{f3}) were the next most frequently calibrated coefficients. Although calibrated values for β_{f3} ranged between 0.6 and 1.233, remaining in close proximity to the global calibration value of 1.0, substantial deviations from the global calibration were seen for β_{f1}. The largest value for β_{f1} reported as a result of local calibration was nearly 250 times larger than the global calibration factor of 1.0. The β_{f2} term was altered in three calibration efforts, as noted in Table 6. The k_f-terms were calibrated independently for twelve asphalt mixtures in the effort conducted for North Carolina, the only effort to consider altering these calibration coefficients.

Table 6 Summary of Calibration Efforts for Fatigue (Alligator) Cracking Model (*4, 10, 11, 13, 15, 17, 19, 23*)

Calibration Coefficient	Global Model (Manual of Practice)	AZ	CO	IA	NE States	NC Calibration	NC Validation	OR	WA
k_{f1}	0.007566	0.007566	0.007566	N/A	NR	By mix	By mix	NR	NR
k_{f2}	-3.9492	-3.9492	-3.9492						
k_{f3}	-1.281	-1.281	-1.281						
β_{f1}	1	249.0087	130.3674			3.5	3.5		0.96
β_{f2}	1	1	1			0.72364	0.72364		0.97
β_{f3}	1	1.2334	1.218			0.6	0.6		1.03
C_1	1	1	0.07		-0.06883	0.24377	0.24377	0.56	1.071
C_2	1	4.5	2.35		1.27706	0.24377	0.24377	0.225	1
C_4	6,000	6,000	6,000		6,000	6,000	6,000	6,000	6,000
Statistical Parameters - Calibration									
R^2, %	27.5	50 (58)[1]	62.7	N/A	NR	"poor"	"poor"	NR	NR
S_e (%)	5.01	14.8 (13)[1]	9.4		NR	17.11	10.239	2.644	
N	405	419	56		NR	124	124	NR	
p-value (paired t-test)	NR	0.0837	0.7566		NR	0.001	0.034		
p-value (slope)	NR	0.9897	0.3529		NR	NR	NR		

NR: Not Reported

N/A: Not Applicable

[1] See section C.1.1

4.3 Rutting

4.3.1 Rutting – Verification Results

Verification efforts conducted for total rutting predictions of the nationally calibrated rutting model are summarized in Table 7, along with the nationally calibrated coefficients in the global model, as shown in Version 2.1 of the AASHTOWare Pavement ME Design software. Statistics are summarized for the comparison of field-measured data with total rutting predictions with the global model. Statistical parameters for the development of the global model are also shown, as reported in the Manual of Practice. It is important to note that the nationally calibrated coefficients reported in the Manual of Practice are not consistent with those reported and presumably used as the default coefficients for the AC and granular materials in the AASHTOWare Pavement ME Design Software Version 2.1. While the Manual of Practice reports values of 0.4791 and 1.5606 for AC layer coefficients k_{r2} and k_{r3}, respectively (also listed as k_{2r} and k_{3r} in the Manual of Practice, respectively), the values for these coefficients are reversed in the software. It is likely that the values for coefficients k_{r2} and k_{r3} were mistakenly reversed in previous versions of the Manual of Practice as previous versions of the software list the coefficients as they are shown in Table 7 for the Global Model. Verification efforts for three State Highway Administrations (SHAs) (Missouri, Ohio, and Utah) reported nationally calibrated coefficients for rutting in the HMA layer consistent with the Manual of Practice, while the remaining efforts either did not report the values of the nationally calibrated coefficients or they were consistent with values shown in the software.

Both the Manual of Practice and current and previous versions of the software show the same value for k_{s1} of the fine-graded material as 1.35. However, the Research Results Digest 308, which summarizes the changes made to the MEPDG as a result of the NCHRP Project 1-40D, shows "BrSG = 1.67" under "new rutting calibration factors" (*26*). Given the context of the document, this is believed to represent the k_{s1} term for the fine-graded material.

As was the case with the two k-terms for the AC layers and the k_{s1} term for the fine-graded submodel, there also appears to be erroneous values reported in the Manual of Practice for the granular submodel. The k_{s1} term for granular material is listed as 1.67 in the Manual of Practice (*4*); however, this term is shown as 2.03 in previous versions of the MEPDG (1.003 and 1.1), and the Research Results Digest reports "BrGB = 2.03" under "new rutting calibration factors", believed to represent the k_{s1} term for granular material (*26*). The current version of the Pavement ME Design software (2.1) lists the k_{s1} value as 2.03. It is presumed that the value 2.03 is the correct value as it is shown in the software; therefore, Table 7 lists this value for the k_{s1} term for granular material in the global model.

While there are obvious discrepancies in the reporting of the calibration coefficient values in the documents reviewed in this study, the statistical comparison between measured and predicted rutting are consistent. The plot of average measured total rutting versus predicted total rutting in the Research Results Digest 308 is identical to that shown on page 38 of the Manual of Practice (*4, 26*). The statistical parameters, R^2, number of data points (N), S_e, and S_e/S_y are also consistent between the two documents. Therefore, the statistical parameters for

the global model are reported in Tables 7 and 8 as they are shown in the Manual of Practice. Only one effort, which was completed for North Carolina DOT, utilized k_{s1} values consistent with the Research Results Digest 308. Coefficients listed in Table 7 reflect the reports from which they are referenced unless otherwise noted. Efforts for four SHAs (Colorado, Missouri, Ohio, and Utah) reported nationally calibrated coefficients (k_{s1}) consistent with the Manual of Practice for the fine graded and granular submodels. Two of these studies are the same two SHAs for which reported k_{r2} and k_{r3} values were also consistent with the Manual of Practice; coincidentally, many of the same authors were common among the calibration reports for these studies. The remaining efforts either did not report the values or reported values consistent with the software.

Verification efforts were conducted for eleven SHAs and regionally for the northeastern states. As part of the effort for Tennessee DOT, rutting was evaluated for three different categories of pavements: asphalt pavements and asphalt pavement overlaid with AC; concrete pavements overlaid with AC for low volume traffic (0-1,000 AADTT), listed as "AC+PCC low" in Tables 7 and 8; and concrete pavements overlaid with AC for heavy traffic (1,000-2,500 AADTT), listed as "AC+PCC high" in Tables 7 and 8. For AC overlays on PCC pavements, predicted rutting in the AC overlay was compared with measured rutting of the pavement surface, and for asphalt pavements and asphalt pavement overlaid with AC, predicted total rutting was compared with measured rutting on the pavement surface.

In the study conducted for Utah DOT, verification of the national rutting model was conducted for older pavement that used viscosity graded asphalt (68 data points), and for newer pavement that used Superpave mixes (86 data points), as well as both mix types pooled together. While the predictions for the older pavements were adequate, predictions for the newer Superpave mixes were poor and therefore required local calibration. Statistical values listed in Table 7 reflect the verification effort for the Superpave mixes, as these pavements were the focus of the local calibration effort.

Results ranged widely among the verification efforts for the comparison of measured total rutting and predicted total rutting using the global model. It should be noted that the level of statistical evaluation conducted to assess the quality of the predictions with the nationally calibrated model varied. While three studies did not report any statistical parameters, the coefficient of determination, R^2, was reported for approximately half of the evaluations, and the majority reported the standard error of the estimate (S_e). Six of the efforts quantified the evaluation of the goodness of fit with the coefficient of determination, R^2. The highest R^2 value was reported for the verification efforts for Ohio DOT, in which the coefficient of determination was reported as 64%. This was also the only study to show an R^2 at or above 57.7%, the R^2 found in the development of the global model. The worst fit was reported in efforts for Arizona DOT, for which R^2 of 5% was reported. Efforts conducted for North Carolina DOT did not quantify the goodness of fit of the global model; rather, it was simply reported as "poor."

In evaluating bias in the existing global model, only four studies conducted hypothesis testing. If the hypothesis is rejected, meaning the p-value was less than the significance level, than it is

implied that the parameter being tested (slope or intercept) is significantly different from the value for which it is being tested. In the case of the slope, this value would be 1.0 and the intercept value would be 0. For the hypothesis test for the paired t-test, rejecting the null hypothesis implies that the measured and predicted distress are from different populations, and as such, bias exists. As shown in Table 7, hypothesis testing for the intercept showed that the p-values for all four studies (Colorado, Missouri, Ohio, and Utah) were very small. This implies that the intercept was significantly different from zero; therefore, bias exists in the nationally calibrated model when applied to data specific to these states.

Although not all of the studies specifically commented on existing bias in the MEPDG default total rutting model, an examination of the reports and any plots presented in each study enabled the following summary to be made. The total rutting estimates made with the global model were found to be overestimates of measured rutting in many of the verification efforts (Arizona, Missouri, Ohio, Oregon, Wisconsin, and for new AC pavements and AC overlays on AC pavements in Tennessee, and Superpave mixes in Utah). Although the majority of the efforts conducted showed over-prediction of total rutting, under-prediction of total rutting was reported in efforts for North Carolina and Washington State. Bias, in the form of over- or under-prediction, was reported in many of the studies; however, the direction of the bias was not as evident in some studies. Authors of the Colorado verification/calibration report stated (page 128) there was "a significant bias" in the global rutting model (*11*). However, there is no obvious trend evident in the plot of measured versus predicted total rutting (Figure 82, page 126 of the referenced document (*11*)), as it appears that for some magnitudes of rutting the global model over-predicts and for other magnitudes (high and low) the global model tends to under-predict total rutting. In Iowa, verification efforts indicated the model provided good predictions of total rutting (*13*). The authors did not state whether any bias existed in the total rutting predictions; however, the plot of measured versus predicted total rutting shown in Figure 10 (*13*) appears to show some over-prediction. Authors of the Northeastern states study did not explicitly state whether under- or over-prediction was noted with the national model (*15*). However, based on the plot of measured versus predicted total rutting using the global coefficients (shown in Figure 4.7 of (*15*)), it appears the national default rutting model under-predicted total rutting at the high end (above 0.4 inches).

Table 7 Summary of Verification Efforts for Total Rutting Predictions (*10, 11, 13, 14, 17-24, 27*)

Parameter Category	Calibration Coefficient	Global Model (Software Version 2.1)	AZ	CO	IA	MO	NC	NE States	OH	OR
HMA Rutting	k_{r1}	-3.35412	-3.35412	-3.3541	NR	-3.35412	-3.35412	NR	-3.35412	-3.35412
	k_{r2}	1.5606	1.5606	1.5606		0.4791	1.56061		0.4791	1.5606
	k_{r3}	0.4791	0.4791	0.4791		1.5606	0.47911		1.5606	0.4791
	β_{r1}	1	1	1	1	1	1	1	1	1
	β_{r2}	1	1	1	1	1	1	NR	1	1
	β_{r3}	1	1	1	1	1	1		1	1
Fine Graded Submodel	k_{s1}	1.35	1.35	1.35	NR	1.35	1.67	NR	1.35	1.35
	β_{s1}	1	1	1	1	1	1	1	1	1
Granular Submodel	k_{s1}	2.03	2.03	1.673	NR	1.673	2.03	NR	1.673	2.03
	β_{b1}	1	1	1	1	1	1	1	1	1
Goodness of Fit	R^2, %	57.7	4.6	45.1	NR	32	Poor	NR	64	NR
	S_e	0.107	0.31	0.134	0.08	0.11	0.129		0.035	0.568
	N	334	479	155	NR	183	235		101	NR
Bias	p-value (paired t-test)	NR	NR	<0.0001	NR	<0.0001	0.000	NR	<0.0001	NR
	p-value (intercept)			NR		0.0003	NR		<0.0001	
	p-value (slope)			<0.0001		<0.0001			<0.0001	

Table 7 (Continued) Summary of Verification Efforts for Total Rutting Predictions (*10, 11, 13, 14, 17-24, 27*)

Parameter Category	Calibration Coefficient	Global Model (Software Version 2.1)	TN	TN (AC+PCC) Low	TN (AC+PCC) High	UT	WA	WI
HMA Rutting	kr1	-3.35412	-3.35412	-3.35412	-3.35412	-3.35412	NR	-3.35412
	kr2	1.5606	1.5606	1.5606	1.5606	0.4791*		1.5606
	kr3	0.4791	0.4791	0.4791	0.4791	1.5606*		0.4791
	$\beta r1$	1	1	1	1	0.56	1	1
	$\beta r2$	1	1	1	1	1	1	1
	$\beta r3$	1	1	1	1	1	1	1
Fine Graded Submodel	ks1	1.35	1.35	1.35	1.35	1.35	NR	1.35
	β_{s1}	1	NR	NR	NR	NR		1
Granular Submodel	ks1	2.03	2.03	2.03	2.03	1.673	NR	2.03
	β_{b1}	1	NR	NR	NR	NR		1
Goodness of Fit	R^2, %	57.7	NR	NR	45	9.7	NR	14
	S_e	0.107	0.08	0.08	0.05	0.155		0.106
	N	334	94	43	40	86		139
Bias	p-value (paired t-test)	NR	NR	NR	NR	0.0822	NR	0.0018
	p-value (intercept)					<0.0001		<0.0001
	p-value (slope)					NR		<0.0001

* These values were switched in the report (*21*) but were correct in the software.

4.3.2 Rutting – Calibration Results

Only one of the verification efforts indicated that calibration was not necessary, specific to one of three pavement types evaluated for Tennessee. As a result, calibration efforts were conducted for all of the eleven SHAs and the northeastern states. The most common coefficients to be altered in the calibration process were β_{r1} for the HMA layers, β_{s1} for the granular material, and β_{s1} for fine-graded subgrade material. Among those studies that altered the β_{r1} term as part of the local calibration procedure, the resulting term ranged from 0.477 to 2.2. For those efforts that varied the β_{s1} term for the granular submodel, results ranged from 0 to 2.0654; similarly, values for the β_{s1} term for the fine-graded submodel were between 0 and 1.5.

Results of the calibration efforts were varied among the twelve studies. Although most studies reported improvements in the total rutting predictions with locally calibrated coefficients, correlations between measured and predicted rutting were still poor. This was certainly the case with the study for Arizona DOT and the 2009 study for Utah DOT. After the 2013 recalibration, the rutting prediction was significantly improved with R^2 = 43% and SEE = 0.07 in Utah. The highest R^2 reported with locally calibrated coefficients was 63% for Ohio DOT. This was actually slightly lower than the coefficient of determination reported when the nationally calibrated model was applied to the Ohio dataset (as shown in Table 7). However, coefficient of determination is the only measure of fit and does not assess bias. In the study for Ohio, it was reported that improvements were made in the prediction capability of the model by conducting local calibration; this is evident in the reduction in the standard error of the estimate, S_e. Furthermore, the S_e that resulted from local calibration for Ohio DOT was notably less than the S_e reported in the development of the nationally calibrated model. Although R^2 was not reported for the Oregon study, a significant reduction in S_e was reported after local calibration. In the verification effort, a S_e of 0.568 was reported for Oregon; this was reduced through local calibration to a value of 0.180, a value much closer to the S_e reported for the development of the global model.

Although no statistics were reported in the Washington study, it was reported that the locally calibrated model resulted in predictions that matched well with WSPMS data in magnitude and progression and was an improvement over predictions made with the default model. In the Northeastern states study, the only statistic reported, SSE, was reported to have decreased with local calibration, and the regional calibration coefficients gave a better fit between measured and predicted rutting in all layers. In the study conducted for Tennessee, calibration of the rutting model for AC overlays on PCC pavements with low traffic resulted in improved predictions of rutting in the AC overlay (local calibration coefficients in the base and subgrade layers were set to zero to fix predicted rutting in the underlying layers to zero). Additionally, improvements were reported in the Tennessee study for total rutting predictions as a result of local calibration for asphalt pavement and asphalt pavement with AC overlays.

In addition to improvements in overall goodness of fit, some improvements in bias were also realized with local calibration efforts. In the study conducted for North Carolina, significant reduction in bias was reported. No significant bias was found for the locally calibrated model

for Missouri. It was reported for Arizona that the over-prediction bias was removed through local calibration. Additionally, for Colorado and Utah, the “significant bias” reported for the nationally calibrated model was eliminated through local calibration. The p-values reported for Arizona, Colorado, Missouri, and Utah also indicate reduction of bias as a result of local calibration. Although slight, a reduction in bias for the total rutting prediction was noted with local calibration in Iowa. In the same study, improvements were also found in predictions of rutting for each layer with the locally calibrated model. Although there were no statistics reported or hypotheses testing completed for the Oregon DOT study, it was reported that bias was reduced with local calibration of the rutting model. Also, even though there were no statistics reported, it was expected that the predictions were improved for the recalibrated models for the Wisconsin study. While some studies resulted in reduction in bias, Ohio reported that significant bias remained in the predictions despite the recalibration effort.

Table 8 Summary of Calibration Efforts for Total Rutting Predictions (*10, 11, 13, 14, 17-24, 27*)

Parameter Category	Calibration Coefficient	Global Model (Software Version 2.1)	AZ	CO	IA - CAL	IA-VAL	MO	NC - CAL	NC - VAL	NE States	OH
HMA Rutting	kr1	-3.35412	-3.3541	-3.3541	NR	NR	-3.35412	Material-Specific	Material-Specific	NR	-3.35412
	kr2	1.5606	1.5606	1.5606			0.4791				0.4791
	kr3	0.4791	0.4791	0.4791			1.5606				1.5606
	βr1	1	0.69	1.34	1	1	1.07	0.9475	0.9475	1.308	0.51
	βr2	1	1	1	1.15	1.15	1	0.86217	0.86217	NR	1
	βr3	1	1	1	1	1	1	1.35392	1.35392	NR	1
Fine Graded Submodel	ks1	1.35	1.35	0.84	NR	NR	1.35	NR	NR	NR	1.35
	β_{s1}	1	0.37	N/A	0	0	0.01	1.5	1.5	1.481	0.33
Granular Submodel	ks1	2.03	2.03	0.4	NR	NR	1.673	NR	NR	NR	1.673
	β_{s1}	1	0.14	N/A	0	0	0.4375	0.53767	0.53767	2.0654	0.32
Goodness of Fit	R^2, %	57.7	16.5 (21)[1]	41.7	NR	NR	52[2]	15	Poor	NR	63
	S_e (in)	0.107	0.11 (0.12)[1]	0.147	0.07	0.07	0.051	0.122	0.19		0.014
	N	334	497	137	NR	NR	183	235	124		101
Bias	p-value (paired t-test)	NR	0.0568	0.4306	NR	NR	0.943	0.008	0.000	NR	<0.0001
	p-value (intercept)		NR	NR			0.05	NR	NR		0.3395
	p-value (slope)		0.0521	0.0898			0.322				<0.0001

[1]See section C.2.1

[2]See section C.2.4

Table 8 (Continued) Summary of Calibration Efforts for Total Rutting Predictions (*10, 11, 13, 14, 17-25, 27*)

Parameter Category	Calibration Coefficient	Global Model (Software Version 2.1)	OR	TN	TN (AC+PCC) Low	TN (AC+PCC) High	UT*	WA	WI**
HMA Rutting	kr1	-3.35412	-3.35412	-3.35412	-3.35412	-3.35412	-3.35412	NR	NR
	kr2	1.5606	0.4791	1.5606	1.5606	1.5606	1.5606		
	kr3	0.4791	1.5606	0.4791	0.4791	0.4791	0.4791		
	βr1	1	1.48	1.33	2.20	1	0.580	1.05	0.477
	βr2	1	1	1	1	1	1	1.109	1
	βr3	1	0.9	1	1	1	1	1.1	1
Fine Graded Submodel	ks1	1.35	1.35	1.35	N/A	N/A	1.35	N/A	1.35
	β_{s1}	1	0	0.68	0	0	0.28	0	0.451
Granular Submodel	ks1	2.03	2.03	2.03	N/A	N/A	2.03	N/A	2.03
	β_{s1}	1	0	0.12	0	0	0.71	0	0.195
Goodness of Fit	R^2, %	57.7	NR	33	50	N/A	43	NR	NR
	S_e	0.107	0.180	0.05	0.04		0.067		
	N	334	NR	94	43		145		
Bias	p-value (paired t-test)	NR	NR	NR	NR	N/A	0.88	NR	NR
	p-value (intercept)						0.33		
	p-value (slope)						0.55		

* 2013 recalibration results.

** 2014 draft User Manual.

4.4 Transverse Cracking

4.4.1 Transverse Cracking – Verification Results

Nine verification efforts for transverse cracking were conducted and calibration was conducted for four SHAs. Table 9 summarizes the verification and calibration results.

Unlike the other performance prediction models, the calibration coefficients used in the transverse cracking model are dependent on the level of design selected by the user. Listed in Table 9 are two sets of calibration coefficients: those values initially established and presented in the Manual of Practice and older versions of the software, and those values presented in the current software, Pavement ME Design Version 2.1. Although it is unclear at what point in time it occurred (or which version of the software the change was made), an update to the initial equations utilized to determine transverse cracking was completed. Appendix A.3 presents both sets of performance prediction models and the associated calibration coefficients. Only goodness of fit for the transverse cracking model and the associated calibration coefficients, as described in the Manual of Practice, were documented, and as such only R^2 values for the older performance model and calibration coefficients are presented in Table 9.

Verification was attempted for the Northeastern states and eight SHAs: Arizona, Colorado, Iowa, Missouri, Ohio, Oregon, Utah, Washington, and Wisconsin. More than half of the verification efforts were conducted with non-statistical analyses. In the study conducted for Arizona, a Level 3 design was utilized in a non-statistical comparison of measured transverse cracking versus predicted transverse cracking. There were discrepancies in the reported value of the coefficient, K, for Level 3; see Section C.3.1 for more details. Although no statistics were reported in the Arizona study, it was reported that predictions with the global model generally under-predicted measured values. Verification was attempted for the Northeastern states; however, it was speculated that the transverse cracking measurements were made in error; therefore, prediction capabilities could not be quantified. A non-statistical comparison was conducted for Ohio using measured and predicted transverse cracking to verify the model, revealing an adequate performance of the global transverse cracking model. Although the model was verified, it was recommended that due to the limited scale of transverse cracking measurements, a more detailed evaluation should be conducted in the future. Results of the non-statistical approach used for Utah indicated that the nationally calibrated model predicted cracking well for mixes designed with Superpave binders but significantly under-predicted transverse cracking for mixes with conventional binders. Local calibration was not recommended for Utah. Efforts completed for Washington State DOT found the global model to provide reasonable estimates of transverse cracking. For Wisconsin DOT, a non-statistical comparison of measured and predicted transverse cracking indicated that the transverse cracking model using default calibration factors over-predicted transverse cracking in Wisconsin. Local calibration was warranted for Wisconsin.

For the remaining efforts that utilized statistical analyses to evaluate the nationally calibrated model, the predictions were generally not adequate. Verification efforts for Colorado indicated that the global transverse cracking model resulted in poor goodness of fit and bias in the form

of under-prediction of measured values. The nationally calibrated Level 3 model was evaluated for Iowa, resulting in a large S_e, and despite significant measured transverse cracking, the model under-predicted with minimal cracking estimated. Missouri evaluated transverse cracking predictions at both Levels 1 and 3 designs and found that despite the higher R^2 value, Level 3 resulted in significant under-predictions. Predictions at the Level 1 design also showed significant bias, and in some cases resulted in significant over-predictions of measured values. Verification of the Level 3 model was also conducted for Oregon. Although few statistics were reported in that study, from a plot presented in the calibration report, it can be inferred that the nationally calibrated model under-predicted transverse cracking in Oregon.

4.4.2 Transverse Cracking – Calibration Results

Although the recommendation for Arizona was to recalibrate the transverse cracking model, researchers were unable to successfully calibrate the model to local conditions. Ultimately, it was recommended that the model not be used as one of the design criteria in Arizona. Of the four efforts that used statistical analyses to verify the global model, three (Colorado, Missouri, and Oregon) recommended and attempted recalibration. Recalibration was not recommended for Iowa due to the large disparity between measured and predicted values and large bias.

The only statistic reported for Oregon was S_e and it was found that the calibration effort resulted in an increase in S_e. While the predictions in the Oregon study were found to be reasonable, there was no improvement when compared to the nationally calibrated model. Calibration efforts for Colorado and Missouri utilized a Level 1 design and found an improvement in the predictions with the calibrated coefficients. For Colorado, “reasonable” predictions were reported and the significant bias found in the nationally calibrated model was eliminated. In the study conducted for Missouri, excellent predictions were reported but were slightly biased. Finally, the locally calibrated model coefficients reported in a draft user manual are presented in Table 9 for Wisconsin. The draft user manual does not include calibration statistics for the model.

Table 9 Summary of Verification and Calibration Results for the Transverse (Thermal) Cracking Model (*4, 11, 13, 14, 19, 24, 25, 27*)

State	K	R^2, %	S_e (ft/mi)	N	p-value (paired t-test)	p-value (slope)	p-value (intercept)
Global Model (Manual of Practice)	Level 1: 5.0	34.4	NR	NR	NR		
	Level 2: 1.5	21.8					
	Level 3: 3.0	5.7					
Global Model (Software Version 2.1)	Level 1: 1.5 Level 2: 0.5 Level 3: 1.5	NR	NR	NR	NR		
Verification Results							
CO	Level 1: 1.5	39.1	0.00232	NR	0.0123	<0.0001	NR
IA	Level 3: 1.5	NR	1203	NR	NR	NR	NR
MO	Level 1: 1.5[1]	Level 1: 52	Level 1: 459	49	Level 1: <0.0001	Level 1: <0.0001	Level 1: <0.0001
	Level 3: 1.5[1]	Level 3: 78	Level 3: 281[2]		Level 3: 0.0001	Level 3: <0.0001	Level 3: 0.125
OR	Level 3: 1.5	NR	121	NR	NR		
WI	Level 3: 3.0	NR	NR	94	NR		
Calibration Results							
CO	Level 1: 7.5	43.1	194	12	0.529	0.339	NR
IA	N/A						
MO	Level 1[1]: 0.625	Level 1: 91	51.4	49	0.0041	<0.0001	0.907
OR	10	NR	751	15	NR		
WI	Level 1: 3.0 Level 2: 0.5 Level 3: 3.0	NR	NR	NR	NR		

[1] Inconsistency in reporting in reference document, see section C.3.4

[2] Inconsistency in reporting in reference documents, see section C.3.4

4.5 IRI

4.5.1 IRI – Verification Results

Verification efforts were conducted for seven SHAs and the northeastern states for the International Roughness Index (IRI) model. Six of these efforts resulted in recalibration. Table 10 summarizes the results of the verification and calibration efforts by agency.

In verifying the nationally calibrated IRI model in Missouri, researchers utilized predictions from individual locally calibrated distress models for the inputs rather than the nationally calibrated distress models. This resulted in reasonable predictions with a slight bias (under-estimates at higher magnitudes of IRI). A similar approach was taken in Iowa. First, verification efforts were conducted using predictions with individual global distress models (rut depth, load related cracking, and thermal cracking) for inputs to the nationally calibrated model. This resulted in good estimation of field measurements. Iowa also utilized a separate dataset to conduct a validation of the nationally calibrated IRI model with global distress model inputs. Additionally, distresses predicted with locally calibrated models were used in conjunction with the nationally calibrated coefficients in the IRI model for the two datasets. This method resulted in good estimation of measured values for both the verification dataset and the validation dataset. Researchers took a similar approach in verifying the IRI model in Washington State. Estimates for cracking and rutting made with locally calibrated models were used as inputs in the default IRI model, resulting in under-predictions of actual roughness, although it was noted that the differences were small. While it was believed the under-prediction in Washington could be resolved through calibration of the IRI model, software bugs reportedly prevented such efforts. The verification effort conducted for Utah utilized rutting predictions from the locally calibrated models to estimate the rut depth input needed for the nationally calibrated IRI model. Based on adequate goodness of fit, acceptable standard error, and only slight bias evident at IRI approaching zero, recalibration of the nationally calibrated model was not necessary for Utah. Similar to the approach used in Utah, the IRI prediction model was verified for Wisconsin. While the R^2 and SEE were found to be reasonable, the nationally calibrated IRI model generally over-predicted IRI when it was less than 70 in/mi and under-predicted IRI when it was greater than 70 in/mi, and the difference between the predicted and measured IRI values was statistically significant. A local calibration of the IRI model was recommended for Wisconsin. The remaining efforts did not state that any locally calibrated distress models were utilized to determine inputs for the nationally calibrated IRI model.

The correlation between predicted and measured IRI described by R^2 varied widely, from 0.8% to 67%, with all but two efforts resulting in a value less than the 56% reported for the development of the nationally calibrated model. Despite the variation in reported R^2 values, all reported S_e values fell below that determined in the model development, indicating that there is more precision when applied to a state's conditions but less accuracy than when developed at the national level. Predicted IRI was converted to PSI and compared with measured PSI values in Tennessee. For lower levels of traffic (cumulative ESALs less than 4.5 million), pavement roughness (IRI converted to PSI) was under-predicted by the default model. For higher levels (cumulative ESALs between 4.5 million and 9 million), high variability in predicted

roughness was observed. Based on the results of the verification of the default IRI model (no statistics were reported), calibration was recommended but not conducted for Tennessee. Only SSE was reported in the Northeastern states study, which indicated that the correlation between measured and predicted values for the nationally calibrated IRI model was very poor.
In the Arizona study, bias in the form of large over-predictions for lower IRI and under-predictions for higher IRI was observed. Similarly, in the Ohio study, the default IRI model was found to over-predict IRI for lower magnitudes (less than 80 inches/mile) and under-predict at higher measured IRI values (greater than 80 inches/mile). Although it was reported in the Colorado study that the nationally calibrated model over-predicts for higher magnitudes of IRI, the plot of measured versus predicted IRI values shown in the referenced document shows that the model under-predicts for higher magnitudes of IRI.

4.5.2 IRI – Calibration Results

For the following states, calibration was recommended after verification: Arizona, Colorado, Missouri, Northeastern States, Ohio, and Wisconsin. Local calibration resulted in improvements in the goodness of fit for Arizona, Colorado, Missouri, and Ohio. In Arizona, local calibration resulted in a very good R^2 value and the elimination of the large over-prediction bias found with the global model. Although S_e increased slightly after calibration in Colorado, the R^2 saw a notable increase and the under-prediction bias was removed, resulting in improved predictions with the calibrated model. A reasonable correlation between measured and predicted IRI with the locally calibrated IRI model was found for Missouri. Additionally, some bias in the locally calibrated model was reported, however, it was considered reasonable.

Only the SSE was reported for the Northeastern states study, and it was reported that the regional calibration improved the IRI predictions and resulted in a reduction in SSE. Improvements in IRI predictions were reported with local calibration in Ohio, resulting in an R^2 of 69%, and although bias remained, it was reported to be more reasonable than the bias in the nationally calibrated model. Finally, the IRI model was recalibrated to remove the bias identified in the model verification process. The recalibrated model coefficients presented in Table 10 for Wisconsin are from a draft user manual, which does not report model calibration statistics.

Table 10 Summary of Verification and Calibration Results for the IRI Model (*4, 10, 11, 13, 14, 15, 18, 21, 24, 25, 27*)

State	C_1	C_2	C_3	C_4	R^2, %	S_e (in/mi)	N	p-value (paired t-test)	p-value (intercept)	p-value (slope)
Global Model	40	0.4	0.008	0.015	56	18.9	1,926	NR		
Verification Results										
AZ	40[1]	0.4	0.008	0.015[1]	30	18.7	675	NR		
CO	40[2]	0.4	0.008	0.015[2]	35.5	15.9	343	0.553	0.0001	NR
IA - Global Distress + Natl Cal Coeff	40	0.4	0.008	0.015	NR	12.35	NR	NR		
IA - Validation Global Distress + Natl Cal Coeff	40	0.4	0.008	0.015	NR	13.23	NR	NR		
IA - Locally Cal Distress + Natl Cal Coeff	40	0.4	0.008	0.015	NR	10.79	NR	NR		
IA – Validation Locally Cal Distress + Natl Cal Coeff	40	0.4	0.008	0.015	NR	12.83	NR	NR		
MO	40	0.4	0.008	0.015	54	13.2	125	0.0182	0.0037	0.0953
NE States	40	0.4	0.008	0.015	NR			NR		
OH	40	0.4	0.008	0.015	0.8	9.8[4]	134	<0.0001	<0.0001	0.78
UT	40	0.4	0.008	0.015	67	16.6	162	0.179	0.1944	<0.0001
WI	40	0.4	0.008	0.015	62.7	5.694	142	0.0004	<0.0001	0.0161
Calibration Results										
AZ	1.2281	0.1175	0.008	0.028	82.2 (80)[3]	8.7 (8)[3]	559	0.1419	NR	0.7705
CO	35	0.3	0.02	0.019	64.4	17.2	343	0.1076	NR	0.3571
MO	17.7	0.975	0.008	0.01	53[5]	13.2[5]	125[5]	0.6265	0.0092	0.225
NE States	51.6469	0.000218	0.0081	-0.9351	NR			NR		
OH	0.066	1.37	0.01	17.6	69	15.9	134	0.455	<0.0001	<0.0027
WI	8.6733	0.4367	0.00256	0.0134	NR			NR		

[1] Inconsistency in reporting in reference document, see section C.4.1
[2] Inconsistency in reporting in reference document, see section C.4.2
[3] Inconsistency in reporting in reference document, see section C.4.1
[4] Inconsistency in reporting in reference document, see section C.4.6
[5] Inconsistency in reporting in reference document, see section C.4.4

4.6 Longitudinal Cracking

Verification and calibration efforts of the longitudinal cracking model were conducted for three SHAs and the Northeastern states, as summarized in Table 11. This table also shows the nationally calibrated coefficients in the global model as shown in the AASHTOWare Pavement ME Design software Version 2.1. While these efforts evaluated the current default longitudinal cracking model, it is anticipated a new longitudinal cracking model will be developed as part of the NCHRP 1-52 project.

4.6.1 Longitudinal Cracking – Verification Results

Few statistics were reported in the verification efforts conducted for the three SHAs and the Northeastern states. In the case of Washington State, no statistics were reported; however, it was reported that the default model tended to under-predict measured values. In the study for the Northeastern states, the only statistic reported was SSE (58.18) for the nationally calibrated model, which was shown to severely under-predict measured longitudinal cracking. In verification efforts for the other two agencies, Iowa and Oregon, only S_e was reported. For Iowa, it was reported that the global model severely under-predicted the extent of longitudinal cracking.

4.6.2 Longitudinal Cracking – Calibration Results

As a result of the verification efforts, calibration was conducted in all four studies. Of the four studies that carried out local calibration, only one varied all three coefficients, and the remaining three varied only C_1 and C_2. In the Northeastern states study, an improvement in longitudinal cracking predictions was found through regional calibration exhibited by a reduction in SSE from 58.18 to 25.67. For Iowa and Oregon, it was reported that the predictions and bias after calibration were improved, but standard error was still large. It was indicated that for Washington, the calibrated model was able to reasonably estimate longitudinal cracking and showed the level and progression of cracking agreed with measured values. Iowa attempted a validation of the locally calibrated model; however, the only statistical parameter reported, S_e, remained much larger than the S_e determined in the development of the default model. The Iowa study recommended that predictions of longitudinal cracking in the MEPDG should be used only for experimental or informational purposes until the ongoing refinement of the model is complete and it is fully implemented. In general, “improvements” in the model predictions were reported, but the lack of statistics or poor statistics presented make the assessment of the calibration process difficult.

Table 11 **Summary of Verification and Calibration Results for the Longitudinal (Top-down) Cracking Model (*4, 13, 15, 19, 23*)**

State	C_1	C_2	C_4	R^2, %	S_e (ft/mile)	N	p-value (paired t-test)	p-value (intercept)	p-value (slope)
Global Model	1	1	1,000	54.4	582.8	312	NR		
Verification Results									
IA	7	3.5	1,000	NR	3,039	NR	NR		
NE States	7	3.5	1,000	NR			NR		
OR	7	3.5	1,000	NR	3,601	NR	NR		
WA	7	3.5	1,000	NR			NR		
Calibration Results									
IA - Cal.	0.82	1.18	1,000	NR	2,767	NR	NR		
IA - Val.	0.82	1.18	1,000	NR	2,958	NR	NR		
NE States	-1	2	1,856	NR			NR		
OR	1.453	0.097	1,000	NR	2,569	NR	NR		
WA	6.42	3.596	1,000	NR			NR		

5 SUMMARY AND CONCLUSIONS

This report summarized evaluations of the nationally calibrated distress models in the MEPDG for the purpose of conducting local calibration. The results of such local calibration efforts were also summarized. The nationally calibrated transfer models evaluated included fatigue cracking, rutting, transverse cracking, IRI, and longitudinal cracking.

The *Guide for the Local Calibration of the MEPDG* describes the recommended procedures for verification and local calibration of the nationally calibrated distress models embedded within the MEPDG (*5*). For this report, verification refers to the process of predicting distress through the MEPDG using the nationally calibrated models along with project-specific information and then comparing those predictions to measured values. These comparisons are necessary to evaluate the accuracy of the model, the spread, and whether bias exists in the predictions. In doing so, it indicates whether a model's prediction capabilities are acceptable or if a local calibration is required. If local calibration is necessary, the results of the verification provide insight into which coefficients should be focused on to improve accuracy and/or bias.

The following can be summarized regarding the verification efforts for default nationally calibrated models for the MEPDG:

- For the majority of the studies summarized herein, the fatigue (alligator) cracking model resulted in poor or inadequate estimates of measured values, with seven of the ten efforts recommending local calibration. Additionally, the default fatigue cracking model commonly showed bias with seven of the ten efforts reporting under-prediction of observed values.

- All twelve studies evaluated the default rutting models, and as a result, local calibration of the total rutting model was recommended in all twelve studies. Bias in the nationally calibrated model was commonly observed with the majority reporting over-prediction. Only two studies reported under-prediction of total rutting with the global model.
- In evaluating the nationally calibrated transverse cracking model, more than half of the studies conducted non-statistical analyses. One study was not able to properly evaluate the model due to possible errors in measured values. Only three studies reported the model predictions to be adequate. Bias was reported in more than half of the studies, and was generally in the form of under-prediction of transverse cracking, although one study did find the model over-predicted at a Level 1 design in some cases.
- In verifying the nationally calibrated IRI model, four studies utilized locally calibrated distress models for inputs in the nationally calibrated model. Of these four studies, three found the predictions to be adequate, two of which reported adequate predictions with only slight bias. The one study that found the predictions to be inadequate reported small under-predictions of IRI.
- In the remaining IRI studies, all five studies found the predictions to be inadequate or poor. Bias was reported in four of these five studies: two studies reported bias in the form of over-predictions at low magnitudes and under-predictions at high magnitudes, one study showed under-predictions for higher magnitudes of IRI, and one study found pavement roughness to be under-predicted for routes with cumulative ESALs between 0 and 4.5 million.
- The default longitudinal cracking model was found to produce inadequate estimates in all four of the studies summarized herein. Three studies reported under-prediction of longitudinal cracking, while one study reported considerable over and under-prediction.

While the AASHTO calibration guide details a step-by-step procedure for conducting local calibration, the actual procedures utilized for calibration of the models to state-specific conditions vary from agency to agency. Differences in the calibration procedures relative to the AASHTO calibration guide were observed:

- Minimum number of roadway segments necessary to conduct the local calibration for each distress model is provided in the AASHTO calibration guide; however, the step for estimating sample size for assessing the distress models was not always reported. For those efforts that did report a sample size, some were smaller than the minimum recommended in the calibration guide. For example, the rutting model for Tennessee was calibrated utilizing 18 pavement segments, two less than the recommended minimum of 20. Furthermore, as shown in the verification and calibration summary tables (Tables 5-11), the number of data points used in the evaluations were not reported in many of the efforts summarized herein, making it difficult to assess if the resulting statistical parameters are meaningful.
- Typically, calibration was attempted by looking at the predicted values and associated measured distress for a collective set of roadway segments and reducing the error between measured and predicted values by optimizing the local calibration coefficients. However, other approaches were taken. For North Carolina DOT, the global calibration

coefficients in the rutting and fatigue cracking models were recalibrated for the state, while the material-specific coefficients for the rutting and fatigue cracking models were recalibrated to better predict performance for the asphalt concrete mixes commonly used in North Carolina.

- The AASHTO calibration guide recommends conducting statistical analyses to determine goodness of fit, spread of the data, as well as the presence of bias in the model predictions (*5*). Three hypothesis tests are recommended in the calibration guide: 1) to assess the slope, 2) to assess the intercept of the measured versus predicted plot, and 3) a paired t-test to determine if the measured and predictions populations are statistically different. Despite these recommendations, the number of hypothesis tests conducted varied from all three to none, with many of the efforts relying on qualitative comparisons of measured versus predicted distresses.

The differences noted above may be due, in part, to the timing of the publication relative to the initiation of such efforts and the release of new versions of the software. The AASHTO calibration guide was published in 2010; however, appendices detailing the models embedded within the MEPDG were published in 2004 as part of the draft Final Report, and the Manual of Practice was published in 2008. From 2004 to present, the software supporting the MEPDG has undergone many iterations from the initial version (which saw a number of versions before the release of the DarWIN ME software), to the current AASHTOWare Pavement ME Design software, now in its second version. This presents challenges for state agencies, as local calibration is a cumbersome and intensive process and the software and embedded distress models are evolving faster than local calibration can be completed.

The hypothesis testing mentioned is used to evaluate the level of bias in the existing global models as they apply to state-specific data and the level of bias in the locally calibrated models. These hypothesis tests include tests for bias by the deviation from the line of equality in slope, and intercept. Additionally, the AASHTO calibration guide recommends a paired t-test to determine if the populations are significantly different (*5*). However, few calibration efforts summarized in this report conducted such testing for bias. Rather, in those efforts that did evaluate bias, it was often qualitatively evaluated through visual examination of measured versus predicted plots.

The results of the calibration efforts reviewed in this report are summarized for each distress model in Table 12. The table also denotes the number of verification, calibration, and validation efforts conducted for each performance model and the state for which the efforts were completed. The rutting model was the most commonly calibrated model. The transverse and longitudinal cracking models were calibrated the least, with the longitudinal cracking model having a significant spread.

The main goal of local calibration is to improve model predictions by reducing bias and increasing precision. The coefficient of determination (R^2) and standard error of the estimate (S_e) are typically used to assess precision of the model predictions. Comparing the S_e and R^2 values for the locally calibrated model with those values determined in verification (applying

the nationally calibrated model to state specific data) can be done to assess the relative improvement in model predictions. The R^2 and S_e associated with the development of the nationally calibrated models are also important benchmarks that should be used for assessing improvement. Reasonable S_e values for each model have been suggested in previous literature and are listed in Table 2 (*3, 5*). These values should be taken into consideration when evaluating the performance of the locally calibrated models. Generally, local calibration should result in an increase in the R^2 value and a reduction in the S_e value. With these guidelines in mind, the following observations regarding the calibration studies were made:

- Only two studies reported R^2 prior to and after local calibration for fatigue cracking. In both cases, applying state-specific data to the nationally calibrated model resulted in R^2 values below 27.5%, the R^2 value reported in the development of the global fatigue model. For these two studies, a significant improvement in the R^2 for the locally calibrated model was noted. Despite these improvements, three of the four local calibration efforts resulted in S_e values greater than that found in the development of the model, as well as the 7% considered reasonable. Thus, even with local calibration, substantial spread exists in those model predictions.
- Opposite of the fatigue cracking model, results for the local calibration of the rutting model showed that only one of the eight reported R^2 values was greater than that reported for the development of the global model. However, three of the five efforts that reported an R^2 value for both the verification and calibration efforts showed higher R^2 values for the local calibration results than the verification results, indicating that some level of relative improvement was found through local calibration. One effort indicated the global model resulted in a "poor" goodness of fit and only 15% of the variability in the data was explained by the locally calibrated model. For the ten local calibration efforts conducted that reported S_e, all but one effort saw an improvement in S_e (S_e was reduced). Six calibration efforts (including two efforts completed for Tennessee) resulted in S_e values less than the reasonable S_e value (as shown in Table 2) of 0.10 inches.
- Only two of the four local calibration attempts of the transverse cracking model reported R^2 values, both were greater than that determined in the development of the model at a Level 1 analysis. Due to the limited R^2 values reported for local calibration, it is difficult to surmise the relative improvement experienced by conducting local calibration. However, a S_e value of 250 feet for the transverse cracking model is considered reasonable, and all three efforts that reported S_e were less than 250 feet.
- In locally calibrating the IRI model, only three of the five efforts reported R^2. All three R^2 values were higher for the local calibration model than the application of the nationally calibrated model, indicating a relative improvement in model accuracy. Additionally, the S_e value for the locally calibrated model was reported in four efforts, revealing that in two cases the calibrated model was less precise (an increase in S_e). All but one study that reported S_e resulted in reasonable precision (S_e less than 17 in/mile).
- The only parameter (R^2, bias, or S_e) reported for the local calibration efforts of the longitudinal cracking model was S_e, and it was only reported for two of the four efforts. In both cases, a significant amount of spread existed in the calibrated model, with both far exceeding 600 ft/mile, an S_e value that is considered reasonable.

While the AASHTO calibration guide recommends a quantitative statistical analysis that provides insight into not only accuracy, but also spread and bias of the predictions relative to measured values, prediction capability was in large part assessed with qualitative descriptions (*5*). In many cases in which qualitative analysis was conducted, the data were inadequate to perform the recommended statistical analysis. For those efforts that did complete statistical analysis, the number of parameters that were reported ranged, and included one or more of the following: R^2, S_e, S_e/S_y, SSE, or p-values for the hypothesis tests for bias. Goodness of fit, characterized by R^2, was reported in some of the efforts. Generally, R^2 values were found to be low when the global models were applied to state data, with the highest R^2 value of 67% found for the IRI model in the study completed for Utah. While in many cases it was improved over the global models, the R^2 values (where reported) for the locally calibrated models were also relatively low, with the highest value reported as 82% for the locally calibrated IRI model for Arizona. However, due to the inconsistency in reporting for the statistical parameters it is difficult to assess the quality of the predictions or effectiveness of local calibration for each model.

Table 12 Number of Verification/Calibration Studies and Summary of Calibration Results

Model	Verification	Calibration	Validation	Results of Calibration
Fatigue Cracking	[10] AZ, CO, IA, MO, NE states, NC, OR, UT, WA, WI	[6] AZ, CO, NE states, NC, OR, WA	[1] NC	• All seven efforts resulted in improvements in predictions. • Two studies (AZ, CO) resulted in sizeable increases in R^2 compared to R^2 in verification effort. Both studies had R^2 values much greater than the development of the global model (R^2 = 27.5%) but were only moderately high (50% and 62.7%). • Reduction or elimination of bias was reported in four (AZ, CO, NC, OR) of the seven studies. • One study (NE states) reported only the Sum of the Squared Error (SSE), which was reduced with calibration. • Two efforts (WA, WI) were qualitative analyses. Both resulted in predictions closely matching measured data.

Model	Verification	Calibration	Validation	Results of Calibration
Total Rutting	[12] AZ, CO, IA, MO, NE states, NC, OH, OR, TN, UT, WA, WI	[12] AZ, CO, IA, MO, NE states, NC, OH, OR, TN, UT, WA, WI	[2] IA, NC	• Generally, improvements in predictions were reported with calibrated models. • Four efforts (AZ, MO, NC, UT) resulted in an increase in R^2. Two efforts (CO, OH) saw decreases in R^2. • Overall, R^2 remained low for the efforts that reported it, ranging from 14.4% to 63%, with only one greater than the R^2 (57.7%) reported in the development of the default model. • Eight studies (AZ, IA, MO, NC, OH, OR, TN, UT) resulted in improvements in standard error of the estimate, S_e, while one study (CO) resulted in an increase in S_e. • Even though most saw improvements in standard error, S_e remained greater than 0.107, the S_e for the development of the default model, in four studies (AZ, CO, NC, OR). • Bias was eliminated or reduced in at least seven studies (AZ, CO, IA, MO, NC, OR, UT) One effort (OH) showed bias remained despite calibration. • Four efforts (NE states, TN, WA, WI) did not report on bias, but all four resulted in improvements in predictions.

Model	Verification	Calibration	Validation	Results of Calibration
Transverse Cracking	[10] AZ, CO, IA, MO, NE states, OH, OR, UT, WA, WI	[5] AZ, CO, MO, OR, WI	[0]	• Two studies (CO, MO) resulted in improvements in R^2 with both values (43.1% and 91%) greater than the R^2 reported in the development of the default model at a Level 1 analysis (34.4%). • Two calibration attempts (AZ, OR) were unsuccessful in improving transverse cracking predictions, and therefore, were not recommended for use. • Predictions were reasonable for one study (CO) with the elimination of bias. Two studies (MO, WI) resulted in good predictions with slight bias.
IRI	[10] AZ, CO, IA, MO, NE states, OH, TN, UT, WA, WI	[5] AZ, CO, MO, NE states, OH, WI	[1] IA	• Generally, improvements in IRI predictions were realized with locally calibrated models. • Three efforts (AZ, CO, OH) resulted in an improvement in R^2, ranging from 64.4% to 82.2%, all of which were greater than the R^2 for the development of the default model (56%). • Only SSE was reported for one study (NE states), which indicated an improvement in predictions with the calibrated model • Bias was removed through three efforts (AZ, CO, WI), while the bias that remained in two efforts (MO, OH) was considered reasonable.

Model	Verification	Calibration	Validation	Results of Calibration
Longitudinal Cracking	[4] IA, NE states, OR, WA	[4] IA, NE states, OR, WA	[1] IA	• For two studies (IA, OR) the predictions and bias were reportedly improved, however, the S_e remained large in both calibrated models. • Despite improved predictions through calibration in one study (IA), the model was recommended for use only for experimental or informational purposes. • One study (NE states) only reported SSE, which was reduced with calibration, resulting in improved predictions. • One qualitative analysis (WA) was conducted and resulted in reasonable predictions.

6 RECOMMENDATIONS

The following recommendations are made based on the summary of existing literature documenting local calibration efforts across the country:

- As noted in this document, there are discrepancies between models and/or model coefficients as they are listed in the first edition of the Manual of Practice relative to the values presented in the corresponding version of the software. Specifically, differences were found between the software and the first edition of the Manual of Practice for the model coefficients in the rutting model, including the k_{r2} and k_{r3} coefficients for the AC layer and the k_{s1} coefficients for the fine-graded and granular sub-models. Additionally, differences in the form of the transverse cracking model were found, as shown in Appendix A. In order to effectively calibrate the embedded distress models for state-specific conditions, it is necessary to understand the form of the model and the correct model coefficients as they are utilized in the software. Additionally, it is necessary to understand the statistical parameters associated with the national calibration of such models in order to better understand if local calibration is necessary, and if so, which model coefficients should be addressed. It is recommended that these coefficients and the form of the model be verified before conducting verification and calibration exercises.
- With a significant spread reported for the current longitudinal cracking model, this model should not be used for design. It is anticipated that a new model will be developed under the ongoing NCHRP 1-52 project.

- As recommended in the *Guide for the Local Calibration of the MEPDG*, the mechanism of the distress, particularly, fatigue cracking, should be accurately identified. In evaluating and calibrating cracking models, it is important to differentiate between top-down and bottom-up cracking in measured field performance.
- The software continues to evolve, as noted above the form for at least one transfer function has been revised in newer versions of the software. Future refinements of transfer models, such as the longitudinal cracking model, are expected. Furthermore, local calibration can be a time consuming and intensive process, therefore an agency may need to consider the implications of conducting local calibration efforts while the embedded models and software continue to be refined. One approach agencies may consider is to estimate the return on investment of calibrating each performance model.
- Although the calibration efforts for some agencies began prior to the publication of AASHTO's guide for local calibration, it is recommended that any future or on-going calibration efforts be completed in accordance with the current AASHTO *Guide for the Local Calibration of the MEPDG*. With a consistent process being utilized among SHAs, a more complete evaluation of pavement performance could be conducted on the national level.
    - Although it was not conducted in the development of the transfer models (*5*) or explicitly stated in the calibration efforts reviewed for this report, it should be noted that forensic investigations are necessary to adequately identify the type and location of distress within the pavement structure. The *Guide for Local Calibration of the MEPDG* addresses this in Step 6 of the step-by-step procedure in which forensic investigations, such as coring, and trenching, are suggested.
- Conducting statistical analyses as outlined in the *Guide for the Local Calibration of the MEPDG* is recommended, such analyses enable a quantitative assessment of the calibration results. Specifically, the parameters help to determine if local calibration has reduced bias and improved precision, as well as any weaknesses that may exist in the model that must be considered during the design process.

REFERENCES

1. *AASHTO Guide for Design of Pavement Structures*. American Association of State and Highway Transportation Officials, Washington D.C., 1993.
2. ARA, Inc., ERES Consultants Division. *Guide for Mechanistic-Empirical Design of New and Rehabilitated Pavement Structures.* Final report, NCHRP Project 1-37A. Transportation Research Board of the National Academies, Washington, D.C., 2004. http://www.trb.org/mepdg/guide.htm
3. Pierce, L. M. and G. McGovern. *NCHRP Synthesis Report 457: Implementation of the AASHTO Mechanistic-Empirical Pavement Design Guide and Software*. TRB, National Research Council, Washington, D.C., 2014, pp. 81.
4. *Mechanistic-Empirical Pavement Design Guide, Interim Edition*: *A Manual of Practice.* AASHTO, Washington, D.C., 2008.
5. Guide for the Local Calibration of the Mechanistic-Empirical Pavement Design Guide. AASHTO, Washington, D.C., 2010.
6. FHWA. TPF-5(178): Implementation of the Asphalt Mixture Performance Tester (AMPT) for Superpave Validation. http://www.pooledfund.org/details/study/405. Accessed May 13, 2013.
7. FHWA. DGIT Workshops and Training: AASHTOWare Pavement ME Design Webinar Series. http://www.fhwa.dot.gov/pavement/dgit/dgitwork.cfm#nhi. Accessed May 13, 2013.
8. Carvalho, R. and C. Schwartz. Comparisons of Flexible Pavement Designs: AASHTO Empirical Versus NCHRP Project 1-37A Mechanistic-Empirical. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1947*, TRB, National Research Council, Washington, D.C., 2006, pp. 167-174.
9. Timm, D., X. Guo, M. Robbins, and C. Wagner. M-E Calibration Studies at the NCAT Test Track. *Asphalt Pavement Magazine*, Vol. 17, No. 5, National Asphalt Pavement Association, 2012, pp. 45-51.
10. Darter, M. I., L. Titus-Glover, H. Von Quintus, B. Bhattacharya, and J. Mallela. *Calibration and Implementation of the AASHTO Mechanistic-Empirical Pavement Design Guide in Arizona*. Report FHWA-AZ-14-606, Arizona Department of Transportation, Phoenix, Ariz., 2014.
11. Mallela, J., L. Titus-Glover, S. Sadasivam, B. Bhattacharya, M. Darter, and H. Von Quintus. *Implementation of the AASHTO Mechanistic Empirical Pavement Design Guide for Colorado*. Report CDOT-2013-4, Colorado Department of Transportation, Denver, Colo., 2013.
12. Ceylan, H., S. Kim, K. Gopalakrishnan, and O. Smadi. *MEPDG Work Plan Task No. 8: Validation of Pavement Performance Curves for the Mechanistic-Empirical Pavement Design Guide*. CTRE Project 06-274 Final Report, Center for Transportation Research and Education, Iowa State University, Ames, Iowa, 2009.
13. Ceylan, H., S. Kim, K. Gopalakrishnan, and D. Ma. *Iowa Calibration of MEPDG Performance Prediction Models*. InTrans Project 11-401, Institute for Transportation, Iowa State University, Ames, Iowa, 2013.
14. Mallela, J., L. Titus-Glover, H. Von Quintus, M. Darter, M. Stanley, C. Rao, and S. Sadasivam. *Implementing the AASHTO Mechanistic Empirical Pavement Design Guide in Missouri, Vol. 1*

Study Findings, Conclusions and Recommendations. Missouri Department of Transportation, Jefferson City, Mo., 2009.
15. Momin, S. A. *Local Calibration of Mechanistic Empirical Pavement Design Guide for North Eastern United States*. MS thesis. University of Texas at Arlington, 2011.
16. Muthadi, N. R. and Y. R. Kim. Local Calibration of Mechanistic-Empirical Pavement Design Guide for Flexible Pavement Design. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2087*, Transportation Research Board of the National Academies, Washington, D.C., 2008, pp. 131-141.
17. Kim, Y. R., F. M. Jadoun, T. Hou, and N. Muthadi. *Local Calibration of the MEPDG for Flexible Pavement Design*. Final Report, FHWA-NC-2007-07, North Carolina Department of Transportation, Raleigh, N.C., 2011.
18. Glover, L. T and J. Mallela. *Guidelines for Implementing NCHRP 1-37A M-E Design Procedures in Ohio: Volume 4 – MEPDG Models Validation and Recalibration*. Final Report: FHWA/OH-2009/9D, Ohio Department of Transportation, Columbus, Ohio, 2009.
19. Williams, C. and R. Shaidur. *Mechanistic-Empirical Pavement Design Guide Calibration For Pavement Rehabilitation*. Report FHWA-OR-RD-13-10, Oregon Department of Transportation, Salem, Ore. and Federal Highway Administration, Washington D.C., 2013.
20. Zhou, C. *Investigation into Key Parameters and Local Calibration on MEPDG*. PhD dissertation. University of Tennessee, Knoxville, 2013.
21. Darter, M. I., L. T. Glover, and H. L. Von Quintus. *Draft User's Guide for UDOT Mechanistic-Empirical Pavement Design Guide.* Report UT-09.11a, Utah Department of Transportation, Salt Lake City, Utah, 2009.
22. Titus-Glover, L., B. Bhattacharya, H. Von Quintus, and M. Darter. *Utah Recalibration of Flexible Pavement Rutting Model AASHTO ME Design Procedure*. Utah Department of Transportation, Salt Lake City, Utah, 2013.
23. Li, J., L. M. Pierce, and J. Uhlmeyer. Calibration of Flexible Pavement in Mechanistic-Empirical Pavement Design Guide for Washington State. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2095*, Transportation Research Board of the National Academies, Washington, D.C, 2009, pp. 73–83.
24. Mallela, J., L. Titus-Glover, H. Von Quintus, M. Darter, F. Fang, and H. Bahia. *Implementation of the Mechanistic-Empirical Pavement Design Guide in Wisconsin*. Final Report, Wisconsin Department of Transportation, Madison, WI, 2009.
25. Mallela, J., L. Titus-Glover, and B. Bhattacharya. AASHTO Pavement ME Design User Manual for Wisconsin. Draft User Manual, ARA, Champaign, Ill., 2014.
26. Darter, M. I., J. Mallela, L. Titus-Glover, C. Rao, G. Larson, A. Gotlif, H. Von Quintus, L. Khazanovich, M. Witczak, M. El-Basyouny, S. El-Badawy, A. Zborowski, and C. Zapata. Changes to the Mechanistic-Empirical Pavement Design Guide Software Through Version 0.900, July 2006. *NCHRP Research Results Digest 308*, Transportation Research Board, Washington, D.C., 2006.
27. Mallela, J., L. Titus-Glover, H. Von Quintus, M. Stanley, and C. Rao. *Implementing the AASHTO Mechanistic-Empirical Pavement Design Guide in Missouri, Volume II: MEPDG Model Validation and Calibration.* Missouri Department of Transportation, Jefferson City, MO, 2009.

APPENDIX A PERFORMANCE MODELS FOR FLEXIBLE PAVEMENT DESIGN

A.1 Introduction

The MEPDG includes several performance (transfer) models to predict the following distresses:

- Rut depth—total, asphalt, and unbound layers (in)
- Transverse (thermal) cracking (non-load related) (ft/mi)
- Alligator (bottom-up fatigue) cracking (percent lane area)
- Longitudinal (top-down) cracking (ft/mi)
- International roughness index (IRI) (in/mi)

These models are presented in this appendix to facilitate the discussion of the local calibration results. The information is adapted from the Manual of Practice (*4*) and the AASHTOWare Pavement ME Design software Version 2.1. When discrepancies are found between the two references, information in the software is presented.

A.2 Rut Depth for Asphalt and Unbound Layers

Two performance models are used to predict the total rut depth of flexible pavements and asphalt overlays: one for the asphalt layers and the other one for all unbound aggregate base layers and subgrades. Equation A.1 shows the asphalt rutting model developed based on laboratory repeated load plastic deformation tests.

$$\Delta_{p(AC)} = \varepsilon_{p(AC)} h_{(AC)} = \beta_{r1} k_z \varepsilon_{r(AC)} 10^{k_{r1}} n^{k_{r2}\beta_{r2}} T^{k_{r3}\beta_{r3}} \quad \text{(A.1)}$$

where:

- $D_{p(AC)}$ = Accumulated permanent or plastic vertical deformation in the asphalt layer or sublayer, in
- $\varepsilon_{p(AC)}$ = Accumulated permanent or plastic axial strain in the asphalt layer or sublayer, in/in
- $\varepsilon_{r(AC)}$ = Resilient or elastic strain calculated by the structural response model at the mid-depth of each asphalt layer or sublayer, in/in
- $h_{(AC)}$ = Thickness of the asphalt layer or sublayer, in
- n = Number of axle load repetitions
- T = Mix or pavement temperature, °F
- k_z = Depth confinement factor shown in Equation A.2
- $k_{r1,r2,r3}$ = Global field calibration parameters (from the NCHRP 1-40D recalibration; k_{r1} = -3.35412, k_{r2} = 1.5606, k_{r3} = 0.4791)
- $\beta_{r1,r2,r3}$ = Local or mixture field calibration constants; for the global calibration, these constants were all set to 1.0

$$k_z = (C_1 + C_2 D)0.328196^D \quad \text{(A.2)}$$

$$C_1 = -0.1039(H_{HMA})^2 + 2.4868 H_{HMA} - 17.342 \quad \text{(A.3)}$$

$$C_2 = 0.0172(H_{HMA})^2 - 1.7331 H_{HMA} + 27.428 \quad \text{(A.4)}$$

where:

- D = Depth below the surface, in

$H_{(AC)}$ = Total asphalt thickness, in

Equation A.5 shows the field calibrated transfer function for the unbound layers and subgrade.

$$\Delta_{p(soil)} = \beta_{s1} k_{s1} \varepsilon_v h_{soil} \left(\frac{\varepsilon_o}{\varepsilon_r}\right) e^{-\left(\frac{\rho}{n}\right)^{\beta}} \quad \text{(A.5)}$$

where:

$D_{p(Soil)}$ = Permanent or plastic deformation for the layer or sublayer, in
n = Number of axle load applications
e_o = Intercept determined from laboratory repeated load permanent deformation tests, in/in
e_r = Resilient strain imposed in laboratory test to obtain material properties ε_o, β, and r, in/in
e_v = Average vertical resilient or elastic strain in the layer or sublayer and calculated by the structural response model, in/in
h_{soil} = Thickness of the unbound layer or sublayer, in
k_{sl} = Global calibration coefficients; k_{s1} = 2.03 for granular materials and 1.35 for fine-grained materials
β_{s1} = Local calibration constant for the rutting in the unbound layers; the local calibration constant was set to 1.0 for the global calibration effort

$$log\beta = -0.6119 - 0.017638(W_c) \quad \text{(A.6)}$$

$$\rho = 10^9 \left(\frac{C_o}{1-(10^9)^{\beta}}\right)^{\frac{1}{\beta}} \quad \text{(A.7)}$$

$$C_o = Ln\left(\frac{a_1 M_r^{b_1}}{a_9 M_r^{b_9}}\right)^{\frac{1}{\beta}} = 0.0075 \quad \text{(A.8)}$$

where:

W_c = Water content, percent
W_r = Resilient modulus of the unbound layer or sublayer, psi
$a_{1,9}$ = Regression constants; a_1=0.15 and a_9=20.0
$b_{1,9}$ = Regression constants; b_1=0.0 and b_9=0.0

A.3 Transverse (Thermal) Cracking

The following is taken from the Manual of Practice (*4*):
The amount of thermal cracking is estimated using Equation A.9 based on the probability distribution of the log of the crack depth to asphalt layer thickness ratio.

$$TC = \beta_{tl} N\left[\frac{1}{\sigma_d} log\left(\frac{C_d}{H_{HMA}}\right)\right] \quad \text{(A.9)}$$

where:

TC = Observed amount of thermal cracking, ft/mi
β_{tl} = Regression coefficient determined through global calibration (400)
$N[z]$ = Standard normal distribution evaluated at [z]

σ_d = Standard deviation of the log of the depth of cracks in the pavement (0.769), in
C_d = Crack depth, in
H_{HMA} = Thickness of asphalt concrete layers, in

The crack depth (C_d) induced by a given thermal cooling cycle is estimated using the Paris law of crack propagation, as shown in Equation A.9.

$$\Delta C = A(\Delta K)^n \quad \text{(A.9)}$$

where:

ΔC = Change in the crack depth due to a cooling cycle
ΔK = Change in the stress intensity factor due to a cooling cycle
A, n = Fracture parameters for the HMA mixture, which are obtained from the indirect tensile creep-compliance and strength of the asphalt mixture using Equation A.10

$$A = 10^{k_t \beta_t (4.389 - 2.52 Log(E_{AC} \sigma_m n))} \quad \text{(A.10)}$$

where:

$n = 0.8\left[1 + \frac{1}{m}\right]$
k_t = Coefficient determined through global calibration for each input level (Level 1 = 5.0; Level 2 = 3.0; and Level 3 = 1.5)
E_{AC} = Asphalt concrete indirect tensile modulus, psi
s_m = Mixture tensile strength, psi
m = M-value derived from the indirect tensile creep compliance curve
β_t = Local or mixture calibration factor (set to 1.0)

The stress intensity factor, *K*, is determined using Equation A.11.

$$K = \sigma_{tip}(0.45 + 1.99(C_o)^{0.56}) \quad \text{(A.11)}$$

where:

s_{tip} = Far-field stress from pavement response model at depth of crack tip, psi
C_o = Current crack length, ft

The following equations for transverse (thermal) cracking are according to the AASHTOWare Pavement ME Design software Version 2.1:

$$C_f = 400 \times N\left[\frac{\log\left(\frac{C}{h_{ac}}\right)}{\sigma}\right] \quad \text{(A.12)}$$

where:

C_f = Observed amount of thermal cracking, (ft/500 ft)
$N[z]$ = Standard normal distribution evaluated at [z]
C = Crack depth, in
h_{ac} = Thickness of asphalt concrete layers, in
σ = Standard deviation of the log of the depth of cracks in the pavements

The change in the crack depth due to a cooling cycle, ΔC, is calculated as shown in Equation A.13

$$\Delta C = (k \times \beta_t)^{n+1} \times A \times \Delta K^n \quad \text{(A.13)}$$

where:

- ΔC = Change in the crack depth due to a cooling cycle
- k = Regression coefficient determined through field calibration (Level 1 = 1.5; Level 2 = 0.5; and Level 3 = 1.5)
- β_t = Calibration parameter
- ΔK = Change in the stress intensity factor due to a cooling cycle
- A,n = Fracture parameters for the asphalt mixture, A is determined by Equation A.14

$$A = 10^{\left(4.389 - 2.52 \times Log(E \times \sigma_m \times n)\right)} \quad \text{(A.14)}$$

where:

- E = Mixture stiffness
- s_m = Undamaged mixture tensile strength

A.4 Alligator (Bottom-Up Fatigue) Cracking

Alligator cracking is assumed to initiate at the bottom of the asphalt concrete layers and propagate to the surface under truck traffic. The allowable number of axle load applications needed for the incremental damage index approach to predict both types of load related cracks (alligator and longitudinal) is shown in Equation A.15 as it is shown in the Manual of Practice (*4*).

$$N_{f-AC} = k_{f1}(C)(C_H)\left(\beta_{f1}\right)(\varepsilon_t)^{k_{f2}\beta_{f2}}(E_{AC})^{k_{f3}\beta_{f3}} \quad \text{(A.15)}$$

where:

- $N_{f\text{-}AC}$ = Allowable number of axle load applications for a flexible pavement and asphalt overlays
- ε_t = Tensile strain at critical locations and calculated by the structural response model, in/in
- E_{AC} = Dynamic modulus of the HMA measured in compression, psi
- $k_{f1,f2,f3}$ = Global field calibration parameters (from the NCHRP 1-40D re-calibration; k_{f1} = 0.007566, k_{f2} = -3.9492, and k_{f3} = -1.281)
- $\beta_{f1,f2,f3}$ = Local or mixture specific field calibration constants; for the global calibration effort, these constants were set to 1.0
- C_H = Thickness correction term, dependent on type of cracking

$$C = 10^M \quad \text{(A.16)}$$

$$M = 4.84\left(\frac{V_{be}}{V_a + V_{be}} - 0.69\right) \quad \text{(A.17)}$$

where:

- V_{be} = Effective asphalt content by volume, %
- V_a = Percent air voids in the HMA mixture

The allowable number of axle load applications as it is presented in the AASHTOWare Pavement ME Design software Version 2.1 is shown in Equation A.18. Equations A.16 and A.17 are applied in the same manner as in Equation A.15.

$$N_{f-AC} = 0.00432(C)(\beta_{f1})(k_1)\left(\frac{1}{\varepsilon_1}\right)^{k_2\beta_{f2}}\left(\frac{1}{E}\right)^{k_3\beta_{f3}} \quad \text{(A.18)}$$

where:

- $N_{f\text{-}AC}$ = Allowable number of axle load applications for a flexible pavement and asphalt overlays
- ε_l = Tensile strain at critical locations and calculated by the structural response model, in/in
- E = Dynamic modulus of the HMA measured in compression, psi
- $k_{1,2,3}$ = Global field calibration parameters (k_1 = 0.007566, k_2 = 3.9492, and k_3 = 1.281)
- $\beta_{f1,f2,f3}$ = Local or mixture specific field calibration constants; for the global calibration effort, these constants were set to 1.0

The allowable axle load applications were then used to determine the cumulative damage index (*DI*), which is a sum of the incremental damage indices over time as shown in Equation A.19.

$$DI = \sum(\Delta DI)_{j,m,l,p,T} = \sum\left(\frac{n}{N_{f-AC}}\right)_{j,m,l,p,T} \quad \text{(A.19)}$$

where:

- n = Actual number of axle load applications within a specific time period
- j = Axle load interval
- m = Axle load type (single, tandem, tridem, quad, or special axle configuration)
- l = Truck type using the truck classification groups included in the MEPDG
- p = Month
- T = Median temperature for the five temperature intervals or quintiles used to subdivide each month, ^{o}F

The area of alligator cracking is calculated from the cumulative damage index at the bottom of the AC layer over time using Equation A.20.

$$FC_{bottom} = \left(\frac{C_4}{1+e^{\left(C_1*C_1'+C_2*C_2'*log(DI_{bottom}*100)\right)}}\right)*\left(\frac{1}{60}\right) \quad \text{(A.20)}$$

where:

- FC_{bottom} = Area of alligator cracking that initiates at the bottom of the AC layers, percent of total lane area
- DI_{bottom} = Cumulative damage index at the bottom of the AC layers
- $C_{1,2,4}$ = Transfer function regression constants; C_4= 6,000; C_1=1; and C_2=1

$$C_2' = -2.40874 - 39.748(1 + h_{AC})^{-2.856} \quad \text{(A.21)}$$

$$C_1' = -2 * C_2' \quad \text{(A.22)}$$

where:

h_{AC} = total thickness of asphalt layer, in

A.5 Longitudinal (Top-Down) Cracking

Longitudinal cracks are assumed to initiate at the surface and propagate downward. The Manual of Practice uses Equations A.15 and A.19 to calculate the allowable number of axle load applications and cumulative damage index for fatigue and longitudinal cracks. The AASHTOWare Pavement ME Design software Version 2.1 uses A.18 and A.19 to calculate the allowable number of axle load applications and cumulative damage index for fatigue and longitudinal cracks. The length of longitudinal cracking is then determined using Equation A.23.

$$FC_{top} = \left(\frac{C_4}{1+e^{\left(C_1 - C_2 * log\left(DI_{top}\right)\right)}}\right) * 10.56 \quad \text{(A.23)}$$

where:

FC_{top} = Length of longitudinal cracking that initiates at the surface, in

DI_{top} = Cumulative damage index at the surface, percent

$C_{1,2,4}$ = Transfer function regression constants; $C4$= 1,000; $C1$=7; and $C2$=3.5

A.6 International Roughness Index (IRI)

The MEPDG uses Equation A.24 to predict IRI over time for AC pavements. This regression equation was developed based on data from the LTPP program.

$$IRI = IRI_0 + C_1(RD) + C_2(FC_{Total}) + C_3(TC) + C_4(SF) \quad \text{(A.24)}$$

where:

IRI_0 = Initial IRI after construction, in/mi

RD = Average rut depth, in

FC_{Total} = Total area of load-related cracking (combined alligator, longitudinal, and reflection cracking in the wheel path), percent of wheel path area

TC = Length of transverse cracking (including the reflection of transverse cracks in existing HMA pavements), ft/mi

$C_{1,2,3,4}$ = Regression constants; $C1$ = 40; $C2$ = 0.4; $C3$ = 0.008; $C4$= 0.015

SF = Site factor (Equation A.25)

$$SF = Frosth + Swellp * Age^{1.5} \quad \text{(A.25)}$$

where:

IRI_0 = Initial IRI after construction, in/mi

Age = pavement age, year

$$Frosth = Ln[(Precip + 1) * Fines * (FI + 1)] \quad \text{(A.26)}$$

$$Swellp = Ln[(Precip + 1) * Clay * (PI + 1)] \quad \text{(A.27)}$$

$$Fines = F_{sand} + Silt \quad \text{(A.28)}$$

where:

PI = subgrade soil plasticity index, percent

Precip = average annual precipitation or rainfall, in

APPENDIX B SUMMARY OF CALIBRATION METHODOLOGIES

B.1 Methodology used in Efforts for Arizona (*10*)

1. DARWin ME (version not stated)
2. Material properties
    a. Asphalt materials included conventional and Superpave mixes. HMA thicknesses included: <8 in and ≥ 8 in.
    b. Base materials typically included granular materials (A-1 and A-2); subgrade typically included coarse grained material (A-1 through A-3)
    c. HMA dynamic modulus Level 2
    d. HMA creep compliance Level 1
    e. Indirect tensile strength Level 3
    f. Effective binder content Level 3
    g. HMA coefficient of thermal contraction Level 2 and 3
    h. Base type/Modulus Level 2
    i. Subgrade type Level 1
3. Pavement sections included: new pavements (AC/granular, thin AC/JPCP) and rehabilitated (AC/AC and AC/JPCP)
4. Climate locations included: northern, central, and southern regions of the state with low and high elevations.

B.2 Methodology used in Efforts for Colorado (*11*)

1. Version 1.0 of the MEPDG (Report date: July 2013)
2. Material properties hierarchical input levels for MEPDG calibration:
    a. HMA dynamic modulus Level 2 (Computed using material gradation, air void, binder type, etc. data)
    b. HMA creep compliance & indirect tensile strength Level 2 (Computed using material gradation, air void, binder type, etc. data)
    c. Volumetric properties Level 3 (CDOT defaults)
    d. HMA coefficient of thermal contraction Level 3 (MEPDG defaults)
    e. Unit weight Level 3 (MEPDG defaults)
    f. Poisson's ratio Level 3 (MEPDG defaults)
    g. Other thermal properties; conductivity, heat capacity, surface absorptivity Level 3 (MEPDG defaults)
3. A variety of new and overlay HMA sections were used for model calibration:
    a. HMA thicknesses included: less than 4 inches; thicknesses between 4 and 8 inches and greater than 8 inches with most sections less than 8 inches thick.
    b. Binder type: neat and modified
    c. Climate zones: hot/moderate, cool, and very cool.
4. Calibration of the MEPDG global models was done using nonlinear model optimization tools (SAS statistical software).

5. The criteria that were used for determining models adequacy for Colorado conditions is presented in Table B.1.

Table B.1 Criteria for Determining Models Adequacy for Colorado Conditions (*11*)

Criterion	Test Statistics	R^2 Range/Model SEE	Rating
Goodness of Fit	R^2, percent (for all models)	81-100	Very Good
		64-81	Good
		49-64	Fair
		<49	Poor
	Global HMA Alligator Cracking model SEE	<5 percent	Good
		5-10 percent	Fair
		>10 percent	Poor
	Global HMA Total Rutting model SEE	<0.1 in	Good
		0.1-0.2 in	Fair
		>0.2 in	Poor
	Global HMA IRI Model SEE	<19 in/mi	Good
		19-38 in/mi	Fair
		>38 in/mi	Poor
Bias	Hypothesis testing-Slope of Linear measured vs. Predicted Distress/IRI Model (b1=slope) H0:b1=0	p- value	Reject if p-value <0.05
	Paired t-test between measured and predicted distress/IRI	P -value	Reject if p-value is <0.05

B.3 Methodology used in Efforts for Iowa

1. Initial verification of the HMA performance models was conducted in 2009 using MEPDG Version 1.0 (*12*).
    a. Level 3 analyses to evaluate the MEPDG globally calibrated performance for Iowa conditions based on statistical measures were conducted.
    b. Performance predictions were compared with actual performance data for the five HMA sections used for the verification study. Performance predictions evaluated included rutting and IRI. Alligator and transverse cracking were not included due to the difference in measurements of these distresses by Iowa DOT and the unit of measurement used in the MEPDG. Longitudinal cracking was excluded due to lack of accuracy in the predictions, as found in their literature review.
    c. A paired t-test was used to check for bias between predictions from the globally calibrated performance models and measured distress values.

    d. Bias was reported for rutting and IRI, although there was good agreement between actual IRI measurements and IRI predictions.
2. Local calibration was performed using the MEPDG Version 1.1
3. Procedure used for local calibration (*13*):
    a. Select typical pavement sections around the state.
    b. Identify available sources to gather input data and determine hierarchical input level.
    c. Prepare input data from available sources: Iowa DOT PMIS, material testing records, design database, and previous research reports relative to MEPDG implementation in Iowa.
    d. Assess local bias associated with national calibration factors.
    e. Determine local calibration factors by conducting sensitivity analysis and optimization of calibration coefficients.
    f. Determine the adequacy of local calibration factors.
4. A total of 35 representative HMA sections were chosen for the local calibration effort, one of which was an Iowa LTPP section (*13*).
5. For HMA material properties, inputs necessary for the MEPDG were taken from an Iowa DOT mix design database. For sections where the mix design information was not available, LTPPBind was used to determine the asphalt binder grade, and typical aggregate gradation based on average HMA aggregate gradations in the Iowa DOT mix design database was used (*13*).
6. A sensitivity analysis was conducted to understand the effect of each calibration coefficient on performance predictions and to more easily identify coefficients that should be optimized (*13*).
7. Non-linear optimization was utilized for local calibration of the cracking and IRI performance models. Linear optimization was utilized for fatigue, rutting, and thermal fracture (*13*).
    a. Linear optimization was used to reduce the large number of computations associated with the trial-and-error procedure.
8. Local calibration was attempted for the following distresses (*13*):
    a. Alligator cracking
    b. Rutting
    c. Thermal (transverse) cracking
    d. IRI
    e. Longitudinal cracking
9. Although it is not explicitly stated in the 2013 report, it can be inferred that the same percentage of the data used for calibration and validation used for JPCP pavements was also used for HMA pavements. It is stated that "about 70% of the total selected sections were utilized to identify the local calibration factors while the remaining 30%, as an independent validation set, were utilized to verify the identified local calibration factors" (*13*). The number of data points in each set was not reported.
10. Accuracy of the performance predictions were evaluated by plotting the measured performance measures against the predicted performance measures and observing the deviation from the line of equality. Additionally, the average bias and standard error

were determined and used to evaluate the nationally calibrated and locally calibrated models using the following equations (*13*):

a. $Average\ Bias = \frac{\sum_{j=1}^{n}(y_j^{measured} - y_j^{predicted})}{n}$

b. $Standard\ error = \sqrt{\frac{\sum_{j=1}^{n}(y_j^{measured} - y_j^{predicted})^2}{n}}$

11. As part of the same study, simulations were completed in the MEPDG and DARWin-ME software to compare performance predictions for the same designs. In some cases, significant differences exist, suggesting the need for further investigation/verification of performance prediction models in the DARWin-ME (currently referred to as AASHTOware PavementME software) (*13*).

B.4 Methodology used in Efforts for Missouri (*14*)

1. Version 1.0 of the MEPDG (Report date: July 2013)
2. Material Properties hierarchical input levels for MEPDG calibration:
    i. HMA dynamic modulus Level 2
    ii. HMA creep compliance & indirect tensile strength Level 3
    iii. Volumetric properties Level 1 (MoDOT specific defaults and LTPP materials database and MoDOT testing program)
    iv. HMA coefficient of thermal contraction Level 3
    v. Unit weight Level 1
    vi. Other thermal properties; conductivity, heat capacity, surface absorptivity Level 3
3. HMA pavement included: New or reconstructed HMA, HMA overlaid over HMA, HMA overlaid over PCC.
4. All HMA thicknesses were included.
5. The general procedure for model validation-calibration (either statistical approach or non-statistical approach) was as follows:

Statistical approach:

1. Evaluate models prediction capabilities by determining the correlation between measure and predicted values using global calibration (default) coefficients. The statistics to make this comparison included coefficient of determination R^2 and the standard error of the estimate (S_e).
2. If predictions are not adequate, local calibration is recommended.
3. Model coefficients are locally calibrated and the same statistics, R^2 and S_e, are used to assess the adequacy of the new coefficients.
4. Bias is determined by performing linear regression using the measured and MEPDG predicted distress; three hypothesis tests are evaluated:
    a. Hypothesis 1: Assess if the linear regression model developed has an intercept of zero.

b. Hypothesis 2: Assess if the linear regression model has a slope of one.
c. Hypothesis 3: Assess if the measured and predicted distress/IRI represents the same population of distress/IRI using a paired t-test.

The level of significance used was 0.05. A rejection of any hypothesis indicates that the model is biased.

Non-Statistical Approach

This approach was used when the measured distress/IRI was zero or close to zero for the sections under evaluation. Comparisons between predicted and measured distress/IRI was conducted by categorizing them into groups. The evaluation consisted on determining how often measured and predicted distress/IRI remained in the same group. This is an indication of reasonable and accurate predictions without bias.

B.5 Methodology used in Efforts for Northeastern States (*15*)

1. Version 1.1 of the MEPDG was utilized.
2. Seventeen LTPP pavement sections from GPS-1 and GPS-2 projects in the northeastern (NE) region of the United States were selected for use in the calibration procedure to best represent conditions in New York State. LTPP sites from the following states were chosen: Connecticut, Maine, Massachusetts, New Jersey, Pennsylvania, and Vermont.
3. Verification was completed by executing the MEPDG models with the default, nationally calibrated coefficients and comparing the predicted distresses with the measured distresses for each model.
4. Verification exercises were performed for the following models:
    a. Permanent deformation model (rutting)
    b. Bottom-up fatigue (alligator) cracking model
    c. Top-down fatigue (longitudinal) cracking model
    d. Smoothness (IRI) model
    e. Transverse (thermal) cracking model
5. Calibration was performed by minimizing the difference between predicted and measured distress values represented by the sum of the squares of the error (SSE).
6. Local calibration was performed for four of the five models evaluated:
    a. Alligator cracking
    b. Rutting
    c. IRI
    d. Longitudinal cracking
7. This work was reported in the form of a thesis conducted at University Texas at Arlington and was sponsored by New York State Department of Transportation (*15*).

B.6 Methodology used in Efforts for North Carolina

Previous Level 3 verification was conducted using MEPDG Version 1.0 to determine if the national calibration coefficients could capture the rutting and alligator cracking on North Carolina asphalt pavements (*16*).

More recently, verification was conducted as part of a calibration effort completed in 2011 (*17*). The following details the methodology utilized in that effort.

1. Version 1.1 of the MEPDG was utilized.
2. Local calibration was performed for the following models:
    a. Permanent deformation (rutting)
    b. Alligator cracking
3. Material-specific calibration of the global field calibration coefficients (k_{r1}, k_{r2}, and k_{r3}) in the rutting model, as shown in equation A.1, was completed for the twelve most commonly used asphalt mixtures in North Carolina.
    a. Triaxial repeated load permanent deformation (TRLPD) testing was conducted at three temperatures, from which the permanent and resilient strain values were determined at each loading cycle, N.
    b. Plots were developed for the logarithm of the ratio of permanent strain to resilient strain, $log(\varepsilon_p/\varepsilon_r)$, versus $log(N)$, from which the slopes were employed in numerical optimization to, in turn, determine a constant that closely replicates TRPLD results.
    c. The constant, $log(A)$, was then plotted with the logarithm of the test temperature, $log(T)$, to determine the slope and intercept.
    d. The slope and intercept were then used to determine the material-specific calibration coefficients, k'_{r1}, k'_{r2}, k'_{r3} unique to each of the twelve common asphalt mixtures used in North Carolina.
4. Material-specific calibration was also conducted for the fatigue model coefficients, (k_{f1}, k_{f2}, and k_{f3}) as shown in equation A.12 for the same twelve asphalt mixtures as were considered for the rutting coefficient calibration.
    a. Viscoelastic continuum damage (VECD) fatigue testing was conducted at various temperatures and strain levels.
    b. The simplified VECD (S-VECD) model was applied to the results to simulate strain-controlled direct tension fatigue testing and traditional bending beam fatigue testing.
    c. Material-specific coefficients were determined from the empirical model developed for the direct tension fatigue testing simulation.
5. Using the same twelve common asphalt mixtures and the associated material-specific calibration coefficients, local calibration was then conducted by calibrating coefficients for the rutting model, β_{r1}, β_{r2}, β_{r3}, β_{s1} (for granular base, and subgrade) and the alligator cracking model, β_{f1} β_{f2}, β_{f3}, C_1, and C_2. Two approaches were taken in the local calibration efforts:

a. Approach I involved using a large factorial of calibration coefficients: $\beta_{r2}$and β_{r3} for the rutting model, and β_{f2} and β_{f3} for the alligator cracking model, and executing the software numerous times for each model. In doing so, the remaining calibration coefficients were optimized using Microsoft Excel Solver by determining the combination of coefficients that produced the smallest sum of squared errors (SSE) between measured and predicted distress. This procedure was conducted for each model investigated.
b. Approach II involved a simultaneous optimization procedure. This procedure included the genetic algorithm (GA) optimization technique conducted using MATLAB® such that all model coefficients were optimized simultaneously.

6. Since MEPDG Version 1.1 does not incorporate material-specific coefficients for the rutting model, a hybrid version of the MEPDG was developed and utilized to account for material specific rutting calibration coefficients.
7. Model verification, calibration, and validation were evaluated with the following parameters:
    a. The standard error of the estimate, S_e
    b. The ratio of the standard error of the estimate (also referred to as the standard deviation of the residual error) to the standard deviation of the measured performance, S_y described by S_e/S_y
        i. "A ratio that is smaller than one indicates that the variability in the predicted residual error is smaller than that in the measured data."
    c. Coefficient of determination, R^2
    d. The p-value for the null hypothesis
        i. The null hypothesis, H_0, was that the average bias, or residual error, between the predicted and measured values is zero at the 95% confidence level, as described below:

$$H_0: \sum(Measured - Predicted) = 0$$

8. Level 2 inputs were utilized for asphalt mixtures and subgrade materials and Level 3 inputs (national default values) were applied for unbound base materials.
9. Twenty-two LTPP sites, 6 SPS sites, and 16 GPS sites were utilized for calibration of the rutting and alligator cracking models.
10. Twenty-four non-LTPP asphalt pavement sections were reserved for validation of the locally calibrated models; as a result, 25 distress datasets from 1993 and 1999 were utilized in addition to more recent distress data collected in 2010.

B.7 Methodology used in Efforts for Ohio (*18*)

1. Version 1.0 of the MEPDG
2. Globally calibrated models were evaluated and re-calibration was conducted using data from 13 LTPP projects.
3. Verification exercise conducted for new or reconstructed flexible pavements to determine if nationally (globally) calibrated models were sufficient in predicting

performance for selected pavements in Ohio with available and high-quality traffic, foundation, design, materials, and performance data.
    a. 13 LTPP projects were used with AC thicknesses of 4 or 7 inches.
    b. Two LTPP flexible pavement project categories: SPS-1 and SPS-9
    c. One location in the state was selected: a 3.3 mile section of US 23 in Delaware County, about 25 miles north of Columbus.
4. Asphalt concrete material properties hierarchical input levels for MEPDG calibration:
    a. HMA dynamic modulus Level 2, using LTPP data
    b. HMA creep compliance and Indirect Tensile Strength Level 3 (from Ohio MEPDG related literature or MEPDG defaults)
    c. Volumetric properties, Level 1 from LTPP
    d. HMA coefficient of thermal expansion, Level 3 (using MEPDG defaults)
    e. Unit weight, Level 1, determined from LTPP data
    f. Poisson's ratio, Level 1 and 3, computed from (from Ohio MEPDG related literature or MEPDG defaults)
    g. Other thermal properties; conductivity, heat capacity, surface absorptivity Level 3 (MEPDG defaults)
5. The following distresses were evaluated:
    a. Load-related alligator cracking, bottom initiated cracks
    b. Total rut depth
    c. Transverse "thermal" cracking
    d. Smoothness (measured as International Roughness Index [IRI])
6. Model adequacy was evaluated either by non-statistical methods or statistical methods, for statistical methods the following procedure was followed:
    a. MEPDG executed for each LTPP projects to predict distresses and IRI
    b. Predicted distress and roughness data (IRI) extracted for comparison with measured LTPP distress/IRI
    c. Statistical analyses conducted to check adequacy of performance models (prediction capability, accuracy, and bias)
    d. Where necessary, local calibration of MEPDG models conducted
    e. Sensitivity analyses of recalibrated models performed
    f. Results of verification of global models and local calibration (including revised model coefficients) are summarized
7. Model prediction capability was assessed by determining the correlation between measured and predicted distress/IRI using coefficient of determination, R^2. The reasonableness of the estimated R^2 was determined based on the following categories, such that a poor correlation implied the model was not predicting distress or IRI reasonably and may require recalibration:
    a. Excellent: > 80%
    b. Very good: 75 to 85%
    c. Good: 65 to 75%
    d. Fair: 50 to 65%
    e. Poor: < 50%

8. Model accuracy was assessed by the standard error of the estimate (S_e), which was taken as the square root of the average squared error of prediction. For S_e much greater than that reported from the NCHRP 1-40D distress/IRI predictions, recalibration was necessary.
9. Bias was determined by a linear regression using measured and predicted distress/IRI. Hypothesis tests were conducted for the following hypotheses using a significance level, α, of 0.05 or 5%.
    a. Hypothesis test 1: determine whether the linear regression model has an intercept of zero
        i. The null (H_0) and alternative hypothesis (H_A) are
            1. H_0: Model intercept = 0
            2. H_A: Model intercept ≠ 0
        ii. If the p-value < 0.05, the null hypothesis is rejected and it is implied that the linear regression model had an intercept significantly different from zero at the α-level, therefore, the prediction model is biased and should be recalibrated.
    b. Hypothesis test 2: determine whether the linear regression model has a slope of 1.0
        i. The null (H_0) and alternative hypothesis (H_A) are
            1. H_0: Model intercept = 1.0
            2. H_A: Model intercept ≠ 1.0
        ii. If the p-value < 0.05, the null hypothesis is rejected and it is implied that the linear regression model had a slope significantly different from 1.0 at the α-level, therefore, the prediction model is biased and should be recalibrated.
    c. Hypothesis test 3: determine whether the measured and predicted distresses/IRI represent the same population of distress/IRI using the following paired t-test:
        i. The null (H_0) and alternative hypothesis (H_A) are
            1. H_0: mean measured distress/IRI = mean predicted distress/IRI
            2. H_A: mean measured distress/IRI ≠ mean predicted distress/IRI
        ii. If the p-value < 0.05, the null hypothesis is rejected and it is implied that the measured and MEPDG distress/IRI are from different populations at the α -level, therefore the MEPDG distress/IRI predictions are biased and should be recalibrated.

B.8 Methodology used in Efforts for Oregon (*19*)

1. Darwin M-E Version 1.1
2. Since pavement work conducted by ODOT involves the rehabilitation of existing pavements, calibration was conducted for rehabilitation of existing structures.
3. Pavement sections were selected based on location (Coastal, Valley, and Eastern), type (HMA over aggregate base, HMA inlay or overlay over aggregate base, HMA inlay or

overlay over cement treated base, and HMA overlay of CRCP), traffic level (low and high) and pavement performance (very good/excellent, as expected, and inadequate).
4. Primary effort for calibration was on a Level 3 analysis.
5. The authors do not clearly specify how many data points were used to conduct local calibration of the different models, and the only statistic reported was the standard error of the estimate (S_e) before and after calibration.
6. The IRI model was not calibrated.

B.9 Methodology used in Efforts for Tennessee (*20*)

1. Version 1.100 of the MEPDG was utilized for an initial verification of the rutting and roughness models for the design of new AC pavements.
    a. Accuracy of the predictive performance models was evaluated using HMA Dynamic modulus Level 1 (obtained from laboratory testing) and Level 3 (estimated using the Witczak model).
2. Nineteen HMA pavement sections were utilized for initial verification, including 18 interstate highway pavement sections and one state route pavement section.
    a. Two sections were AC pavements.
    b. Six sections were PCC pavements that had received an AC overlay.
    c. The remaining sections included AC pavements that had received an AC overlay
3. Two input levels were defined and analyzed:
    a. "Level 2.5" considered Level 3 inputs for AC layers and Level 2 for base and subgrade properties.
    b. "Level 1.5" considered Level 1 for material properties of AC layers and Level 2 inputs for the base and subgrade.
4. Roughness is characterized by Tennessee DOT (TDOT) in terms of PSI, rather than by IRI, therefore, initial IRI was determined from the average initial PSI and an IRI-PSI relationship previously developed.
5. Local calibration was attempted on the permanent deformation (rutting) transfer function for asphalt pavements and comparisons were drawn between predicted and measured rutting by grouping pavement sections by traffic levels.
    a. In verifying and calibrating the rutting model, three different categories of pavements were considered: asphalt pavements and asphalt pavement overlaid with AC; concrete pavements overlaid with AC for low volume traffic (0-1,000 AADTT); and concrete pavements overlaid with AC for heavy traffic (1,000-2,500 AADTT).
6. Using Microsoft Solver, elimination of bias was attempted by minimizing the S_e between measured and predicted rutting values by varying the β_{1r} and β_{s1} (for base and subgrade) coefficients.
    a. Comparisons were drawn between predicted and measured rutting for the two pavement types and then further broken down by traffic levels (Annual Average Daily Truck Traffic (AADTT) between 0 and 1,000, and AADTT between 1,000 and 2,500).

7. A total of 18 pavement sections were utilized in the calibration effort using Version 1.100 of the MEPDG.
    a. Six pavement sections included PCC pavements that had received an AC overlay.
    b. Ten pavement sections included AC pavements that had received an AC overlay.
        i. The AC overlays ranged in thickness from two to eight inches.
    c. Two new AC pavement sections were included.
8. Calibration of the sections with an AC overlay atop PCC pavements included:
    a. Level 1 inputs were used for the dynamic modulus of the asphalt mixtures and the complex modulus of the asphalt binders.
    b. Level 3 inputs utilized for the PCC, base, and subgrade layers.
    c. Because no rutting occurred in the PCC and underlying layers, only rutting in the asphalt overlay was considered and compared with predicted rutting.
    d. Calibration coefficients were varied using Microsoft Excel Solver to minimize S_e.
    e. Due to limited number of PCC pavements overlaid with AC, no validation effort was conducted.
9. Calibration of the sections with an AC overlay atop AC pavements and the new AC pavements included the following:
    a. AC pavements overlaid with AC were assumed as new pavements (the AC overlay pavements and new AC pavement design utilized the same rutting transfer model in the MEPDG).
    b. The level of inputs for the asphalt layers was not explicitly stated.
    c. Total predicted rutting was compared with measured rutting.
    d. Calibration coefficients were varied using Microsoft Excel Solver to minimize S_e.
    e. Validation was conducted using two additional AC pavement sections that had been overlaid with AC.

B.10 Methodology used in Efforts for Utah (*21, 22*)

1. Version 0.8 of the MEPDG (Report date: October 2009)
    a. HMA dynamic modulus Level 3 (Computed using material gradation, binder grade and Witczak model)
    b. HMA creep compliance & indirect tensile strength Level 3 (Based on binder type and material type)
    c. Effective binder content- Level 3 (Estimated from past construction projects QA/QC)
    d. HMA coefficient of thermal contraction Level 3 (MEPDG defaults)
    e. Base type/Modulus-Level 3, based on material type
    f. Subgrade type (Level 1, backcalculated using deflection data)
2. HMA sections included:
    g. New HMA (including thin overlays): 26 projects
    h. HMA overlaid existing HMA: 4 projects
3. All HMA thicknesses were included although most sections were between 4 and 8 inches.

4. Local calibration was conducted using linear and non-linear regression procedures (SAS statistical software). Optimization was performed to select local calibration coefficients to maximize R^2 and minimize S_e, goodness of fit and bias was checked, and limited sensitivity analysis was performed.
5. Recalibration of the rutting models was conducted in 2013. The recalibration analysis was conducted in the same manner and using most of the test sections used in the 2009 calibration with four more years of rutting data.

B.11 Methodology used in Efforts for Washington (*23*)

1. Version 1.0 of the MEPDG (Paper date: July 2009)
2. It was reported that the calibration process follows a combination of a split-sample approach and a jackknife testing approach per recommendation in the draft report to NCHRP Project 1-40A (Recommended Practice for Local Calibration of the ME Pavement Design Guide).
3. For the calibration procedure, data from the Washington State Pavement Management System (WSPMS) was used.
4. Their calibration efforts focused on fatigue damage, longitudinal cracking, alligator cracking, and rutting models.
5. An elasticity analysis was conducted to assess the effect of the different calibration factors on the distress models. Two representative calibration sections were used.
6. Final calibration factors are reported, but no statistics are presented.

B.12 Methodology used in Efforts for Wisconsin (*24*)

1. Version 1.0 of the MEPDG (Report date: April 2009)
2. The study was conducted using information collected from the LTPP sections in Wisconsin.
3. Most of the inputs required for the verification and calibration were from the LTPP database with some inputs being national defaults in the software.
4. A design was conducted for each LTPP section to predict pavement distresses and IRI.
5. The Pavement ME Design predictions were then compared with the measured distresses in the LTPP database to develop recommendations for Wisconsin DOT.
6. Calibration was attempted for rutting, transverse cracking and IRI models.
7. Work was sponsored by Wisconsin DOT and completed by ARA in 2009.

APPENDIX C SUMMARY OF VERIFICATION AND CALIBRATION RESULTS

C.1 Fatigue Cracking (Alligator/Bottom-up)

C.1.1 Arizona (10)

Verification of the Darwin ME global alligator cracking models consisted of running the MEPDG with the global coefficients for all selected projects and evaluating goodness of fit and bias. A total of 363 data points were used. "Obvious bias" in the form of under-prediction was reported. The goodness of fit was reported as very poor with R^2 of 8.2% and S_e of 14.3%. Based on this finding, local calibration was recommended. Local calibration was conducted using 419 data points, which included both the calibration and validation datasets, and the authors reported adequate goodness of fit with an R^2 of 50% and S_e of 14.8%. The bias was reduced if not eliminated through local calibration. It should be noted that the authors of the study for Arizona DOT reported different statistics in Chapter 3 and 8. From Chapter 3 (Page 61, Table 20) R^2 was reported as 50% and S_e was listed as 14.8% of the lane area. In Chapter 8 (Page 175, Table 65), R^2 was reported as 58% and S_e was shown as 13% of the lane area (*10*).

C.1.2 Colorado (11)

The evaluation of the global alligator cracking models consisted of running the MEPDG with the global coefficients for all selected projects and evaluating goodness of fit and bias. The goodness of fit was poor with an R^2 of 17.5 percent, which indicates a weak alligator cracking prediction. The bias was reported in terms of p-value for paired t-test and slope (see statistics available in Table 5). Both terms indicated bias in the prediction. The model consistently under-predicted alligator cracking with increasing HMA fatigue damage. Since the fatigue model did not adequately predict alligator cracking for Colorado conditions, local calibration was recommended. A total of 50 data points were used for this effort.

The local calibration of HMA alligator cracking and fatigue damage models was conducted using the same goodness of fit and bias parameters that were reported for verification. The results showed an adequate (fair) goodness of fit with minimal bias. For the fatigue damage and alligator cracking local models, 56 data points were used and for the reflection cracking model 87 data points.

C.1.3 Iowa (13)

The global alligator cracking model was evaluated by plotting the measured against the predicted cracking and by calculating the bias and standard error for the global model. The global model resulted in good estimates of the measured alligator cracking, with only two of the 327 data points underestimated by the model. Researchers concluded that the nationally calibrated model for alligator cracking did not require local calibration.

C.1.4 Missouri (14)

The alligator cracking data showed that approximately 99% of all measured alligator cracking had a value less than 5% (by lane area). Because most of the projects reported no alligator cracking, a non-statistical approach was used to verify the model. In order to verify the adequacy of the global calibration coefficients, cracking was categorized into eight groups to

determine how often measured and predicted cracking fell in the same group. The authors reported that the global model seemed to both slightly under-predict and slightly over-predict alligator cracking. Based on this information, local calibration was not recommended at the time the study was conducted and it was recommended that the model be reevaluated when data show higher magnitudes of alligator cracking.

C.1.5 Northeastern States (15)

Verification of the nationally calibrated model was conducted using measured fatigue cracking from the LTPP sites in the northeastern region. However, the amount of data points considered for this evaluation and for the regional calibration was not reported. The predicted fatigue cracking using the nationally calibrated model resulted in sum of square of the errors (SSE) of 63.48. By plotting the measured fatigue cracking with predicted fatigue cracking from the nationally calibrated model, it was concluded that the global model "severely under-predicts the extent of alligator cracking." As a result, calibration was conducted for the fatigue cracking model using Microsoft Excel Solver to determine the calibrated coefficients, C1, C2, and C4, by minimizing SSE. No other statistical measures were used to evaluate the goodness of fit of either the nationally calibrated model or the locally calibrated model. An improvement in the prediction of fatigue cracking was reported for the regionally calibrated model with a reduction in SSE from 63.48 to 43.48.

C.1.6 North Carolina

Initial verification of the nationally calibrated alligator cracking model was reported in 2008 (*16*). For this effort, the MEDPG Version 1.0 was utilized to evaluate the predicted alligator cracking relative to measured alligator cracking. New and rehabilitated asphalt pavements from LTPP (only sections that were not included in the national calibration) and non-LTPP pavement sections in North Carolina were considered. Considering only LTPP pavement sections, a total of 76 data points were used to evaluate the model. Standard error and SSE were reported as 10.7% and 8,505.51, respectively. When both LTPP and non-LTPP pavement sections were considered, the dataset was expanded to 176 data points and resulted in S_e of 6.02% and SSE of 29,487.1. The nationally calibrated model was found to under-predict alligator cracking and "a significant amount of bias" was reported for the default model.

Verification was also conducted en-route to performing local calibration in efforts completed in the more recent 2011 study, using Version 1.1 of the MEPDG software (*17*) for which results are reported in Table 5 of this report. Additional parameters that were reported for the application of the nationally calibrated model include total SSE of 56,412, bias of -11.034 (percent of the lane area) and S_e/S_y of 1.022. The authors of the 2011 study reported the global calibration coefficients resulted in significant under-predictions of the mean measured fatigue cracking.

Material-specific coefficients (k_{f1}, k_{f2}, and k_{f3}) calibrated for each of the twelve most common asphalt mixtures used in North Carolina were used in calibrating the alligator cracking model for North Carolina (*17*). Two approaches were taken to develop local calibration model coefficients, β_{f1}, β_{f2}, β_{f3}, C_1, and C_2. It was found that the locally calibrated coefficients from Approach II-F improved the predicted mean alligator cracking and reduced the bias found when using the

nationally calibrated coefficients. Approach II-F resulted in statistically better predictions of alligator cracking and as a result, coefficients from Approach II-F were recommended. The calibration coefficients and statistical parameters resulting from Approach II-F are listed in Table 6. Despite the improved predictions, it was found that differences between measured and predicted fatigue cracking were still significant at the 95% confidence level, thus rejecting the null hypothesis. Additional parameters were reported describing the adequacy and bias of the calibrated model: the total SSE of 38,752, bias of -5.153 (percent of lane area) and S_e/S_y of 0.949.

Validation of locally calibrated models from Approach I-F and II-F with use of the material-specific coefficients was conducted with sections from the pavement management system of the North Carolina Department of Transportation (*17*). For the application of the recommended locally calibrated coefficients determined from Approach II-F to the validation dataset, it was reported that both bias and standard error were reduced relative to the calibration and verification results. As was the case with the calibration set, the null hypothesis was also rejected when the local calibration coefficients were applied to the validation set, indicating that there were statistically significant differences between the measured and predicted fatigue cracking values. Statistical parameters for the validation of Approach II-F calibration coefficients are also listed in Table 6. Additional parameters reported but not listed in Table 6 include bias of 1.973 (percent of lane area) and S_e/S_y of 1.690.

C.1.7 Ohio (18)

The alligator cracking model was not evaluated or recalibrated due to premature longitudinal cracking associated with construction defects in the LTPP project sites.

C.1.8 Oregon (19)

The verification of the global calibration factors using Darwin ME software showed that the software under estimates the amount of alligator cracking, therefore, local calibration was recommended. After calibration, the authors indicated that the model was improved in terms of bias and standard error, but there was a high degree of variability between the predicted and measured distresses. The authors only reported S_e, with a value of 3.384 before calibration and 2.644 after calibration; no other statistics were presented. The number of data points used to conduct local calibration was not explicitly stated, however, it was noted that only one year of distress data was available for the verification and calibration procedures.

C.1.9 Utah (21)

The alligator cracking data showed that 95% of the data points had less than 2% of cracking (% lane area). Due to the limited data available, a non-statistical approach was used to verify the model. In order to verify the adequacy of the global calibration coefficients, cracking was categorized into eight groups to determine how often measured and predicted cracking fell in the same group. It was found that the global model predicted cracking relatively well, but the model could not be evaluated for sections with significant amount of cracks; therefore, local calibration of the model was not required.

C.1.10 Washington (23)

Verification of the alligator cracking model was conducted by plotting MEPDG predictions using default calibration factors versus Washington State Pavement Management System (WSPMS) data. No statistics were reported on the prediction capability of the default alligator cracking model, however, the authors did state that it tended to under-predict alligator cracking.

The elasticity analysis identified the individual calibration coefficients that needed to be recalibrated. Once recalibrated, alligator cracking predictions made with the locally calibrated model were plotted with time. On the same plot, WSPMS alligator cracking data were also plotted over time. It was found that the locally calibrated model resulted in predictions that matched well with WSPMS data, but no statistics were presented.

The authors explained a key assumption in their evaluation: since alligator cracking is located in the wheelpath area (about half of the total lane area) and the MEPDG considers alligator cracking as a percent of the total lane area, WSPMS alligator cracking was divided by two. Based on this assumption, MEPDG estimation should be greater than or equal to WSPMS corrected values.

C.1.11 Wisconsin

Verification was completed using inputs from the LTPP database in Version 1.0 of the MEPDG and performance data of the LTPP sections. A design was conducted for each section, and predicted cracking was compared with field performance, from which it was reported that alligator cracking predictions were reasonable for pavements that were less than 10 years old. There were not enough data to arrive at any conclusions for moderately to highly distressed pavements (*24*).

C.2 Rutting

C.2.1 Arizona (10)

Verification of the rutting model was conducted with a total of 479 data points. It was found that the goodness of fit statistics was very poor (R^2 = 4.6 % and S_e = 0.31 inch), with obvious bias (large over-prediction) when using the national default rutting model.

A total of 497 data points were used in the calibration of the three rutting submodels. The calibration coefficient, β_{r1}, was reduced from the default value of 1.0 to 0.69. Additionally, new calibration coefficients, β_{s1} and β_{b1}, were determined as a result of the calibration of the base and subgrade rutting submodels. Despite the calibration effort, the goodness of fit remained poor with R^2 of 16.5%, while the standard error was reduced to an S_e of 0.11 in. The bias (over-prediction) found in the application of the national default model was removed as a result of the local calibration. The authors attributed the poor goodness of fit to excessive variability in the year-to-year measured rutting and not to major weakness in the models.

Several important discrepancies were discovered in the literature. First, the Manual of Practice reports global calibrations factors for k_{2r} as 0.4791 and k_{3r} as 1.5606, however the report

detailing the verification and calibration effort for Arizona reports show k_2 as 1.5606 and k_3 as 0.4791 (*4, 10*). Secondly, Arizona reported different statistics for the locally calibrated model in Chapter 3 and 8. From Chapter 3 (Page 67, Table 23) R^2 was listed as 16.5% and S_e was reported 0.11 in, however, in Chapter 8 (Page 175, Table 65), R^2 was shown to be slightly higher at 21% and S_e was reported as 0.12 in (*10*).

C.2.2 Colorado (11)

Since the MEPDG predicts HMA pavement total rutting using separate submodels for the surface HMA, granular base, and subgrade, evaluation of the global total rutting model consisted of the following steps:

a. Run the three rutting submodels using global coefficients for all sections to obtain estimates of total rutting.
b. Perform statistical analysis to determine goodness of fit and bias in estimated total rutting.
c. The goodness of fit was evaluated using the same parameters that were used for alligator cracking.

A total of 155 data points were used to evaluate the national default rutting submodels. It was reported that there was "significant bias" when the nationally calibrated rutting model predictions were compared with measured rut depth. In the plot of measured versus predicted rutting shown in Figure 82 (*11*), it appears that the global model over-predicts for some magnitudes of rutting and tends to under-predict total rutting for other magnitudes (high and low). Rutting prediction was poor, therefore, local calibration was recommended.

Local calibration of the HMA rutting, unbound aggregate base rutting, and subgrade rutting global model coefficients was conducted using a total of 137 data points. After calibration, reasonable predictions of rutting were obtained. It was also reported that the significant bias associated with the nationally calibrated model was eliminated through local calibration, based on the reported p-values.

It should be noted that the k_{r2} and k_{r3} values are reported differently in the body of the text, specifically Table 59 of the report, which shows values consistent with the Pavement ME Design software Version 2.1. However, in the Appendix of the referenced report (*11*), the values are reversed and as such are consistent with the values reported in the Manual of Practice (*4*). The values listed in Tables 7 and 8 are the values listed in Table 59 of the referenced document, which are consistent with the current version of the software.

C.2.3 Iowa (13)

Comparisons were drawn between predicted and measured rutting in each layer (HMA, granular base, and subgrade) and total rutting. The nationally calibrated model was found to under-estimate HMA layer rutting while over-predicting rutting in the granular base and subgrade. The total rutting prediction bias and standard error for the nationally calibrated model were reported as 0.05 and 0.08, respectively.

As a result of local calibration, reductions in both bias and standard error were reported. Overall improvements relative to the globally calibrated model in rutting predictions for each layer and total rutting were also reported for the locally calibrated rutting model. Bias and standard error for the locally calibrated model were reported as 0.03 and 0.07, respectively, for the total rutting prediction.

Similar to the results reported for the calibration dataset, predictions made for the validation dataset using the locally calibrated coefficients showed good agreement with measured values and an improvement over the globally calibrated rutting model in terms of bias and standard error. However, as was the case with the calibration dataset, there is limited data available for the evaluation of rutting in the granular base layer and the subgrade layer, although the number of data points in the validation dataset was not reported. Applying the nationally calibrated model to the validation dataset resulted in a bias and standard error of 0.04 and 0.07, respectively, for the total rutting prediction. The application of the locally calibrated model to the validation dataset resulted in bias and standard error of 0.02 and 0.07, respectively, for the total rutting prediction. As a result, the locally calibrated rutting models were recommended for use over the existing global models.

C.2.4 Missouri (14)

The verification of the national rutting model was conducted using 183 data points. It was found that the model over-predicts total rutting and a poor correlation was observed (R^2=0.32 and S_e=0.11 in). Local calibration of the three rutting submodels (HMA, base and subgrade) was conducted using the same data points.

Calibration was conducted by modifying the β_{r1} in the HMA submodel, β_{s1} for the unbound base submodel, and β_{s2} for the subgrade submodel. After local calibration, a fair correlation between measured and predicted rutting was reported (R^2=0.52 and S_e=0.051 in). It should be noted that the coefficient of determination was listed in the plot of measured versus predicted total rutting for the locally calibrated model (Figure I-124) as 53% and 52% in Table I-69 of the reference document (*14*). Values listed in Table 8 of this report are consistent with Table I-69 of the reference document (*14*), which described the statistics for the locally calibrated model.

C.2.5 Northeastern States (15)

Comparisons were drawn between measured rutting and predicted rutting in the AC, base, subgrade, and total rutting using the nationally calibrated model. The proportion of predicting rutting in each layer to the total rutting was determined and applied to measured total rutting to draw comparisons for total rutting, rutting in the AC, base, and subgrade. The authors did not explicitly state whether under- or over-prediction was noted with the national model. However, based on the plot of measured versus predicted total rutting using the global coefficients (shown in Figure 4.7), it appears the national default rutting model under-predicted total rutting at the high end (above 0.4 inches).

Calibration was completed for the permanent deformation model by using the nationally calibrated model to determine the ratio of predicted deformation to total rutting in each layer.

Once the proportions were determined, the measured total rut depth was multiplied by the corresponding ratio to estimate the measured rutting in each layer. Calibration was then performed using a simple linear regression with no intercept, where the measured rutting served as the independent variable. The calibration coefficients were then determined as the inverse of the slope. The number of data points used for the comparison were not reported, nor were any statistical measures reported relative to the accuracy or goodness of fit of the nationally and locally calibrated models. However, it was reported that the SSE for total rutting decreased with local calibration and that the regional calibration coefficients give a better fit between measured and predicted rutting in all layers. It was suggested that the much larger regional coefficients could be due to the small data set of 17 sections.

C.2.6 North Carolina

Initial verification of the nationally calibrated rutting model was reported in 2008 (*16*). MEDPG Version 1.0 was utilized to evaluate the predicted total rutting, AC rutting, rutting in the granular base, and rutting in the subgrade relative to measured rutting. New and rehabilitated asphalt pavements from LTPP (only sections that were not included in the national calibration) and non-LTPP pavement sections in North Carolina were considered. By including only LTPP pavement sections, a total of 161 data points were used to evaluate the model. The R^2, S_e, and SSE for the comparison of measured rutting with predicted total rutting using the nationally calibrated coefficients were reported as 0.340, 0.111, and 1.962, respectively. When both LTPP and non-LTPP pavement sections were included, the dataset consisted of 255 data points. Applying the nationally calibrated coefficients and comparing with measured rutting resulted in the following parameters: R^2 of 0.142, S_e of 0.153, and SSE of 10.387. The predicted rut depths matched well with the measured rut depths when only LTPP pavement sections were considered.

Verification was also conducted as part of local calibration efforts completed in 2011 using MEPDG Version 1.1 (*17*). In applying the nationally calibrated model, it was found that it under-predicted the total rut depth as well as the rut depth in the HMA layers. While the differences reported for the total and HMA rut depths were found to be statistically significant at the 95% confidence level, the difference in predicted and the (estimated) measured rut depths for the unbound base and subgrade were not found to be statistically significant. Although comparisons were made for the individual models (HMA rutting, base rutting, and subgrade rutting) as well as the total rutting model, only the results for the total rutting are reported in Table 7. In addition to the parameters shown in Table 7, the following parameters were also reported: total SSE of 4.110, bias of -0.031, and S_e/S_y of 1.027.

The material-specific calibration coefficients were used in the local calibration process to calibrate β_{r1}, β_{r2}, β_{r3}, β_{gb}, and β_{sg} for North Carolina (*17*). Two approaches were taken in the local calibration process, Approach I-R and II-R. The locally calibrated coefficients from Approach II-R were reported to "significantly reduce bias and standard error between the predicted and measured rut depth values for all layers types, except for the subgrade." Bias for total rutting was also reduced through the local calibration coefficients resulting from Approach II-R. Furthermore, Approach II-R resulted in improvements over the nationally calibrated model

for all of the statistics evaluated, therefore, coefficients determined in Approach II-R were recommended. Although comparisons were made for the individual models (HMA rutting, base rutting, and subgrade rutting) as well as the total rutting model, only the results for the total rutting predicted using Approach II-R coefficients are listed in Table 8. The following parameters were also reported for the predicted total rutting using the recommended local calibration coefficients from Approach II-R relative to the measured rutting: total SSE was 3.604, bias of -0.021, and S_e/S_y of 0.975. Additionally, it was reported that at the 95% confidence level, there were differences between the measured and predicted total rut depth values.

Validation of locally calibrated models from Approach I-R and II-R with use of the material-specific coefficients was conducted with sections from the pavement management system of the North Carolina Department of Transportation (*17*). For the recommended locally calibrated coefficients determined from Approach II-R, it was reported the calibrated coefficients over-predicted the total rut depth in the validation dataset. It was found that the differences between predicted and measured total rut depth were statistically significant at the 95% confidence level for the validation set. Parameters regarding the model adequacy and bias when applied to the validation dataset are also listed in Table 8. Additional parameters that were reported included bias of 0.248 and S_e/S_y of 2.317.

C.2.7 Ohio (18)

A statistical comparison of measured total rutting and predicted total rutting using the nationally calibrated rutting model was conducted to evaluate the total rutting predictions (*18*). The coefficient of determination was calculated and reported as 64%, falling into the fair category, although it was reported that the there was a "poor correlation between measured and MEPDG predicted rutting" (*18*). Using a linear regression line for the measured and predicted total rutting plot, three hypothesis tests were conducted to determine if bias exists in the predicted rutting models. The hypothesis tests looked at the slope and intercept of the linear regression model and if the measured and predicted total rutting represent the same population. All three tests rested in the rejection of the null hypothesis, indicating bias existed in the nationally calibrated model. Over-prediction of the total rut depth is apparent in the plot of measured versus predicting total rutting shown in Figure 21 of the referenced document (*18*).

Local calibration was attempted due to the bias in the form of over-prediction of total rutting and poor correlation between measured and predicted rutting. First, a thorough review of the rutting submodel (HMA, base, and subgrade) predictions was completed to check for reasonable predictions based on engineering judgment. From this, it was found that rutting in the HMA layer contributed an unreasonable amount to the total rutting despite relatively thick asphalt concrete layers, indicating a need to adjust the rutting accumulated in the unbound layers. Local calibration was completed by modifying the local calibration coefficient, β_{1r}, of the HMA rutting submodel, and β_{S1} and β_{S2} of the base and subgrade rutting submodels.

Once local calibration was complete, statistical comparisons between measured and predicted rutting using the recalibrated submodels were conducted. The coefficient of determination was

calculated and showed a fair correlation (R^2 of 0.63). The three hypothesis tests were completed to check for model bias, revealing significant bias remained in the model despite recalibration. Specifically, the intercept of the linear regression model was statistically significantly different than zero and the population of measured totally rutting was statistically significantly different than the MEPDG predicted rutting. S_e was calculated to check for model accuracy, which showed an accurate model in the sense that the S_e (0.014 in.) was much less than that reported for the global total rutting model. Due to the noted bias after local calibration, it was recommended that a larger data set be used in calibration and a more comprehensive set of HMA pavement mixtures in Ohio be evaluated.

C.2.8 Oregon (19)

The authors reported that in Oregon, rutting in the base and subgrade layers is not a problem, since most of the rutting comes from the HMA layers only. The approach was to set the calibration factors to zero for base and subgrade layers. No hypotheses testing was conducted to determine if bias was present, however, the authors reported that bias was reduced through local calibration. The only statistic that was reported, S_e, was found to be 0.568 before calibration for the global model, and it was reduced to 0.180 after calibration. In evaluating the global model it was also reported that the much of the estimated rutting was predicted for the subgrade layer.

C.2.9 Tennessee (20)

An initial verification of the nationally calibrated rutting models was attempted for both AC pavements overlaid with AC and PCC pavement overlaid with AC, ignoring rutting in the base and subgrade for AC overlays on PCC pavements. Comparisons between measured and predicted rutting were drawn for the two input levels considered ("Level 2.5" and "Level 1.5").
For the AC overlays on PCC pavements, the predicted rutting in the AC followed a similar trend with time as was observed for measured rutting. In looking at two input levels, "Level 1.5" resulted in more accurate rutting predictions, whereas "Level 2.5" tended to over-predict rutting in the AC overlay. For the AC overlays on AC pavements, the nationally calibrated models over-predicted total rutting for input "Level 1.5" and "Level 2.5." The predictions of rutting in the AC layers were higher than measured rutting for both input "Level 1.5" and "Level 2.5", although predictions at "Level 1.5" were reported to be more reasonable than the latter.

As part of the calibration effort, verification was completed such that predictions from the nationally calibrated model were compared with measured rutting. For AC overlays on PCC pavements, predictions with the global model were made using Level 1 inputs for the AC overlay and Level 3 inputs for the existing PCC pavement, base, and subgrade material. Only rutting in the AC overlay was considered where the overlay was placed on PCC pavements. Total rutting was considered for AC pavement and AC overlays on AC pavements. For PCC pavements with AC overlays, in looking at two traffic levels, low (0-1,000 AADTT) and high (1,000-2,500 AADTT), it was found that the low traffic levels resulted in poor rutting predictions that under-estimated rutting in the AC overlay. In the case of high traffic levels, predictions were reasonable relative to measured rut depths, therefore, calibration was deemed unnecessary. For AC overlays on AC pavements, comparisons between rutting predictions and

measured rut depths indicated that the nationally calibrated model over-predicted total rutting for AC pavements and AC overlays on AC pavements. Results from the verification effort are summarized in Table 7.

Based on the results of the more detailed verification effort, model coefficients were calibrated for each of three categories of pavements. For AC overlays on PCC pavements with low traffic, calibration was conducted by varying the β_{r1} coefficient, while β_{r2} and β_{r3} were held to their default value of 1.0; coefficients for β_{BS} and β_{SG} were assigned a value of zero. AC overlays on AC pavements and new AC pavements were analyzed together. The local coefficients, β_{r1}, β_{BS}, and β_{SG} were varied to minimize the S_e. Although it was reported in the verification effort that calibration was deemed unnecessary for AC overlays on PCC pavements with high traffic, calibration was attempted. The local calibration coefficient, β_{r1}, was varied in the calibration procedure and attempts to minimize the S_e only resulted in the same value (1.0, which is consistent with the nationally calibrated model), it was concluded that no local calibration was necessary for the high traffic volume. While not necessarily calibration, the authors later reported in Table 5.3 (page 125 of the referenced document) coefficients of zero for the β_{BS} and β_{SG} for AC overlays on PCC pavements with high traffic.

Results of the local calibration for the other two pavement categories are listed in Table 8. As a result of the local calibration for AC overlays on PCC pavements with low traffic levels, the β_{r1} increased from 1.0 in the nationally calibrated model to 2.20 in the locally calibrated model. For AC overlay on AC pavements and new AC pavements, the local calibration resulted in new coefficients for β_{r1}, β_{BS}, and β_{SG} and also improved the rutting predictions

C.2.10 Utah (21)

The verification of the national rutting model indicated that the rutting predictions were adequate for older pavement that used viscosity graded asphalt (68 data points), but were poor for newer pavement that used Superpave mixes (86 data points). Statistical parameters for the verification effort of the newer Superpave mixes are listed in Table 7 of this report. Local calibration of the three rutting submodels (HMA, base, and subgrade) was conducted to predict rutting for current HMA mixes using Superpave mix design and using mostly Level 3 inputs. After local calibration, a poor correlation between measured and predicted rutting was found. This was attributed to the measured rutting data obtained from the UDOT PMS database, which was measured using a laser system and converted to LTPP standards. No improvement was found on S_e, but the significant bias was eliminated.

C.2.11 Washington (23)

Verification of the MEPDG using global calibration factors showed that the rutting was under-predicted; hence, the next step was to conduct calibration. After calibration, the authors reported that the calibrated model estimation matched well with performance data in magnitude and progression. However, as it was mentioned in the description of methodologies, statistics were not reported. The authors also indicated that WSDOT does not typically

experience rutting in the base and subgrade; hence, rutting should be presented as only occurring in the surface layer. This was corrected by setting subgrade rutting factors to 0.

C.2.12 Wisconsin

Verification was completed in Version 1.0 using inputs and performance data of LTPP sections in Wisconsin. Predicted rutting was compared with field performance data, from which it was reported that the default calibration coefficients produced total pavement rutting predictions that were statistically different from rutting measured in the field. Thus, local calibration was then conducted to modify the local calibration coefficients of the HMA, base, and subgrade rutting models (*24*). The new model coefficients obtained through local calibration are as follows (*25*):

AC Rutting β_{r1} = 0.477
Granular Base Rutting β_{s1} = 0.195
Subgrade Rutting β_{s1} = 0.451

C.3 Thermal (Transverse) Cracking

C.3.1 Arizona (*10*)

Verification of the transverse cracking model using Level 3 was conducted for all HMA sections at the LTPP SPS-1 site in Arizona. The coefficient, K, for Level 3 was shown as 3.0 in the Appendix of the Arizona report (*10*). However, in discussions of the verification effort (page 71), K was listed as 1.5 for the Level 3 designs completed to evaluate the nationally calibrated model. Therefore, it is assumed a K-value of 1.5 was the value utilized in the nationally calibrated model. A non-statistical analysis was completed by comparing measured versus transverse cracking predicted using the global MEPDG model. The authors indicated that the software under-predicted transverse cracking for a range of HMA sections. As a result, local calibration was deemed necessary.

Local calibration was attempted at a Level 3 analysis by modifying the calibration parameter, K. Researchers varied the value of K (up to 100) until on average, predictive transverse cracking matched measured cracking. At a value of K = 100, predictions roughly matched measured cracking at high magnitudes but over-predicted at low magnitudes of transverse cracking. Predictions at K = 50 were not consistent for all climatic locations. The authors' recommendation was to not use transverse cracking as design criteria.

C.3.2 Colorado (11)

Model verification of the transverse cracking model was completed using 12 data points. Since the transverse cracking model is very sensitive, the authors recommended that only Level 1 HMA creep compliance and indirect tensile strength inputs be used for local calibration. The global model was found to under-predict measured transverse cracking. Since poor goodness of fit was found, calibration of the transverse cracking model was recommended.

Calibration was conducted using the same 12 data points and indicated that predictions were reasonable. By varying the K coefficient from 1 to 10 incrementally and evaluating the predicted

transverse cracking, researchers arrived at a locally calibrated coefficient of 7.5, which produced the best goodness of fit and the least bias. The resulting R^2 value of 43.1% was an improvement over the nationally calibrated model and was deemed adequate. A large increase in S_e was reported with the nationally calibrated model at 0.00232 ft/mi and the locally calibrated model at 194 ft/mi, however, researchers considered both values to be adequate. Additionally, it was reported that the significant bias found in the nationally calibrated model was eliminated with local calibration.

C.3.3 Iowa (13)

Iowa conducted transverse cracking model verification by plotting the measured transverse cracks in feet per mile against transverse cracks predicted by the nationally calibrated thermal cracking model. Bias and standard error were also reported. Although significant amounts of thermal cracking were measured in the field, the nationally calibrated thermal cracking model predicted minimal levels of thermal cracking. As a result, a large standard error of 1,203 and a large bias of -446 were reported. Due to the large disparity between measured and predicted thermal cracking, the thermal cracking model was not considered for local calibration.

C.3.4 Missouri (14, 27)

In verifying the transverse cracking model, two levels of inputs, Level 1 and Level 3, were considered for the HMA creep compliance and indirect tensile. A total of 49 data points were used to verify the model at both levels. When Level 3 designs were used, the authors reported that the MEPDG under-predicted the measured cracking; when a Level 1 design was used, predictions were slightly improved but still showed a significant bias. Although predictions were reported to improve when a Level 1 design was used, the R^2 and S_e were actually worse than statistics for the Level 3 design according to the report. In the Volume I report (*14*), which summarizes the findings of the calibration efforts, the S_e for the verification of the Level 3 nationally calibrated model is listed as 0.15 ft/mi (Table I-66 on page 186); however, the Volume II report (*27*), which details the model verification and calibration efforts, has a much larger value of 281 ft/mi listed for S_e (Table 17 on page 56). This is the only statistic that is reported differently between the two reports for the transverse cracking model. The recommendation was to recalibrate the model.

Calibration was conducted using the same 49 data points and indicated that predictions were excellent but slightly biased (R^2=0.91 and S_e=51.4 ft/mi).

It is not clearly stated which level (1, 2, or 3) was used in the local calibration effort. It was noted that default HMA creep compliance and HMA tensile strength values were replaced with MoDOT specific values and the local calibration coefficient, β_t, was modified from 1.5 to 0.625 to reduce bias. The use of MoDOT specific values would imply that a Level 1 design was used in the calibration effort. It should also be noted that β_t is the local calibration coefficient and has a default value of 1.0 in the software; therefore, it is assumed this was meant to describe the K-value. However, various values have been reported for K at a Level 1 design (as shown in A.3): in the Pavement ME software (Version 2.1), K has a default value of 1.5; in the Manual of

Practice the default value for K is reported as 5.0. In describing the nationally calibrated default model, the K-values were listed in the referenced document on page 7 (*27*) as follows:

- At Level 1, K = 5.0
- At Level 2, K = 1.5
- At Level 3, K = 3.0.

However, the statement on page 57, which implied a Level 1 design was utilized in local calibration while also stating the coefficient was changed from 1.5 to 0.625, indicates that the above values may have been reported incorrectly on page 7 (*27*). Therefore, Table 9 of this report, for both verification and calibration, reflects the default K-values as they are presented in A.3 for the Pavement ME software.

C.3.5 Northeastern States (15)

Transverse cracking predicted by the nationally calibrated model was compared with measured transverse cracking from 17 LTPP sections in the NE region (*17*). It was found that in many cases, the measured transverse cracking did not increase over time as the model predicts. It was believed that measurements were made in error; as a result, calibration was not performed on the transverse cracking model.

C.3.6 Ohio (18)

A non-statistical comparison of measured and predicted transverse cracking was used to evaluate the nationally calibrated thermal (transverse) cracking model. The measured transverse cracking from the LTPP projects (SPS-1 and SPS-9) was divided into four groups based on the extent of cracking, as listed below. The nationally calibrated model was used to predict transverse cracking to determine how frequently the predicted values fell in the same category as the measured values.

a. 0-250 ft/mile
b. 250-500 ft/mile
c. 500-1000 ft/mile
d. 1000-2000 ft/mile

It was found that all predicted transverse cracking fell into the same range as the measured cracking (0-250 ft/mile). Despite the adequate performance of the global transverse cracking model, it was recommended that due to the limited scale of transverse cracking measurements (maximum measurement of 110ft/mile), a more detailed review should be completed on the adequacy of the default HMA creep compliance and tensile strength within the MEPDG.

C.3.7 Oregon (19)

Verification of the nationally calibrated model was attempted for a Level 3 design with value of K equal to 1.5. The nationally calibrated model resulted in a S_e of 121. No other statistics were reported. The authors reported that the Darwin M-E considerably under-predicted transverse cracking relative to the actual measured cracking in the field.

In an effort to calibrate the model, iterative runs to optimize the thermal cracking model were conducted by changing K from 1.5 to 12.5 for 15 projects. Reasonable estimates of thermal cracking were found with a value of 10, but the locally calibrated model did not improve the prediction compared to the nationally calibrated model. The S_e before calibration was 121, however, the S_e increased to 751 after calibration. Based on the results of the local calibration effort, researchers recommended that additional projects with more variation in cracking be included in future calibration efforts.

C.3.8 Utah (21)

Similar to the procedure followed for alligator cracking, a non-statistical approach for validating the nationally calibrated model was conducted using Level 3 inputs. The predictions showed that the national model predicted cracking well for the fairly newer pavements constructed using Superpave binders. For older pavement sections that used conventional binders, the predictions were very poor and under-predicted measured transverse cracking. The authors indicated that local calibration of the model was not required.

C.3.9 Washington (23)

Verification of the transverse cracking model was conducted by plotting WSPMS data and the MEPDG predictions using default calibration coefficients over time. No statistics were presented, but the authors indicated that the MEPDG transverse cracking model using default calibration factors can reasonably estimate WSDOT transverse cracking.

C.3.10 Wisconsin

Verification of the transverse cracking model was conducted for the LTPP sections over time. A design was conducted for each of the 94 data points, and a non-statistical comparison of measured and predicted transverse cracking was performed. The results indicated that the nationally calibrated transverse cracking model using default calibration factors over-predicted transverse cracking in Wisconsin. Thus, the transverse cracking model was recalibrated by varying the calibration parameter (*24*). For the locally calibrated model, the K-values were reported on page 92 of the draft user manual (*25*) as follows:

- At Level 1, K = 3.0
- At Level 2, K = 0.5
- At Level 3, K = 3.0

C.4 IRI

C.4.1 Arizona (10)

Verification of the IRI model was conducted using a total of 675 data points. The goodness-of-fit reported was poor with an R^2 of 30% and S_e reported as 18.7 in/mi. Bias in the form of large over-predictions for lower IRI and under-predictions for higher IRI was observed. As a result, local calibration was deemed necessary.

It should be noted that the default calibration coefficients, C_1 and C_4 for the model, were listed differently than reported in the Manual of Practice. In the Appendix (page 188) of the

referenced report, C_4 is reported as 40 and is listed as the model coefficient for the rutting term, and C_1 is listed as 0.015 as the model coefficient for the site factor term. This is contrary to the coefficients reported in the Manual of Practice (as shown in A.6), where C_1 is listed as the model coefficient for rut depth with a value of 40, and C_4 is listed as the model coefficient for the site factor with a value of 0.015. However, on page 77 in Table 28 of the reference report, the terms are consistent with the Manual of Practice: C_4 is reported as the model coefficient for site factor and C_1 as the model coefficient for rutting. In order to be consistent, Table 10 lists C_1 as 40 and as the model coefficient for the rutting term and C_4 is shown as the value 0.015 and as the model coefficient for the site factor (for the global model). The model coefficient reported in Table 10 after calibration represent the same terms: C_1 for the rutting term and C_4 for the site factor term.

Local calibration was conducted using 559 data points. The goodness-of-fit reported was very good with an R^2 of 82.2% and S_e of 8.7 in/mi. It should be noted that the authors of the referenced study (*10*) reported different statistics in Chapters 3 and 8. In Chapter 3 of the document (Page 77, Table 28), R^2 was reported as 82.2% and S_e was shown as 8.7 in/mi. However, in Chapter 8 (Page 175, Table 65), R^2 was reported as 80% and S_e was listed as 8 in/mi. Regardless of the discrepancies in statistics, both sets show the local calibration resulted in very good predictions of IRI. Additionally, the statistics reported on page 77 in Table 28 and the plot of measured versus predicted IRI revealed the over-prediction bias in the global model was removed through local calibration.

C.4.2 Colorado (11)

Colorado conducted verification of the IRI model using 343 data points. It was reported that the nationally calibrated model had poor goodness of fit and bias in the predictions. Although it was reported that "the model over-predicts IRI for higher measured IRI values," the plot of measured versus predicted IRI shown in the referenced document (Figure 98, page 139) indicates under-prediction for higher magnitudes of measured IRI values. Calibration was recommended and was conducted using the same number of data points.

It was found that the goodness of fit was significantly improved. Discrepancies were also found in this study. On page 140, Equation 12, of the reference report (*11*) C_4 is reported as the model coefficient for the rutting term and C_1 as the model coefficient for the site factor term. On page 140, in Table 67 of the report, C_4 is referred to as the model coefficient for site factor and C_1 as the model coefficient for rutting. In order to be consistent with the Manual of Practice, the MEDPG default coefficients are listed in Table 10 for C_1 as 40 for the model coefficient for the rutting term, and C_4 as 0.015 for the model coefficient for the site factor (for the global model). Additionally, the model coefficient reported in Table 10 after calibration represent the same terms: C_1 for the rutting term and C_4 for the site factor term.

C.4.3 Iowa (13)

Measured IRI values were plotted against IRI values predicted by the nationally calibrated IRI model using the distress inputs predicted by the corresponding nationally calibrated distress models. This approach resulted in good estimation of field measurements. Alternatively, IRI

values were also predicted using nationally calibrated coefficients in the IRI model with inputs (rut depth, fatigue cracking, and thermal cracking) predicted by locally calibrated distress models. These values were plotted with the measured IRI values. This approach also resulted in good estimation of measured IRI values. Researchers did not consider the modification of the nationally calibrated coefficients due to the good estimation of measured IRI values by using either inputs from nationally calibrated distress models or inputs from locally calibrated distress models. Researchers also cite the expected improvement in longitudinal cracking and thermal cracking through other national studies as a reason for not calibrating the IRI model.

C.4.4 Missouri (14)

A total of 125 data points were used to verify the adequacy of the nationally calibrated IRI model. As shown in Appendix A.6, the IRI model utilizes predicted transverse cracking (TC), fatigue cracking (FC), and rut depth (RD) to compute IRI. In verifying the nationally calibrated model for MoDOT, the results of locally calibrated rutting and transverse cracking models were utilized to predict IRI. The authors reported that a reasonable prediction was found (R^2 = 0.54, S_e = 13.2 in/mi), but a slight bias was present. For higher magnitudes of IRI, slight under-estimates, although "not very significant," were reported for the nationally calibrated IRI model (*14*).

Local calibration was conducted to remove the bias and improve the accuracy of the predictions. Two sets of statistics were reported in the reference document (*14*) regarding the results of the local calibration, with no explanation for the differences. In Figure I-127 the coefficient of determination, R^2, was reported as 58%, the S_e was shown as 12.8 in/mi, and the number of datapoints utilized in the local calibration was shown as 121. However, in Table I-71 the following statistics were reported: R^2 of 53%, S_e of 13.2 in/mi, and number of datapoints as 125. The values shown in Table 10 of this report reflect the values shown in Table I-71 of the reference document (*14*), as this is the most complete set of statistics reported for the locally calibrated model. Although there are differences in statistics reported, they are small, and both show a reasonable correlation between measured and predicted IRI with the locally calibrated IRI model. Some bias in the predicted IRI was found with the locally calibrated model, however, the amount of bias was considered reasonable. The authors concluded that the model can be used in routine designs.

C.4.5 Northeastern States (15)

Measured IRI values from 15 LTPP sections were compared with predicted IRI values from the nationally calibrated model. A very poor correlation was reported, particularly for high measured IRI values. Additionally, the nationally calibrated model was reported to have an SSE of 1.557. Regional calibration was completed and resulted in improvements over predictions made by the nationally calibrated model with a reported SSE of 0.799.

C.4.6 Ohio (18)

A statistical comparison between measured IRI and IRI predicted by the nationally calibrated IRI model was conducted. The coefficient of determination was determined to be 0.008, indicating a poor correlation between measured and predicted IRI. The default IRI model was found to

over-predict IRI for lower magnitudes (less than 80 inches/mile) and under-predict at higher measured IRI values (greater than 80 inches/mile). Additionally, significant bias was reported in the predicted IRI. The S_e was found to be lower than that reported for the development of the nationally calibrated model (18.9 in/mile). Although this is reported in the text as both 19.8 in/mile and 9.8 in/mile, it is believed to be the lower of these two values, since it was reported to be lower than the nationally calibrated model.

Local calibration was conducted and a statistical comparison with measured IRI was completed to evaluate the calibrated model. It was found the S_e of the locally calibrated model was very similar to the nationally calibrated model. Hypothesis testing was also conducted, which showed that the null hypothesis for the intercept and slope were rejected while the null hypothesis for the paired t-test was accepted. The level of bias was reduced with the locally calibrated model and was considered more reasonable than the nationally calibrated model

C.4.7 Tennessee (20)

A verification exercise was conducted using 19 pavement sections in Tennessee; however, no statistics were reported on the accuracy or bias of the predicted roughness. As PSI is used to characterize roughness in Tennessee; IRI predicted by the nationally calibrated model was converted to PSI using a previously established PSI-IRI relationship. Two input levels were considered in the verification, “Level 1.5” and “Level 2.5”, which resulted in similar predictions of IRI. It was suggested that since IRI is dependent on rutting, fatigue cracking, thermal cracking, site, and other factors, the similarities in predictions were both a result of no transverse or longitudinal cracking predicted for either “Level 1.5” or “Level 2.5” and the small influence of AC layer properties on alligator cracking.

The effect of traffic level on PSI prediction was also evaluated and compared with measured PSI. It was found that for cumulative ESALs over a 20-year design period between 0 and 4.5 million, pavement roughness was under-predicted. However, the rate of decrease in PSI was similar to the rate for measured PSI. For cumulative ESALs between 4.5 and 9 million over the same design period, predicted and measured PSI agreed well, although measured PSI was reported to have high variability. Although local calibration was recommended, it was not conducted.

C.4.8 Utah (21)

The nationally calibrated IRI model was evaluated. The authors indicated that for the rutting input, the locally calibrated model was used. It is assumed that the remaining inputs resulted from nationally calibrated distress models. A total of 162 data points were used to verify the adequacy of the model. A good correlation between measured and MEPDG-predicted IRI was found, and the standard error of estimate (S_e) was about the same as that reported for the national MEPDG IRI model. Some bias in the predicted IRI was found, but it was considered insignificant. Local calibration was not required.

C.4.9 Washington (23)

It was found that when calibrated cracking and rutting estimates were used with the default IRI model, the results under-predicted actual WSDOT roughness, although the differences were small. This was believed to be due to the effect of studded tire wear in Washington, which is not modeled in the MEPDG. It was noted that the differences could be resolved through calibration of the model and that studded tire wear could be adjusted through the site factor in the model. However, the authors indicated that the IRI model could not be calibrated because of bugs in the MEPDG software.

C.4.10 Wisconsin

Verification of the IRI model was conducted for the LTPP sections. As for the verification of the distress models, a design was conducted for each of the 142 data points, and the predicted IRI results were compared with the field measured data. The results indicated that while the R^2 (0.6273) and SEE (5.694 in/mi) were reasonable, the nationally calibrated IRI model generally over-predicted IRI when it was less than 70 in/mi and under-predicted IRI when it was greater than 70 in/mi. In addition, a statistical test showed that the difference between the predicted and measured IRI values was statistically significant (*24*). Thus, the IRI model was recalibrated, and the locally calibrated coefficients are as follows (*25*):

IRI C1 = 8.6733
IRI C2 = 0.4367
IRI C3 = 0.00256
IRI C4 = 0.0134

C.5 Top-Down (Longitudinal) Cracking

C.5.1 Iowa (13)

The global model was evaluated by plotting measured versus predicted longitudinal cracking and determining the bias and standard error of both models. It was reported that the global model severely under-predicted the extent of longitudinal cracking in both the calibration and validation datasets. Although the locally calibrated model resulted in improved predictions and bias relative to the global model, bias was still present. A large standard error was also reported in predictions for both the calibration and validation sets of 2,767 and 2,958, respectively. It was recommended that predictions of longitudinal cracking in the MEPDG be used only for experimental or informational purposes until the ongoing refinement of the model is complete and it is fully implemented.

C.5.2 Northeastern States (15)

Calibration was performed by minimizing the SSE between predicted and measured longitudinal cracking using Microsoft Excel Solver. In comparing measured longitudinal cracking with predicted cracking, it was found that the nationally calibrated model severely under-predicts the extent of longitudinal cracking. Additionally, the SSE for was found to be 58.18 for the Northeastern LTPP sections used in the study. By performing the regional calibration, the SSE was reduced to 25.67 and an improvement was realized in longitudinal cracking predictions relative to measured longitudinal cracking.

C.5.3 Oregon (19)

Verification of the Darwin ME using global calibration factors showed that the software either under-estimates or over-estimates the distress considerably. Local calibration was recommended. After calibration, the model was improved, but there was also a high degree of variability between the predicted and measured distresses. The standard error of the estimate (S_e) was 3601 before calibration and 2569 after calibration.

C.5.4 Washington (23)

Verification of the longitudinal cracking model was conducted by plotting MEPDG predictions using default calibration factors and Washington State Pavement Management System (WSPMS) data over time. The default model tended to under-predict WSPMS data. Local calibration coefficients were used to predict longitudinal cracking and were also plotted with WSPMS data over time. No statistics were presented, but plots of predicted longitudinal cracking over time showed that the model estimations by the calibrated and default calibration factors were significantly different. The authors indicated that the predictions using calibrated factors show a similar level and progression of distress as the WSPMS longitudinal cracking data. They concluded that the calibrated model is able to reasonably estimate longitudinal cracking for WSDOT.