

TG 50-02

GUIDELINES FOR REPORTING UNCERTAINTY IN MEASUREMENT AND VERIFICATION

Approved By:	Executive Accreditation: Mpho Phaloane
Author:	Inspection STC Working Group
Date of Approval:	2017-10-17
Date of Implementation:	2017-10-17

CONTENTS

1. Introduction	3
2. Purpose.....	3
3. Reducing Risk and Managing Uncertainty from a Regulator's Perspective.....	4
4. Uncertainty levels	4
5. Reporting Savings with Symmetrical Savings Distributions	5
6. Reporting Savings with Skewed Savings Distributions	6
7. Calculation of Savings Uncertainty	7
8. Statements on Uncertainty	8
9. Unquantifiable Uncertainties.....	9
10. Data Validation.....	10
11. Model Selection.....	13
12. Model Validation.....	14
13. Model Prediction Validation.....	17
14. Savings Validation.....	22
15. Acknowledgements	22
Appendix A: Combining Uncertainties	23
Appendix B: ASHRAE Guideline 14 Method for Combining Uncertainties	29
Appendix C: References to relevant sections in other guidelines	32
Appendix D: References and Further Resources	33
Appendix F: Statistical Software packages for M&V Uncertainty Quantification.....	34
Addendum 1: Amendment Record	35

1. Introduction

This document clarifies the uncertainty requirements contained in South African National Standard (SANS) 50 010 Clause 7 and related International Standards Organisation (ISO) standards. It is not intended to place too onerous a burden on M&V reporting or to exclude projects, but to codify best practice as well as provide a fair requirement for the uncertainty with which savings are reported. **The uncertainty figure given by an energy model is only credible if the assumptions for the model have been verified and these results are shown.**

*"The evaluation of uncertainty is neither a routine task nor a purely mathematical one; it depends on detailed knowledge of the nature of the measurand and of the measurement method and procedure used. The quality and utility of the uncertainty quoted for the result of a measurement therefore ultimately depends on the understanding, critical analysis, and integrity of those who contribute to the assignment of its value."*¹

Credible savings should also be defined in alignment to the applicable standards and stakeholder expectations.

2. Purpose

It is the purpose of this guidance document to provide clarifications and guidance on the use of statistics and statistical modelling in the M&V industry. Although the latest International Performance Measurement and Verification Protocol (IPMVP), American Society of Heating, Refrigeration, and Air conditioning Engineers (ASHRAE) Guideline 14 and SANS 50 010 do mention uncertainty and requires the management thereof, it is only the IPMVP and ASHRAE guidelines that give some guidance and meaning to the required "management of uncertainty". This guideline therefore aims to provide better understanding in the use of statistics and other uncertainty tools in the field of M&V as it is applied in South Africa. Targets for acceptable uncertainty is not set in this guideline (it is not the purpose), however recommendations can be made to assist relevant stakeholders in defining the relevant acceptable levels of uncertainty and what may be required should these targets not be met.

The contributions of the IPMVP, ASHRAE Guideline 14, and Uniform Methods Project (UMP) are hereby acknowledged. Much of this guideline is based on these works.

Part I: Savings Uncertainty Reporting

This section does not make recommendations, but reports results from international best practice and a South African survey on savings uncertainty levels, as well as discussions from a SANAS STC WG meeting on uncertainty quantification, convened for the writing of this guideline. Internationally, savings uncertainty reporting are dealt with either by setting minimum uncertainty limits, or by applying "knock-down" or discount factors, recognise only a portion (e.g. 50% or 25%) of reported savings. The reporting scheme proposed by the WG was such that no project would be excluded, and that the "knock-down" or discount factor applied to reported savings not complying with a given requirement would be fair, and not a one-size-fits-all solution.

¹ Guide, EURACHEM-CITAC. "Quantifying uncertainty in analytical measurement." *Laboratory of the Government Chemist, London* (2000).

3. Reducing Risk and Managing Uncertainty from a Regulator's Perspective

The following is based on the *State and Local Energy Efficiency Action Network's Energy Efficiency Program Impact Evaluation Guide*².

Uncertainty is quantified to manage risk – in the case of M&V, the risk of not achieving savings. The level of effort needs to be balanced against the benefit derived from the uncertainty quantification. Managing risk is therefore a trade-off which inadvertently also affects the associated costs. Resources should not be spent on ineffective actions (energy efficiency programs that don't work, or M&V savings reports with uncertainties that are so large as to be meaningless, or expensive, or even over-specified M&V requirements, which then may then exclude many projects from being evaluated for a certain reward programme, for instance). An important question to ask is therefore, "risky compared to what?"³ or in other words; how much uncertainty is enough and what level of uncertainty is acceptable? The other question to ask then is "what do we do with uncertainty, more so, what action should be taken when uncertainty is not within acceptable levels, but cannot be improved?

The SEE Action guide therefore suggests a number of questions that a regulator should ask when setting uncertainty reporting requirements to manage risk, reproduced below in abbreviated form:

- What are the project budget and savings goals? (Larger budgets tend to have more evaluation).
- Does the savings determination need to indicate how much savings were achieved, or only that the savings were above a specified goal?
- Is it a new programme, or one with a track record, which would require less rigorous evaluation?
- Is it adequate to verify that the measures were installed, or is rigorous field evaluation required? This method falls more to a "deemed savings" approach based on historical detailed M&V facts.
- Is the programme likely to be expanded (requiring more monitoring) or contracted?
- How long, in months or years, does monitoring need to be conducted? Longer periods are more expensive.
- What is the interval on savings reporting? Monthly or annual savings are less expensive than hourly figures.
- Is there a performance-based incentive determined on the basis of the evaluation?
- Are there cost-recovery or lost revenue recovery impacts for the project sponsor?

4. Uncertainty levels

M&V uncertainty is expressed as an "expanded uncertainty", following international best practice as codified in the ISO Guide to the Expression of Uncertainty in Measurement (GUM). An expanded uncertainty is a relative precision, stated at a given confidence interval. Therefore if the 80% confidence limit falls at 110, and the mean is 100, the precision is 10% at the 80% level, or 80/10.

² State and Local Energy Efficiency Action Network, *Energy Efficiency Program Impact Evaluation Guide*, US Department of Energy, December 2012, DOE/EE-0829

³ For some projects, quantifying uncertainty to certain levels do not increase programme effectiveness. See Sonnenblick, R., and J. Eto. "A Framework for Improving the Cost-Effectiveness of DSM Program Evaluations. A Report from the Database on Energy Efficiency Programs (DEEP) Project." *Lawrence Berkeley Laboratory. LBL-37158. March 1995.*

International uncertainty reporting requirements vary from 68/50 (ASHRAE Guideline 14, with additional uncertainty requirements that also need to be met) to 90/10 for states such as California, where M&V funding is very high. M&V costs are very sensitive to the precision level, and not as sensitive to the confidence level. A popular requirement is 80/20.

Note that due to limitations in the data, or the presence of unquantifiable uncertainties (see Section 4.2), a large reported savings uncertainty figure is not always an indicator of poor M&V. Similarly, a small reported savings uncertainty figure is not always an indication of high quality M&V. The validity of the reported figure will depend on the model validation that was performed. Validation is discussed in Part II of this document.

A self-reported survey of uncertainty levels of projects done by South African M&V companies found the following:

Project type	Percentage of previous projects which would have complied at the given reporting level					
	80/10	80/15	80/20	80/30	80/40	80/50
Industrial	30.43%	43.48%	65.22%	82.61%	86.96%	91.30%
Non-Industrial	70.73%	100%	100%	100%	100%	100%
Total	40.63%	56.25%	71.88%	84.38%	87.50%	90.63%

The sample sizes were small, however: 23 industrial projects, most of which from the same company (with rigorous uncertainty quantification) and 11 non-industrial projects from various companies, some reporting 0-10% uncertainty (unlikely).

Note to regulators: Not all M&V studies are equal, and a one-size-fits-all approach isn't helpful. One suggestion from a South African M&V company has been 80/7.5 for up to 12 data points, 80/15 for up to 52 data points, and 80/20 for more than 52 data points. Differentiating between industrial and non-industrial projects could also be useful. Note that leading guidelines such as ASHRAE's Guideline 14 place more emphasis on CV(RMSE) and bias (see Appendix). The ASHRAE requirements are stringent, however.

5. Reporting Savings with Symmetrical Savings Distributions

The scenario described below is that considered by the SANAS STC WG to be one way of reporting savings in a fair manner.

Symmetrical savings distributions are the default case and holds for most projects where normal distributions are assumed. Savings are then reported as the most likely savings value at the x/y confidence/precision level on the savings. For cases where savings are normally or symmetrically distributed, the most likely value is the mean.

If savings cannot be reported with $y\%$ relative precision at the $x\%$ confidence level, the savings reported can be adjusted by the difference between the savings precision at the $x\%$ confidence level, and the $y\%$ requirement.

The following examples assume the 80/20 level.

Note that these calculations pertain to *savings* uncertainty, not overall uncertainty. It may be that the energy use on a project is 1000MWh, and the savings are 200MWh. If the uncertainty is 10MWh, this is 1% of the total energy use, but 5% of the savings. A model which looked accurate on the baseline now seems inaccurate when considering savings. This is to be expected. It becomes more acute when the savings are a small percentage of the total energy use: for smaller differences made by the Energy Savings measure, more accurate instruments are needed to detect this difference with confidence.

5.1. Example 1: Symmetrical savings distribution, uncertainty higher than threshold

Suppose a project saves 200MWh with a 48MWh precision at the 80% confidence level. The relative precision on the savings is then only 24%, not the required 20%. Therefore, the creditable savings are reported as:

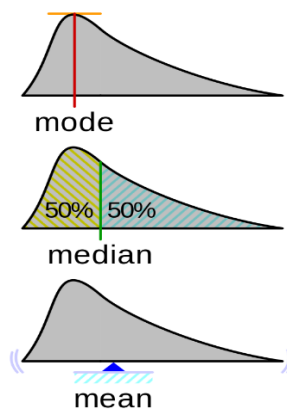
$200 - (0.24 - 0.2) \times 200 = 192\text{MWh}$. Example 2: Symmetrical savings distribution, uncertainty within threshold

Suppose a project saves 200MWh with a 10MWh precision at the 80% confidence level. The relative precision is then 5%, which is better than the required 20%. Therefore the creditable savings are reported as the mean: 200MWh.

6. Reporting Savings with Skewed Savings Distributions

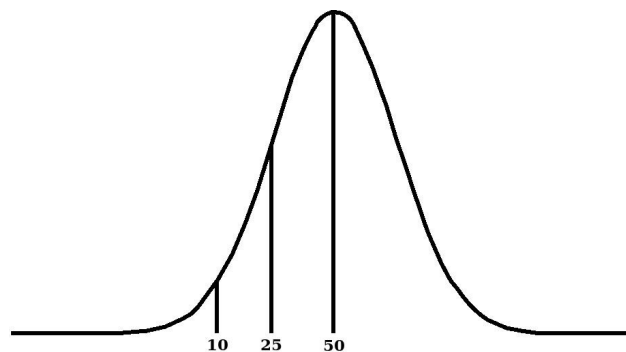
Skewed savings probability distributions are not very common in M&V because normal or Student-T distributions are usually assumed, and are produced by linear regression and other standard techniques. Skewed distributions are therefore not of concern in most projects. However, multiplying or dividing values that are normally distributed, doing yes/no surveys, modelling the times between events such as equipment failure, or yield from renewable sources such as wind, can lead to skewed distributions⁴. For skewed savings distributions, the mean is not the most likely savings value, and the standard equal-tailed confidence interval is not the best representation of uncertainty. In such cases the mode of the distribution can be used as the most likely savings value. The difference between the mode, mean, and median are illustrated below.

⁴ The skewness of a data set can be calculated by most spreadsheet and statistical software packages. Rules of thumb suggest that skewness in the range (-1, 1) can be considered close to symmetrical and therefore normally distributed. However, with small data sets, noise may increase the apparent skewness. A more reliable test is to divide the skewness by the standard error on the skewness estimate. Values in the range (-1.96, 1.96) yield a 95% probability of unskewed data. Note that non-skewness by itself is not a guarantee of normality. For more on this, see Section 4.2



For skewed savings distributions, the lower confidence limit is replaced by the $\frac{100-\alpha}{2}$ percentile, and the mean is replaced by the mode.

Percentiles work as follows: 10% of the savings distribution will be below the 10th percentile. 25% of the savings distribution will be below the 25th percentile (also known as the quartile, meaning quarter). 50% of the savings distribution will be below the 50th percentile: this is referred to as the median. Reporting at the 10th percentile, it means that there is only a 10% chance that the true value is below this number. The figure below illustrates the 10th, 25th, and 50th percentiles of a normal distribution:



6.1. Example 3: Skewed savings distribution, uncertainty higher than threshold

If the mode of the savings is 200MWh, and the 10th percentile of the savings distribution is at 140MWh, the precision is 30%. If the reporting precision requirement is 20%, it means that the precision is 10% higher than the allowable percentage for reporting the mode of the savings. Therefore the creditable savings are reported as

$$200 - (0.3 - 0.2) \times 200 = 180 \text{MWh.}$$

6.2. Example 4: Skewed savings distribution, uncertainty within threshold

If the mode of the savings 200MWh, and the 10th percentile of the savings distribution are at 190MWh, the precision is 5%. Therefore the creditable savings are reported as 200MWh.

7. Calculation of Savings Uncertainty

M&V uncertainty quantification includes a number of considerations. **As a minimum, savings uncertainty usually considers measurement, sampling, and modelling uncertainty.** The uncertainty values for each of these components are then stated. Sampling uncertainty is not present

in every study, but modelling and measurement usually are. If they are not applicable, a motivation supported by auditable proof is usually given. A method for combining uncertainties is given in Part II.

8. Statements on Uncertainty

When reporting the result of a measurement of a physical quantity, it is obligatory that some quantitative indication of the quality of the result be given so that those who use it can assess its reliability. Without such an indication, measurement results cannot be compared, either among themselves or with reference values given in a specification or standard. It is therefore necessary that there be a readily implemented, easily understood, and generally accepted procedure for characterizing the quality of a result of a measurement, that is, for evaluating and expressing its uncertainty⁵

Defining the value of a variable without an expanded uncertainty (confidence and precision) can be misleading. High quality M&V reports will therefore contain statements about the confidence and precision of variables of interest. Accompanying probability distributions on these variables provide even more complete information.

When measurement instruments are used, the make and model of the measurement instrument, together with the manufacturer's accuracy specification and a calibration certificate provide further assurance and motivation of results.

Variables such as the population size, sample size and assumed inter-sample coefficient of variance (CV) are expected in M&V documents where sampling was done, by most guidelines.

Where stipulations are made, their applicability and credibility depends on auditable proof. A timer may indicate operational hours and as such this factor is not measured. However that timer's operation, its control capability and its time-keeping capability are uncertain if not verified.

⁵ Bureau International des Poids et Mesures, Commission électrotechnique internationale, and Organisation internationale de normalisation. *Guide to the expression of uncertainty in measurement (GUM)*. International Organization for Standardization, 1995.

Part II: Validation

Basic checks for linear regression

For $n > 15$

- Collinearity (VIF or condition number)
- Autocorrelation (Durbin-Watson)
- Overall Significance (F -test)
- Explain outliers
- Bias (NDB)
- Goodness of Fit (CV(RMSE))

For $n < 15$, also add:

- Normality of residuals: (Anderson-Darling)
- Residual four-plot
- Leave-one-out cross-validation

Reporting savings with a given uncertainty level does not guarantee that the quoted level is valid. In order to verify the validity of the answer, an M&V practitioner needs to demonstrate that the data and model were obtained with due care. The requirements below describe certain important checks, but are not exhaustive.

These tests are built into most spreadsheet or statistical software packages, with some performed automatically. They consist of calculations and graphical plots that highlight certain features of the data or behaviour of the model. The Analysis of Variance (ANOVA) table is a widely recognised method of reporting the most important information about a model.

9. Unquantifiable Uncertainties

Some Energy Governing Factors (variables) are expensive or impossible to quantify, but do have an effect on the overall savings uncertainty.

A prudent allocation of M&V resources should lower the overall uncertainty while maintaining an improved cost benefit by addressing the uncertainty factors which have the largest influence and can also be mitigated at the (most often) lowest cost. Managing this risk is the art and science of M&V uncertainty quantification.

It is considered bad practice for an analyst to claim that a given uncertainty factor is practically unquantifiable in order to remove its effect from the overall uncertainty calculation. It is more acceptable to list all significant but unquantified factors, and to give reasons for not attempting to quantify them.

Factors vary by project, but could include the following:

- Differences in weather between meteorological station and facility;
- Human errors for data entry, survey response, operations etc.;
- Poor instrument selection, placement, and installation;
- Misestimation of key energy governing factors for any chosen Measurement option;

- Missing data;
- Uncertainty arising from incorrect model form;
- Model extrapolation uncertainty, if the baseline period values do not cover the whole input range;
- Human behaviour in certain kinds of projects, e.g. the size of the non-participant population for spill-over or free ridership calculation in so-called net-to-gross evaluations;
- Self-selection (rather than random selection) and other survey biases;
- Assumptions about the distribution of the data: assuming normality when data is non-normally distributed, for example; and
- Multicollinearity of energy governing factors.

10. Data Validation

10.1. Statement of measurement errors on all variables

The precision of measurement instruments are usually stated at the 95% confidence level, and can be assumed to be so if no further information is provided by the supplier. This is sometimes referred to as “ $k=2$ ” or “a coverage factor of two”. Where figures from the ASHRAE Guideline 14 are used or where instruments are qualified to an international standard, the confidence on the stated precision values is to be 68%.

Often, precision is relative to the full-scale meter value. A Class 1 meter is accurate to 1% of its full-scale reading value, for example. The relative precision of small readings may therefore be much worse than 1%. Therefore first fully understand a manufacturer’s claim of meter error, before applying it to energy impact evaluations.

Typical measurement uncertainty values for different kinds of instruments can be found in the ASHRAE Guideline 14⁶.

When using a utility meter for data, by law that meter should be calibrated. In fact, most of the time it is quite difficult getting that meter’s credentials and error specifications from the utility supplier. Efforts to get such information should not dwindle as the result of the whole M&V process is reliant on the metering data.

For more on good measurement practice in M&V, see The Uniform Methods Project chapter 9: Metering Cross-Cutting protocols⁹

10.2. Mismeasurement

Measurement Uncertainty is often regarded as negligible in M&V reporting, or simply added to the overall uncertainty. This is valid if the error is in the dependent variable (usually energy use). However, this is incorrect if the mismeasurement is in the independent variable (the energy governing factor) such as temperature, occupancy, production volume, etc. – and can be detrimental to the reported accuracy. Normally distributed, unbiased errors in the measurement of independent variables can lead to biased estimates and unreasonably narrow confidence intervals on these parameter estimates. This is called attenuation bias or the errors-in-variables effect.

¹ Additional information can be found in:

https://www.researchgate.net/publication/309180245_Measurement_Uncertainty_in_Energy_Monitoring_Present_State_of_the_Art_preprint

⁴ <https://energy.gov/sites/prod/files/2013/11/f5/53827-9.pdf>

Mismeasurement becomes significant if the error in the measurement of independent variables is greater than 5%.

10.3. Regression Sample Size

The population size, sample size, assumed inter-sample coefficients of variance (CV) are needed to assess the validity of a model where sampling was done.

The sample size needs to be determined with the power of the final test in mind (in this case the regression tests).

Studies have shown that state-of-the-art M&V regression models need only 3-6 months' hourly data to characterise the baseline adequately, given that such data reflects all operating conditions of the facility.¹⁰ Note that this does not mean that *all* M&V models need only 3 months' hourly baseline data.

Three sample size methods are proposed - as below:

a) Using the relationship between the F -test and expected R^2 (adjusted): For each possible model size, power of test between 80% and 90% and confidence level 90% - Minitab, provides a simplified rule for the sample size¹¹

PN= number of regression parameters (excluding the constant)

n(PN) = the suggested sample size given the number of regression parameters.

PN	n(PN)
1-3	40
4-6	45
7-8	50
9-11	55

Notes:

The resulting sample size from above table will result in a high power of the regression test (>80%).

Uncertainty calculations for sample sizes **less than 15** (see linear regression assumptions in section: Distributional Assumptions) are questionable.

b) Using Power Curves (For regression null hypothesis tests)

In M&V, sample sizes are usually calculated to minimise Type I errors, but not Type II errors. A Type I error occurs when there is a false positive: Savings are detected where there is none. In other words, the 90/10 criterion ensures that if savings are detected we can be 90% sure that they actually occurred. There is a low probability of a false positive.

Type II errors happen when there is a false negative: no savings are detected, when there really were savings. In other words, there is a failure to detect an effect or difference, when that effect or difference really is present.

¹⁰ <https://eta.lbl.gov/sites/all/files/publications/lbnl-187225.pdf>

⁶ [http://support.minitab.com/en-us/minitab/17/Assistant/Multiple Regression.pdf](http://support.minitab.com/en-us/minitab/17/Assistant/Multiple%20Regression.pdf)

Just because we control for false positives (Type I), does not mean we control for false negatives (Type II). For Type I error-optimised sampling plans, the sample size is the minimum sample size needed to avoid false positives. However, the sample size needed to avoid false negatives at the same rate is usually much higher, and so the probability of type II errors is still high. It depends on the situation but it could be 50%, meaning that if no savings are detected there is a 50/50 chance that there actually were savings. This is referred to as low statistical power, and is a risk for the Project Developer (PD) in the case of an Energy Performance Contracting, or the owner in the case of a rebate project.

Statistical package tools allow for determining the sample size based on the effect size index, error probability, power, and number of regression model predictors. These are the preferred tools for sample size determination. The only practical problem is that many of these are geared towards tests related to comparing means rather than the regression null hypothesis testing.

Notes:

An excellent tool for this is *G*Power*¹² – when used, closely aligns with the Minitab table above.

Statistical power is not simple to codify into a rule. However, the M&V practitioner's results for sample size determination can be checked by using a standard toolset if required. Common practice is for the power of test value to be targeted at 0.8 or 80%

c) For general sampling: IPMVP and ASHRAE

The IPMVP and ASHRAE recommendation for sample size determination is (from equation B-2 ASHRAE):

$$n = \frac{CV^2 z^2}{e^2}$$

Where n is the sample size, CV is the coefficient of variation, and e is the relative precision.

Notes:

Experience has shown that this calculation does not provide sufficient power and typically understates the sample size required for regression modelling. This method is useful for savings estimation but not regression, and the limitations regarding power should be considered. The above method is suitable when estimating the savings in the performance reporting period.

In general, one has to be extremely cautious when the sample size is less than 15. There are significant implications for regression that need to be tested if $n < 15$.

10.4. Outliers

It is standard practice to discuss and explain all outliers in the data. If an outlier represents normal operating conditions, statistical best practice dictates that it should not be removed, as this may manipulate a model for more favourable outcomes. The removal of outliers must always be discussed.

Robust regression techniques are sometimes used for modelling where outliers are present, so that outliers do not influence the model. This is equivalent to removing outliers. These

⁷ <http://www.gpower.hhu.de/en.html>

include Huber and Theil-Sen regression, and RANSAC. Such techniques cannot be used without discussing the outliers and the reasons for removing them.

For large multidimensional datasets, it may not be possible to detect outliers manually and discuss them one by one. In such cases, methods such as Principal Component Analysis (PCA) with outliers defined according to criteria such as the Mahalanobis distance at a given confidence level could be used. The influence of the data on the model could then be determined using tools such as Cook's distance, DFFITS, or leverage. Such techniques cannot be used without a detailed discussion of outliers and the reasons for removing them. Otherwise using these methods amount to deleting outliers without justification.

10.5. Independent Variables

Independent variables are energy governing factors that are expected to change, such as temperature, occupancy, production volume, etc. A plot of the magnitudes of the independent variables for the baseline and reporting periods separately gives a visual comparison of the state of the system pre- and post-retrofit. For many cases the independent variables will be different for these two periods, as newer equipment may increase production, or one year may be cooler than the previous one, for example. This is not an automatic disqualification. The energy model is designed to adjust for these changes. However, in cases where pre- and post-retrofit independent variables are so different that significant extrapolation is needed, best practice would dictate that the validity of such an extrapolation be discussed.

11. Model Selection

Many different machine learning techniques can be used to model energy use. These include linear and other types of regression, Bayesian methods, neural networks, nearest-neighbour models, and others. It is beyond the scope of this document to discuss the validation tests for all methods. **Rather, linear regression will be discussed in detail as it is the simplest and most popular choice for M&V.** If a different method is used in an M&V report, tests and threshold values would be described and referenced from textbooks or other credible literature in order to demonstrate that due care was exercised during modelling and that the results are credible.

11.1. Distributional Assumptions for Linear Regression

Linear regression is the most popular modelling method for M&V. It is simple and powerful. However, it only holds under certain assumptions. These assumptions need to be stated and motivated when necessary:

1. The independent variable data have a linear relationship to energy use. This does not mean that only straight lines can be used, but that the $y=Ax+b$ equation is a polynomial, and doesn't contain any terms such as $\cos(x)$, x^2 , $\log(x)$, e^x , etc.
2. The residuals follow normal distributions, especially for sample sizes $n < 15$. The residuals are the model errors: the differences between the predicted model value and the measured data point value. Where the number of data points far exceed the number of energy governing factors considered in the model, it is less critical.
3. There is no autocorrelation. (Note that the ASHRAE method can take lag-1 autocorrelation into account).

4. There is little or no collinearity between independent variables (energy governing factors). (Note that this is extremely important to model coefficient stability and needs to be shown carefully for a model to be considered valid).
5. The data are homoscedastic. In other words, the variance in the data stays constant over the range of the data. If the variance increases as temperature increases, and temperature is an energy governing factor, this assumption is violated. This can be seen by a trumpet-like shape when plotting the residuals against the independent variable.

Points 2 – 5 above are most easily demonstrated in graphical format in the baseline reports. These are otherwise known as the “residual four-plots”.

The requirements in the model validation section are meant to provide assurance that the linear regression model assumptions are not violated and that the model is indeed valid.

12. Model Validation

12.1. General Linear Regression model validation

12.2. Testing Normality of Residuals

The M&V practitioner or analyst usually motivates why he/she has a stated belief that the data or residuals come from a population with a specific distribution – and how this belief impacts on the validity of the model.

Thousands of simulation tests have demonstrated that normality of the residuals is not a concern when $n > 15$. If the sample size exceeds this value, normality tests can be deemed as not required.

The following comment from Professor Robert Nau of Duke University¹³ applies to models with more than 15 data points: *“Technically, the normal distribution assumption is not necessary if you are willing to assume the model equation is correct and your only goal is to estimate its coefficients and generate predictions in such a way as to minimize mean squared error. The formulas for estimating coefficients require no more than that, and some references on regression analysis do not list normally distributed errors among the key assumptions. But generally we are interested in making inferences about the model and/or estimating the probability that a given forecast error will exceed some threshold in a particular direction, in which case distributional assumptions are important. Also, a significant violation of the normal distribution assumption is often a “red flag” indicating that there is some other problem with the model assumptions and/or that there are a few unusual data points that should be studied closely and/or that a better model is still waiting out there somewhere.”*

The implication of this comment is that for small sample sizes, ($n < 15$) the tests listed below are necessary. For larger sample sizes, the F-test is necessary to ensure the structural validity of the model, as part of validating the forecasting power of the model.

For normal distributions and small sample sizes, the following checks can be done to demonstrate normality.

⁸ <http://people.duke.edu/~rnau/testing.htm>. See also: <http://blog.minitab.com/blog/adventures-in-statistics-2/how-important-are-normal-residuals-in-regression-analysis>

12.2.1. *q-q* Plots

One can use the normal probability plots (q-q plots) as an informal means of assessing the non-normality of a set of data. A perfect normal distribution will have all points lie in a straight line. However, for most real-world cases there will be some deviation. You may need considerable practice before you can judge them with any degree of confidence.

12.2.2. Histograms of residuals

For sample sizes greater than 50. The result should approximate a normal distribution.

12.2.3. Anderson-Darling statistic

This test statistic is useful for small to medium sample sizes ($n=10$ to $n=100$), but may test positive falsely for non-normality in large samples. For samples of 10 to 15, this check can be done to provide some assurance of normality.¹⁴

The following are acceptable limits for the Anderson-Darling test. They assume that the mean and variance of the sample are unknown, and are at the 5% significance level. Note that for sample sizes under 10, there is too little data to judge normality.

<i>n</i>	Anderson-Darling limit
10	0.683
20	0.704
50	0.735
100	0.754
>100	0.787

12.3. **Notes:**

In addition to the standard Excel Analysis add-ins, there are Excel resource packs for doing these tests¹⁵. Autocorrelation (serial correlation)

In addition to the residual correlation plot, the Durbin-Watson *d* statistic can be applied¹⁶

⁹ Other tests such as Chi-squared, Shapiro-Wilk or Kolmogorov-Smirnov could also be done. Anderson-Darling is widely accepted as most reliable.

¹⁰ <http://www.real-statistics.com/free-download/real-statistics-resource-pack/>

¹¹ https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic

Acceptance values: $d=2$ indicates no autocorrelation. The value of d always lies between 0 and 4. If the Durbin–Watson statistic is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if Durbin–Watson is less than 1.0, there may be cause for alarm.

Note:

The statistic can be implemented in Excel using the SUMXMY2 and SUMSSQ functions¹⁷

Data and model validation apply to all methods, unless otherwise specified. However, additional demonstrations of model adequacy should hold for other methods. It is beyond the scope of this document to list all methods and their criteria. However, two of the most popular methods will be discussed. If a practitioner selects another method, the onus is on him or her to motivate and demonstrate adequate model convergence.

Making the source code or spreadsheets for calculations available for review. In some cases, sharing the source code might be a sensitive issue. However, the pseudo-algorithm (or flow chart) used for the modelling can then be used to have discussion with the interested parties.

12.4. Collinearity

Collinearity occurs when two of the independent variables in the regression analysis are highly correlated. This makes it difficult for the model to tell which variable is causing the effect. For example, one could regress the energy use of a factory against goods produced and raw material consumed. However, since goods produced would be highly correlated with raw material consumed, regressing against both of them simultaneously will just add noise (variance) to the system. The regression result may show that energy use is not dependent on one of them, or that energy use is negatively correlated with one of them, when one knows that it should be positively correlated (more goods = more energy used).

Indications of collinearity (also called multicollinearity):

1. A Variance Inflation Factor (VIF) > 5 (Used in Minitab).
2. Condition Number > 20 (Used in Python Statsmodels).
3. Large changes in regression coefficients when an independent variable is added or deleted.
4. Small (close to zero) regression coefficients for affected variables, while they also with small p-values (indicating high significance).
5. If a variable has a small coefficient (close to zero) in multiple regression, but a large coefficient when you regress against it alone.

In order to reduce collinearity, remove one of the collinear variables. "Model Selection" or "shrinkage" techniques such as Principal Component Analysis, LASSO (for combinations of a moderate number of important variables), or Ridge Regression (for combinations of many less important variables) are popular ways of identifying the most influential, non-collinear variables.

12.5. Note on assuming normality

The Central Limit Theorem (CLT) is often invoked in popular literature as a 'magic formula' for justifying the simplifying assumption of normal distributions. It does indeed show that under certain conditions, adding many different distributions (possibly of unknown shape) will

¹² <http://paws.wcu.edu/jarrell/mehodwstat.htm>

result in a normal distribution. However, the CLT only holds when the following three conditions are met¹⁸:

- a) A very large number of distributions are added together.
- b) The constituent variances are known to be of the same order of magnitude.
- c) The distributions are independent.

Since normal distributions are so common, making the assumption of normality is often valid, but invoking the CLT for marginal cases is not good practice. If the data is not normally distributed, it would be better to use regression methods such as Generalised Linear Models or Bayesian methods. Monte Carlo simulation or the Mellin Transform Moment Calculator described in Appendix A simplifies the M&V practitioner's task of adding various distributions together.

13. Model Prediction Validation

The usefulness of a model is its performance on future data or “unseen” data (because this is how it is going to be used in M&V).

A key message is that a regression model that is statistically significant is not necessarily a useful one, and something more is needed. This “something more” is difficult to quantify, but some “rules of thumb” can be applied. These are highlighted in the sections that follow.

13.1. *k*-Fold Cross Validation

Cross validation is a technique for assessing the structure of the model, that is, how well it describes the underlying process without “overfitting” to noise in the process.

It can be used for regression or other kinds of models. The method involves dividing the baseline dataset into complementary “training” and “validation” sets. For example, suppose that the baseline is monthly data for January through December of a given year. Exhaustive cross-validation could work by selecting 12 combinations of training/validation sets. Let *V* indicate the validation set, and the blank cells the training months.

¹⁸ For an informative discussion, see *R. Kacker and A. Jones, On the use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent, Metrologia (2003), 40:235*

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1												V
2											V	
3										V		
4									V			
5								V				
6							V					
7						V						
8					V							
9				V								
10			V									
11		V										
12	V											

This is also called “leave-one-out” cross validation. The validation set is predicted by the model trained only on the training set. By choosing different training/validation combinations and checking the results, one can test to see if a specific month has a significant impact on the training model. For example, suppose that July had been an abnormal month, and influences the model significantly. This will be identified by Set 6 having a large prediction error, but a small CV(RMSE) on the baseline data. For hourly data, leave-one-out cross validation will not be economical, and other k -fold cross validation schemes could be considered.

13.2. Useful Regression

A baseline regression model is considered a satisfactory predictor in the sense that the range of response values predicted by the regression model is substantial compared with the standard error of the response. This means that the observed F -ratio (from the ANOVA table) should exceed the critical point of the F -distribution by several points.

Draper and Smith¹⁹ propose two simple “rule of thumb” checks for determining if a regression model is a satisfactory predictor (based on Box and Wetz).

F -Test:

$$F_{obs} \geq 4 \times F_{crit}$$

The only complication is to obtain the F_{crit} value from the F -distribution for the degrees of freedom.

An ANOVA (Analysis of Variance) table is a useful statistical tool for comparing the means of different samples, for determining the significance of energy governing factors in the overall

¹⁹ N.R. Draper and H. Smith, *Applied Regression Analysis*, 3rd Ed. 1998. John Wiley & Sons.

regression model, and performing and reporting F -tests. MS Excel can generate ANOVA tables automatically, as can other statistical packages.²⁰

The probability value of the ANOVA table cannot be taken as statistically significant unless the model is tested for practical significance versus statistical significance.

An alternative form of the test is:

$$\left(\frac{\text{Max } \hat{Y}_i - \text{Min } \hat{Y}_i}{\sqrt{p \times \frac{s^2}{n}}} \right) \geq 4$$

Where:

Max and Min $\hat{Y}_i$: The maximum and minimum predicted values of the regression.

p : The number of regression parameters including the constant.

s : The standard error of the regression

n : The number of samples

The above can be seen to be the same as: the range of the predictions needs to be greater than four times the average spread error of the measured response.

Note:

The above is easily applied since all the necessary statistical performance data is available from the ANOVA table.

13.3. Net Determination Bias

Net determination bias is an important baseline test to check whether the algorithm for savings determination has not added error: overpredicting or underpredicting energy savings. This is defined in ASHRAE 14 in section 4.2.10 with an acceptance limit of NDB $\leq 0.005\%$. The Bonneville Power Administration's Regression Reference Guideline (see Appendix) contains a useful discussion on NDB vs. Normalised Mean Bias Error, as the treatment of these two in ASHRAE G14 is confusing.

The NDB is calculated as the percentage error between the model's predicted baseline energy use ($\hat{E}_i$), compared to the actual baseline energy use (E_i), over all n sampling points:

$$NDB = \frac{\sum_i^n E_i - \sum_i^n \hat{E}_i}{\sum_i^n E_i}$$

This is simply the percentage that the model is over- or underpredicting actual energy.

13.4. Model Goodness of Fit

Goodness of fit describes how closely the baseline model follows the baseline data. It is measured in different ways, with R^2 being popular (on which, see the section below). The

¹³ Many online sources provide tutorials on performing and interpreting ANOVAs. For example, https://en.wikipedia.org/wiki/Analysis_of_variance. Note that the type of ANOVA will depend on the nature of the analysis.

most reliable measure for energy data²¹ is the Coefficient of Variation on the Root Mean Square Error: CV(RMSE). The CV(RMSE) of a model with respect to a data set is equivalent to the standard deviation of a point estimate. CV(RMSE) is measured with respect to the baseline data in M&V.

The CVRMSE calculation is described in Appendix B

ASHRAE Guideline 14 has the following requirements for CV(RMSE) in M&V projects. These are not absolute requirements, but models that do adhere to these bounds are considered very good.

Whole Facility Approach, Prescriptive Path:

For monthly data, with no missing data:

CV(RMSE)		
Data	EE	DSM
< 12 months	20%	30%
12-60 months	25%	35%
>60 months	30%	40%

Whole Facility Approach, Performance Path:

For monthly data, with no more than 25% of the data missing. Savings reported at 68/50 confidence and precision on annual savings.

Retrofit Isolation

Reporting at 68/50 on annual savings

Calibrated Simulation

8760 hourly calculations.

	Monthly Data	Hourly Data
CV(RMSE)	15%	30%
NMBE	5%	10%

Note that by 'calibrated simulation' the ASHRAE guideline (and IPMVP) refers to Building Energy Models calculated by software such as EnergyPlus, and not to spreadsheet or mathematical models. However a simulation can also be built by other software such as spreadsheets and other mathematical models. Building simulations are mostly accepted as these have been peer reviewed and its applicability and accuracy proven in the industry.

¹⁴ as per LBNL <https://eta.lbl.gov/sites/all/files/publications/lbnl-187225.pdf>

13.5. On the use of R^2 and p -values13.5.1. R^2

R^2 is a common and popular metric, usually described as measuring the proportion of data variation explained by the model. However, it should be noted that R^2 is a *relative* measure of goodness of fit for models with the same number of parameters, using the same data set. A perfectly horizontal regression line (in other words a stable process) will have an R^2 value of zero (minimum), even though the model describes the variation in the data perfectly, and it should be one (maximum). This is because R^2 is based on the Pearson moment correlation, which is zero for horizontal lines / stable processes. It is not a measure of uncertainty or of model validity²², and cannot be used to compare one M&V study to another. It is therefore not a valid measure of uncertainty or model precision. Having a model with $R^2=0.8$ *DOES NOT imply 20% precision*.

Adjusted R^2 values may be used as described in the methods outlined elsewhere in this guideline, as part of a broader diagnostic framework, but not as a measure of uncertainty as a sole measure of model validity. CV(RMSE) is a better measure of model validity. However, CV(RMSE)=0.2 *DOES NOT imply 20% precision*.

13.5.2. p -values

The American Statistical Association's statement on interpreting p -values *states that a p -value used in Null-Hypothesis Significance Testing "is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to, or more extreme than its observed value."*²³

The following points from the same document are relevant to M&V uncertainty reporting:

1. P -values can indicate how incompatible the data are with a specified statistical model.
2. P -values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p -value passes a specific threshold
4. A p -value, or statistical significance, does not measure the size of an effect or the importance of a result.
5. By itself, a p -value does not provide a good measure of evidence regarding a model or hypothesis

P -values may be used as stated in this guideline, as part of a broader diagnostic framework.

¹⁵ More information can be found at <https://onlinecourses.science.psu.edu/stat501/node/258>

¹⁶ <http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108>

14. Savings Validation

Measurement, sampling, and modelling uncertainty need to be combined to yield the overall savings uncertainty. For some projects a Monte Carlo simulation or the ASUE method described in Appendix A: Combining Uncertainties, will hold. However, if regression is used, the ASHRAE Guideline 14 method is recommended. It is described in Appendix B

14.1. Sensitivity Analysis

The demonstration of the sensitivity of savings estimate and savings uncertainty to realistic variation in the input parameters and assumptions instils confidence in the energy model.

The simplest and least accurate form of sensitivity analysis is to vary the input parameters one at a time between their maximum and minimum realistic values, and plot how the model responds. In this regard a “Tornado diagram” is a useful tool²⁴.

The one-at-a-time method suffers from the defect that sometimes there is an interactive effect between different parameters. For example, if the model is $a=bc$, and a is between 0 and 10, with a mean of 5, and b is between 0 and 10 with a mean of 5, varying a and b one at a time will vary c between 0 and 50. But if one varies a and b simultaneously, c will vary between 0 and 100. It is therefore useful to do a sensitivity analysis which varies all the input values simultaneously according to a mathematical scheme. Sobol’ sequences, the Morris method, and Latin Hypercube sampling schemes are popular in energy studies.

15. Acknowledgements

This guidance document was developed by the SANAS Specialist Technical Committee (STC) and Working Group members for M&V Inspection Bodies. SANAS would like to acknowledge and thank the members for their contributions in the development of this document.

¹⁷ https://en.wikipedia.org/wiki/Tornado_diagram

Appendix A: Combining Uncertainties

Monte Carlo simulation or the ASUE method is recommended for combining the uncertainties from different sources. The propagation of uncertainties through Taylor Series Approximation as is done in the GUM and ASHRAE RA86 is not recommended, as these methods yield inaccurate results for more complex cases²⁵. Since Taylor series approximation is more difficult to apply than the methods proposed below (it requires integration), it should not be of concern to M&V practitioners.

1. Dominance

In many cases one uncertainty component is much larger than the others. Due to the quadratic addition of uncertainties as shown above, the other uncertainties will then have a small effect on the overall uncertainty.

If the largest uncertainty component is more than four times as large as the next-largest uncertainty component, and can be shown to be normally distributed, that component is said to dominate the overall uncertainty, and other uncertainties may be disregarded. If the dominant uncertainty component is not normally distributed, the *Measurement Good Practice Guide*²⁶ recommends following be added:

“The reported expanded uncertainty is based on a standard uncertainty multiplied by a coverage factor of $k = 2$, which provides a minimum confidence level of 95%. The distribution of the uncertainty is approximately rectangular, U-shaped, etc”. This should be taken into account in the interpretation of the results of this test”

2. Manual addition of Uncertainties

For precision values to be added, they should all be expressed at the same confidence level. It is also assumed that normal distributions describe the uncertainty. Thus, if measurement uncertainty values are at the 95% confidence level, and overall uncertainty is expressed at the x% confidence level, the measurement uncertainty should be converted to the x% confidence level so that the precisions can be added up as

$$p_{\text{total}} = \sqrt{p_1^2 + p_2^2}$$

To convert from uncertainty level z_1/p_1 to z_2/p_2 , one can calculate the new precision p_2 at the desired confidence level z_2 by the formula

$$p_2 = p_1 \frac{z_2}{z_1}$$

where the z-value is the standard score of the distribution.

²⁵ For an informative discussion, see R. Kacker and A. Jones, *On the use of Bayesian statistics to make the Guide to the Expression of Uncertainty in Measurement consistent*, *Metrologia* (2003), 40:235

¹⁸ *Measurement Good Practice Guide No. 36: Estimating Uncertainties in Testing*, Keith Birch, British Measurement and Testing Association <http://www.npl.co.uk/publications/estimating-uncertainties-in-testing>

For example, suppose that the meter precision is 5% at the 95% confidence level, but one wants to express the precision at the 80% confidence level, so that one can add it to other sources of uncertainty. The z-value at the 80% confidence level is 1.28. For the 95% level it is 1.96.

Therefore:

$$p_2 = 0.05 \frac{1.28}{1.96} = 3.26\%$$

This means that 95/5 is equivalent to 80/3.26. If other uncertainties are at the 80% confidence level, the precisions can now be added using the root sum of squares formula given above.

3. Monte Carlo simulation

Monte Carlo (MC) simulation has become a popular method for numerically combining parameter distributions with different shapes, due to its simplicity. Probability distributions are specified for sources of uncertainty. The system is then described by a mathematical model, and this model is evaluated repeatedly for different random draws from the probability distributions specified. The result is a set of numbers forming a distribution which approximates the combination of input probability distributions numerically.

The example below will clarify the process.

According the official guideline applying MC to measurement problems, the following is required for an MC analysis:²⁷

- The mathematical model should be clearly stated and all terms defined.
- The parameters for every input distribution shape specified should be stated and motivated. As with SANS 50 010, parameters should be conservative if there is any doubt as to the values. This is especially true for the variance parameters. For uncertainty analysis, conservatism implies assuming larger variances.
- In cases where the probability distribution shape is uncertain, it should be demonstrated that specifying other plausible distributions does not have a material impact on the output distribution or estimate. Again, conservatism is the guiding principle.
- The number of MC trials required for adequate representation varies with the model. The requirements below do not guarantee convergence, and numerous iterations may be needed to reach convergence. However, as a minimum the number of MC trials should be greater than the maximum of the following:

$$- 10^6$$

$$- 10^4 \times \frac{p}{1-p} \text{ where } p \text{ is the reporting precision}$$

- The model is viewed as converged when the HDI limits do not change with more than 0.5 significant digits when the numbers of MC trials are doubled. [A more cumbersome

²⁸ OIML G 1-101:2008 (E): "Evaluation of measurement data - Supplement 1 to the *Guide to the expression of uncertainty in measurement - Propagation of distributions using a Monte Carlo method*, International Organization of Legal Metrology, Paris, France https://www.oiml.org/en/files/pdf_g/g001-101-e08.pdf

procedure can be used if need be. See Section 7.9.4 of OIML G 1-101:2008(E), see footnote for reference]

- The output probability distribution should be represented by a histogram with no less than 75 bins.

Example

Suppose that a facility's energy use E depends on its baseline energy use per month B , the number of products $n_{products}$ produced by the facility, and the outside air temperature:

$$E = B + n_{products}C + D \times CDD$$

where C is the additional energy used per product and D is the energy used per Cooling Degree Day (CDD). These values could have been obtained by regression. There is some uncertainty in the values of B , C , and D so the following distributions are specified:

$$B \sim \text{Normal}(\mu = 1000, \sigma = 300)$$

$$C \sim \text{Normal}(\mu = 1, \sigma = 0.1)$$

$$D \sim \text{Normal}(\mu = 100, \sigma = 15)$$

If 2000 units are produced in a month, and there were 15 CDDs, the total energy use would be:

$$E = \text{Normal}(\mu = 1000, \sigma = 300) + 2000 \times \text{Normal}(\mu = 1, \sigma = 0.1) + 15 \times \text{Normal}(\mu = 100, \sigma = 15)$$

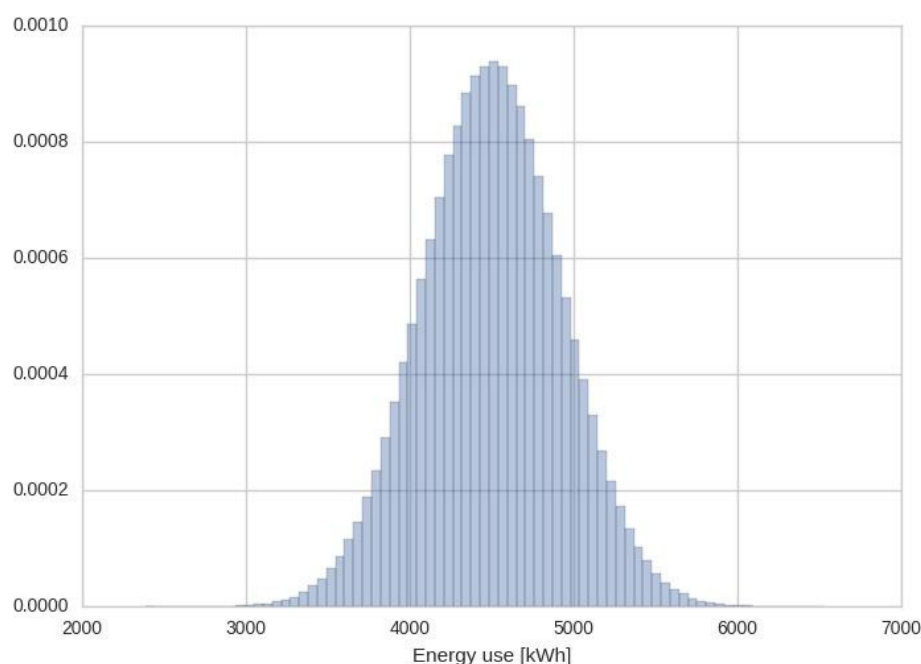
This example was chosen to be simple enough to calculate analytically. The mean of E is:

$$\mu = 1000 + 2000 \times 1 + 15 \times 100 = 4500 \text{ kWh}$$

and the standard deviation is:

$$\sigma_E = \sqrt{300^2 + (2000 \times 0.1)^2 + (15 \times 15)^2} = 425 \text{ kWh}$$

For the MC simulation, distributions with the characteristics described above will be used. 10^6 trials will be conducted, meaning E will be calculated 10^6 times with random draws from the distributions of B , C , and D described above. The mean values for B , C and D will be drawn more often than the extreme values, and over many draws, these will even out to give a distribution on E that reflects the underlying distributions and the mathematical function accurately. The histogram below shows this distribution for E for 10^6 draws, from which the uncertainty can easily be seen and calculated:



The mean for this dataset was given as $\mu=4499.808$, and the standard deviation as $\sigma=425.095$. The 90% confidence interval is (3800.589, 5199.028).

Note that this is a simple example for which MC is not necessary, since the resultant distribution on E can easily be determined by analytical means as shown. However, it is often the case that the distributions on the parameters (such as B , C , and D) are not normal, and that the mathematical function which combines them is difficult. In such cases MC simulation is a useful tool.

There is some variation around the true distribution in a Monte Carlo sample, even with 10^6 trials: different trials will yield slightly different confidence intervals. The values from a single Monte Carlo run can therefore not be used as the final result. OIML Section 7.9.4 Note 7 provides a simple algorithm for deriving these confidence bounds. An algorithm for obtaining 80% equal-tailed confidence interval that will be true for 99% of MC simulations using 10^6 trials, is the following:

1. Obtain the sample as described above. From the example, we call it E .
2. Sort E in ascending order.
3. Let $E_{sorted}[1]$ denote the first value in sorted sequence of E , $E_{sorted}[2]$ the second value, etc. $E_{sorted}[99000]$ is the lower limit of the 80% confidence interval, and $E_{sorted}[901001]$ is the upper limit of the 80% confidence interval. This is true regardless of the distribution, although it is less conservative when E is normally distributed.

Note that these values will change if a different number of Monte Carlo trials are used, or a different confidence interval or coverage factor is desired. For other, less conservative methods, see OIML Section 7.

Note that this example is only for the adjusted baseline energy use. The actual energy use needs to be subtracted from this distribution.

4. ASUE / Mellin Transform Moment Calculator Method

One alternative to MC is to use the ASUE tool²⁸. This tool uses advanced mathematics (the Mellin Transform) in the background to produce equations describing the mean and variance of the resultant probability distribution of a combination of probability distributions, described by a polynomial. For the example above, it yields the equations for the mean and standard deviation as described above. When a more complex function than that of E above is used, and different distributions are combined, the resultant equations may become long. However, they can still be solved by simply substituting the values of the constituent distribution parameters into the resultant equations for mean and variance.

For example, in ASUE terminology, let $B=x1$, $C=x2$, $D=x3$. Each one of these is a distribution defined as:

$$x1 \sim Normal(\mu = m1, \sigma = s1)$$

$$x2 \sim Normal(\mu = m2, \sigma = s2)$$

$$x3 \sim Normal(\mu = m3, \sigma = s3)$$

The mean of the resultant distribution is $m1+2000 \times m2+15 \times m3$

The standard deviation of the resultant distribution is: $\sqrt{s1^2 + (2000s2)^2 + 15s3^2}$

Substituting the values of the individual parameters into these equations yields the same answers as for the Monte Carlo simulation.

The confidence interval can be calculated from the mean and standard deviation by any statistical software package. Doing it by hand, for a normal distribution, one can use the following equation:

$$LCL = \mu - z\sigma$$

Where μ is the mean, σ is the standard deviation, and z is the standard score. The standard scores for different confidence levels are as follows:

Confidence	z-score
68%	1
80%	1.28
90%	1.645
95%	1.96
99%	2.576

For cases where the resultant distribution is non-normal (which can happen even when multiplying and dividing normal distributions), the ASUE method can be extended using the Mellin Transform Moment Calculator (MTMC) as follows. The mean and variance are what are called 'moments' of the distribution. Moments define the shape of a probability distribution. The mean is the first moment, and the variance the second moment. The third moment is skewness, and the fourth moment kurtosis (fat-tailedness). There are higher moments as well,

²⁹ <http://aue.erc.monash.edu.au/webMathematica/Asue/Evaluate.jsp>

but these four moments define a distribution reasonably accurately. The Johnson distribution was developed specifically for fitting a distribution to these four moments. Using the first four moments obtained from the MTMC²⁹, we can find the distribution $\sim \text{Johnson}(\text{mean}, \text{variance}, \text{skewness}, \text{kurtosis})$. There are functions in Matlab/Octave, R, Mathematica, and other software packages for plotting and doing calculations with the Johnson distribution just as with any other distribution.

The MTMC yields $E[y]$, $E[y^2]$, $E[y^3]$, and $E[y^4]$ (which may also be called moments, causing some confusion), although the mean, variance, etc. output may be added soon. To calculate the mean, variance, skewness, and kurtosis from these moments, the following formulae are used:

$$\text{Mean} = E[y]$$

$$\text{Standard Deviation} = \sqrt{E[y^2] - E[y]^2}$$

$$\text{Skewness} = \frac{E[y^3] - 3E[y]E[y^2] + 2E[y]^3}{\text{Standard Deviation}^3}$$

$$\text{Kurtosis} = \frac{E[y^4] - 4E[y]E[y^3] + 6E[y]^2E[y^2] - 3E[y]^4}{\text{Standard Deviation}^4}$$

²¹<http://tc32.ieee-ims.org/content/mellin-transform-based-moment-calculator-multivariate-polynomials>

Appendix B: ASHRAE Guideline 14 Method for Combining Uncertainties

ASHRAE Guideline 14 2002: Measurement and Verification of Energy and Demand Savings contain a useful method for combining uncertainties when linear regression is used.

Using equation 4-7, all the uncertainties that interested and effected parties would be concerned about are included – namely:

- Regression or modelling error
- Sampling error
- Energy instrument measurement error
- Independent variable measurement error

The total uncertainty in equation 4-7 has a simplified way of dealing with progression of the baseline modelling error into the reporting period without the elaborate use of matrix algebra – which is needed to deal with autocorrelation and different sample sizes of the baseline period versus the reporting period. Note that the methods described in the section are model and energy process independent.

Equation 4-7:

Note that the ASHRAE Guideline 14-2014 has an error in equation 4-7, where $U = \frac{t}{F} = \dots$
Comparing to Eq. 4-6 or ASHRAE 14-2002 shows that the version below is correct.

In cases where baseline energy use or demand varies from period to period in response to known independent variables:

$$U = \frac{t}{F} \sqrt{\frac{CV(RMSE)^2}{m} \times \left[\frac{n}{n'} \left(1.6 + \frac{3.2}{n'} \right) \right] + U_s^2 + RE^2_{instrument} + U_{iv}^2}$$

where:

U : Relative uncertainty in reported savings, expressed as a percentage.

CV(RMSE): Coefficient of Variation on the Root Mean Square Error:

$$CV(RMSE) = \frac{\sqrt{\frac{\sum (y_i - \bar{y})^2}{n - p}}}{\bar{y}}$$

t : t-statistic determined at $p=1$.

F : Fractional savings. Approximate percentage of savings achieved relative to baseline.

Note:

CV(RMSE) represents all the baseline model uncertainty - with the following limits: ASHRAE 4.3.2.1. Prescriptive path: CV(RMSE) < 25% if there are 12 months in the reporting period. This is based on a minimum sample size of $n = 9$. ASHRAE 4.3.2.2. Performance path: CV(RMSE) not specified, but total

uncertainty $U < 50\%$ at confidence level 68%. Also, excluded data must be less than 25% of baseline data.

IPMVP and ASHRAE do not deal adequately with model verification and validation (hence the inclusion of sections 3, 4, 5, and 6 of this guideline).

U_s : Sampling Uncertainty:

$$U_s = \frac{100}{\bar{y}} \times \sqrt{\frac{\left(1 - \frac{q}{Q}\right) \left[\sum_{i=1}^n \frac{(y_i - \bar{y})^2}{(q-1)} \right]}{q}}$$

q : size of random sample.

Q : Population size.

m : number of periods in baseline (months, weeks, days).

$\bar{y}$: Arithmetic mean of sampling observations.

y_i : model's predicted value at i .

n : Number of post-retrofit reporting periods.

n' : Number of independent observations in a total of n observations during baseline period, calculated as:

$$n' = n \times \frac{1 - \rho}{1 + \rho}$$

where ρ is the autocorrelation coefficient of a series of n observations at lag 1, derived from performing a regression of the series of n observations against the same data series offset by one time-step. **Noting that the ASHRAE guideline formula contains a typing error**, the correlation coefficient is as follows:

$$\rho = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

For monthly data, this guideline permits an assumption that ρ is 0, so $n'=n$.

p : number of parameters or terms in baseline regression model.

$RE_{instrument}$: Relative error of instrument, expressed as percentage. If multiple instruments are used in series, the overall precision can be calculated as follows, provided that the precisions are all expressed at the same confidence level:

$$RE_{instrument} = \frac{\sqrt{\sum_{i=1}^c (RE_{instrument} \times r_{rating,i})^2}}{\sum_{i=1}^c \bar{r}_i}$$

where r_{rating} refers to the value relative to which the instrument precision is expressed.

U_{iv} : Postretrofit energy savings uncertainty due to independent variable measurement error, expressed as a percentage.

Baseline energy use or demand constant

$$U = \frac{t}{F} \sqrt{\frac{CV(STD)^2}{m} + U_s + RE_{\text{INSTRUMENT}}^2}$$

where:

$$CV(STD) = \frac{\sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}}{\bar{y}}$$

Appendix C: References to relevant sections in other guidelines

The ASHRAE section 4.2.11.4 – Computing Savings Uncertainty may be “codified” and specified for use by any of the interested or effected parties – thereby providing a consistent means of combining uncertainty. The “codified” table could be as follows:

Aspect	EVO-IPMVP	ASHRAE	UMP
Model Uncertainty	Section 2	Section 4.2.11.3. eq 4-4 (< 25%)	Chapters 8,
Sampling Uncertainty	Section 3	Section 4.2.11.1. eq 4-1	Chapter 11
Instrument Measure Uncertainty (Energy)	Section 4	Section 6; Appendix A2.1	Chapter 9
Independent Variable Measure Uncertainty (Energy Drivers)		Appendix A2.2 – A2.13	Chapters 13, 23, also addressed for specific technologies elsewhere
Total Uncertainty	Section 5	< 50% at CL 68%	
Maximum data exclusion		< 25%	

Deviations from the above require that the baseline and assessment reports clarify how these have been dealt with.

Appendix D: References and Further Resources

Various international guidelines provide additional guidance on uncertainty, and M&V professionals are encouraged to familiarise themselves with these documents:

1. Bonneville Power Administration's Regression Reference Guideline
<https://conduitnw.org/layouts/Conduit/FileHandler.ashx?RID=506>
2. The American Society of Heating, Refrigeration, and Air-Conditioning Engineers' Guideline 14 on the measurement and verification of energy and demand savings:
<http://esl.tamu.edu/docs/terp/2005/ESL-IC-05-10-50.pdf>
3. The California Commissioning Collaborative's Guidelines for Verifying Savings from Commissioning Existing Buildings: <http://www.cacx.org/resources/vos-guidelines/>
4. The Clean Energy Ministerial's guiding documents:
<http://www.cleanenergyministerial.org/Our-Work/Initiatives/Energy-Management/Provide-Support-Resources#MVGuidance>
5. The International Performance Measurement and Verification Protocol's Statistics and Uncertainty Guideline (available from the Efficiency Valuation Organisation after logging in: <http://evo-world.org/en>)
6. The State and Local Energy Efficiency Action Network's Energy Efficiency Program Impact Evaluation Guide: <https://www4.eere.energy.gov/seeaction/topic-category/evaluation-measurement-and-verification>
7. The Uniform Methods Project <https://energy.gov/eere/about-us/ump-home>

Appendix F: Statistical Software packages for M&V Uncertainty Quantification

The tests described in this document have been automated in almost all statistical software packages, many of which are freely available. M&V professionals are encouraged to familiarise themselves with at least one package. Once this is done, the danger lies in an abundance of tools, rather than unfamiliarity with them. It is fairly easy to create standard templates or automatic routines for the most common kinds of M&V analyses.

Note that the open-source packages listed below are by no means inferior to the proprietary ones. In fact, the world's leading tech companies and research institutions use Python and R. The speed and level of support offered by their online user communities are also often superior to those of paid packages³⁰.

The most relevant ones for M&V, in alphabetical order, are:

- **Excel:** Good for basic regression models, but limited. There are add-ons that enhance Excel's statistical capabilities. Not considered to be statistical software by professional statisticians, but good enough for many analyses.
- **Mathematica:** Proprietary mathematical programming language.
- **Minitab:** Proprietary statistical software. Popular with industrial statisticians.
- **Python:** Becoming as popular and powerful as R, being more user-friendly and flexible. A fully-fledged programming language, rather than just a statistical language. The Anaconda installation from Continuum Analytics contains all of the modules for standard uncertainty quantification.
- **R:** The original and standard open source statistical and data science programming language. Relatively steep learning curve.
- **SPSS:** Proprietary statistical software. Easy to use for non-statisticians. Graphical User Interface, with programming secondary. (Open source version: PSPP)

³⁰Wikipedia maintains a tabular comparison: https://en.wikipedia.org/wiki/Comparison_of_statistical_packages

ADDENDUM 1: Amendment Record

Proposed By:	Section	Change
STC	Equation 4-7	Corrected