Australia's National
Science Agency

Automated M&V2.0 Current status and challenges

Review of RACE for 2030 White Certificates Project

Mark Goldsworthy

September 2023

RACE for 2030
RELIABLE
AFFORDABLE
CLEAN
ENERGY

Citation

Goldsworthy M, (2023). Automated M&V2.0 Current status and challenges. CSIRO, Australia.

Contents

Figures

Tables

1 Introduction

Measurement and Verification (M&V) refers to the process of using data in combination with some form of modelling or calculation to estimate energy, cost and/or emissions savings due to a site upgrade such as equipment replacement or control system change. Savings are *estimated* and not directly calculated because the counterfactual case (i.e., the energy use of the site without the upgrade) is hypothetical, and so cannot be directly measured.

Traditional M&V approaches include use of equipment energy use measurements combined with engineering calculations, monthly analysis based on energy bills, regression analysis applied to energy meter data, and calibrated building envelope energy simulations to perform a one-time savings analysis for a specific upgrade or 'intervention'. The most common application for traditional M&V is to quantify savings from such interventions to qualify for incentives in energy efficiency schemes.

This review focuses specifically on the status and challenges of M&V2.0 or 'Advanced M&V' (terms used interchangeably here). According to Granderson & Fernandes (1) M&V2.0 is "increasingly understood to refer to the use of automated analytics in combination with higher granularity data to quantify project energy savings". The use of automated data feeds and semi or fully automated analysis aims to reduce the overall M&V transaction cost and hence make assessments of smaller energy upgrades more feasible. However, the combination of automation and higher resolution data (hourly or sub-hourly) opens the possibility for continuous rather than once-off application of the M&V method enabling use cases such as regular performance monitoring, reporting, and benchmarking as well as automated anomaly/fault detection. These ongoing applications may provide greater value to a site than once off savings estimates from a specific intervention.

Use of M&V methods to quantity energy savings is codified in regulations in numerous jurisdictions around the world including across Europe, the United States and Australia. Most regulations refer to the International Performance Measurement and Verification Protocol (IPMVP) (2) which is discussed further in Section 3. In Australia, the NSW Energy Efficiency Scheme (ESS) references IPMVP for general guidance under the Metered Baseline Method (MBM) (3) as well as the Project Impact M&V Method (4) while the non-mandatory guide to the Victorian Energy Upgrades program (VEU) (5) also mentions IPMVP as a source of further information. In SA the Retailer Energy Productivity (REPS) scheme (6) references both the NSW and Victorian M&V methods. While most other states and territories have some form of energy efficiency incentive scheme, they are based on financial incentives typically provided at the point of sale for products deemed eligible, hence they do not reference an M&V approach. Hence in this report, the primarily source of authoritative information or guidelines on M&V analysis is taken to be the IPMVP.

While IPMVP, and in particular the associated IPMVP Uncertainty Assessment Guide (7) in combination with the ASHRAE 14 Guideline (8) which is referenced in several places, provide a good basis for applying an M&V method, neither was written specifically with automated and high-resolution data analysis (i.e. M&V2.0) in mind. The recent IPMVP White Paper on advanced M&V (9) discusses four current challenges with applying M&V2.0 in a generic way to any site. These are; handling of non-routine events ('site-level changes'), calculating savings uncertainty, dealing with difficult to model buildings or sites (so called 'bad-buildings'), and use of models for calculating aggregated savings across multiple sites.

Given here the focus is on M&V applied to individual sites, we do not consider the use of aggregate M&V models. Instead, an additional challenge is specifically identified which is use of net (billing) meter data in the presence of on-site generation. This is expected to be critical issue by itself even though it could be considered to fall within the general challenge of 'difficult to model' buildings. Thus, we consider the four key issues facing site level M&V are:

i) Use of net metering for M&V analysis where onsite generation is present. Given the wide-spread availability of relatively good quality, high frequency data from net interval (billing) electricity meters, it is tempting to use this data directly in an M&V analysis. However, since M&V is based on an analysis of energy *consumption*[1], the presence of any onsite generation means that use of the net meter data directly, without accounting for generation (for example, by installing a separate generation meter) is, strictly speaking, incorrect. For example, Crowe et al. (10) applied M&V analysis to interval meter data for 137 commercial buildings and found that, of the 6 buildings where the baseline model failed to fit the data, 5 failed due to the presence of onsite PV generation. Further work is required to determine the extent to which the presence of onsite generation (and storage) can be tolerated in data used to perform M&V analysis including the impact on savings uncertainty.

ii) Handling of non-routine events. Non-routine events (NRE) are changes to a sites energy use that are not accounted for by the M&V model. That is, they occur when some unexpected event occurs or when a factor that influences energy consumption that is expected to be fixed (a so-called 'static factor' (8)) changes. As discussed by several authors (11) (12) (13), automatically identifying *potential* NRE from the energy data alone and distinguishing them from routine energy use is challenging but not insurmountable. Developing a self-contained M&V2.0 application with a semi-automated process (whereby a user can be guided and prompted to provide additional information and/or confirmation where necessary, and where the program implements the necessary adjustments in a reliable way), is more difficult and will require significant development, testing and refinement.

iii) Quantification of model uncertainty & accuracy. A core component of any M&V analysis is calculation of the uncertainty of estimated savings. While IPMVP and ASHRAE 14 describe methods for calculating uncertainty they are based largely around daily or monthly M&V analyses. As discussed by Touzani et al. (14), when applied to higher resolution (hourly) models the suggested uncertainty estimate approaches are approximate only and tend to overstate model accuracy. Hence, further work is required to develop accurate uncertainty characterisation approaches that can be deployed regardless of model formulation and that also factor in the influence of any non-routine-adjustments applied to the base model.

iv) Dealing with difficult to model sites. For some sites, standard model variables such as weather and time are insufficient. In the best case, additional variables may be required to create a model with sufficient accuracy (noting though that these variables may not have been included in the M&V measurement plan). In the worst case, the energy use variability may be 'intrinsic' to the site (for example behavioural) and cannot be modelled at all at a given time resolution. For example, a study by LBNL (15) applying hourly M&V models based on interval meter data from over 48,000 buildings found that, for office buildings, the models met the minimum performance criteria only 29% of the time (although 69% of large office buildings met the criteria). Measures that can quickly identify sites where automated M&V processes may fail, or where lower resolution (longer time interval) models are required, are needed to avoid unrealistic expectations and to enable timely assessment of the proposed M&V plan.

Central to each of the above four challenges is the question of the extent to which the M&V2.0 analysis can be fully automated, and, if not, the conditions under which intervention by an M&V expert is required. Although the 2020 EVO White Paper (9) anticipated an upcoming advanced M&V guide to be published by EVO in 2020 which

[1] Notwithstanding the use of M&V to calculate peak demand which is generally related to net import of power from the electricity grid.

would provide guidance on these issues, no publication was available at the time of writing (August 2023) suggesting that further work to is still needed to address these challenges.

The remainder of this report provides; a brief overview of M&V2.0 in demand response applications (Section 2); a review of literature including key standards, guidelines and M&V algorithms and the extent to which the M&V analysis has been automated (Section 3); a description of the DCH M&V algorithm developed by CSIRO (Section 4); and finally a more detailed exploration of the current challenges as identified above (Section 5).

2 M&V2.0 for demand response

M&V algorithms may be used to calculate long term demand reduction resulting from equipment upgrades or removal. The primary motivation in this case is for estimating bill savings, though the demand savings may also be of interest to regulators or electricity grid operators. Here the relevant demand reduction may correspond to a particular time window and/or interval (for example half-hour intervals, or workdays during the evening), in which case the M&V model would ideally have a short enough time interval to resolve the period in question. ASHRAE 14 (8) outlines both energy and demand reduction procedures such as the application of representative factors to estimate 24-hour demand profiles from daily or longer interval data (18) and use of 90^{th} percentile monthly demand to estimate monthly peak demand. Estimation of single (peak) events, for example for determination of a 12-month rolling peak demand for a ratcheting capacity charge calculation are likely to be problematic for any M&V method due to the dependence on single events.

Automated M&V2.0 algorithms designed to calculate long term energy and demand savings can also be used to calculate baseline electricity which can be used to calculate avoided electricity use through activation of short-term DR actions (17). In this application, baselines can be used for managing market operations, for example serving as forecasts of 'usual' operation, and/or for settlement (i.e., providing payments for provision of DR services). Given the significant attention at present in the closely related topic of grid interactive buildings and demand response (DR) (16), here we focus on this second application of M&V algorithms involve demand estimation.

In a short-term demand response estimation application, the M&V algorithm is used to construct the baseline model generally from whole of site meter data, and the model estimates are then compared with actual measurements to quantify the DR amount. In contrast to M&V for energy consumption assessment, M&V for demand response may/ may not be concerned with net energy flow to the grid, hence 'behind-the-meter' generation may need to be included in the analysis.

Most DR mechanisms that have been implemented around the world incorporate baseline models that are combined with measurements to calculate the reduction in load provided by the demand response device (see for example (18)). Most of these schemes employ relatively simple baselining approaches that use some combination of averaging of historical electricity demand on reference days either with or without an offset or correction factor to account for demand immediately prior to the DR event. Although IPMVP briefly mentions M&V for demand savings, no detailed methodology is given. A number of studies have been conducted comparing regression-based baseline methods as used by many M&V2.0 algorithms with these simple baseline approaches (18) (19) (20) (21). Most have found that the simpler approaches perform similarly or better while requiring less data and being easier to implement. The latter being a critical consideration for large scale implementation.

In Australia, the Demand Response Mechanism (WDRM) commenced in 2020 and allows 'large' customers (>100MWh/year depending on jurisdiction) to participate in a demand response market (22). Eligible sites (or in certain cases aggregations of sites) provide price-volume bids for demand reduction as 'Wholesale Demand Response Units' in units of integer MW and are dispatched by the market operator in a similar way to large scale generators and scheduled loads. Settlement is calculated based on the estimated DR which is calculating by comparing the metered electricity with the baseline estimate. Currently only 1 type of baseline methodology (CASIO10) is accepted (with four variants) in the scheme (23) and so sites where the load is highly variable or has a significant weather dependence may not be eligible. An additional factor is that the baseline criteria is applied per site, as opposed to at the aggregate level which increases stringency. AEMO provides a calculator tool to determine if a site is likely to meet the baseline model eligibility criteria (24).

Recently a new DR market mechanism, 'Schedule Lite' has been proposed (25). This incorporates two models; the 'visibility' model and the 'dispatch' model. Neither specifically uses baselines to estimate DR. Instead, aggregators of demand response capacity (termed 'traders') provide forecasts of baseline demand coupled with

either indicative (in the case of the visibility model) or actual price-volume bids for DR amounts. In the 'dispatch' model the bids are used in the formal dispatch process, in the 'visibility' model the bids are used to enhance the market operator's visibility of flexible demand and hence improve forecasted demand.

While baseline models are not specifically required, the forecasts of available capacity will presumably need to be based on some sort of baseline model combined with estimates of available DR corresponding to specific market prices. A compliance process run by the market operator will ensure that these forecasts match actual metered electricity demand. Critically it is proposed that the scheme is not restricted to large customers, and that the forecasts are only required for the aggregated DR of many small customers. This suggests that M&V2.0 algorithms might be suitable for use as aggregate load forecasting tools under such a mechanism, and some studies in the literature have evaluated the suitability of M&V2.0 algorithms for aggregate or portfolio estimates of consumption based on site level metering (26) (27) (28). The suitability of different M&V algorithms to model the baseline site load and/or the load variations under different types of demand response is an open research question.

3 Existing literature

A detailed review of the M&V literature is beyond the scope of this document. Instead, here an overview is given of key guidelines related to M&V2.0, a description of different algorithms and tools that have been proposed, and a discussion of automation in the context of M&V2.0. Further information may be found in the cited references.

3.1 Guidelines

The principal guideline for M&V practitioners is the International Performance Measurement and Verification Protocol (IPMVP). First published in 1997, available at no cost and continually updated, it consists of several separate documents including Core Concepts (2), the Uncertainty Assessment Guide (7), the Application Guide for non-routine events (29), and the Renewables Application Guide (30) amongst others. IPMVP makes limited references to ASHRAE Guideline 14 (8) which provides more technical details in certain areas but was last updated in 2014 and was written largely from a conventional M&V perspective.

In the US, the Federal Energy Management Program (FEMP) has published a guideline for applying M&V (31) to calculate savings. The FEMP document, first published in 2000 was aimed to provide more specific guidance given that at the time the IPMVP lacked the necessary detail needed by practitioners. More recent versions of IPMVP have largely addressed this gap. Also in the US, the Uniform Methods Project (UMP) (32) is a 1000+ page document that provides individual guidelines, methods, and examples for applying M&V for specific types of energy efficiency upgrades. However, the UMP does not cover advanced M&V, for example mentioning use of interval meter data only in the context of program level M&V across many sites (See Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol).

In Australia the Australasian Energy Performance Contracting Association published a best practise guide to M&V in 2004 (33). This document was based on early versions of IPMVP, ASHRAE 14 and FEMP, and provided a good overview but has been largely superseded by later versions of the IPMVP. As noted above, the NSW (3) (4) (34) and VIC government (35) provide guidelines on how to apply specific M&V methods to satisfy their scheme requirements. For the case of non-routine adjustments (34) provides detailed information on when and how to apply adjustments.

3.2 Algorithms and tools

A significant body of work exists in the literature on M&V2.0 algorithms, and there are numerous free and commercially available M&V tools targeting different applications and with different features. For example, Granderson & Fernandes (1) reviewed 16 advanced M&V tools and categorised them by multiple features including target users, application area, method, input data, algorithm and metrics, approach to uncertainty calculation, adjustable parameters, and code transparency. The majority of tools apply some form of linear regression model, though machine learning approaches were also common. Surprisingly only approximately half included some form of uncertainty calculation. Application areas were mostly commercial though several considered residential or industrial as well. Approximately half had an open code base with the rest preferring to keep the algorithms proprietary.

Six of the tools involved the tool vendor regularly assessing the actual M&V model for fitness and suitability which may indicate an analysis workflow that is not fully automated or failsafe. A recent updated work (9) published by EVO reviewed 4 additional tools each of which includes a linear model as the default option.

In the scientific literature, many different types of M&V algorithms have been described and evaluated. Grillone et al. (36) review the literature and categorise data-driven M&V methods into i) statistical, ii) machine learning,

and iii) Bayesian methods. The most commonly applied linear regression methods; time-of-week and temperature (TOWT) and change-point fall under the category of statistical methods, while various types of machine learning (artificial neural networks, support vector machine, random forecast and gradient boosting machines) have all been applied to M&V. Alrobaie and Krarti (37) also recently reviewed data driven approaches for M&V and concluded that for non-linear methods there is no general framework or guideline which may be limiting their uptake in practical M&V tools.

In comparison to machine learning and Bayesian methods, the linear statistical models are relatively simple to apply, computationally fast, easily interpretable and have been shown by several authors to give good results (38) (39) and also results comparable to or better than more advanced methods (14) (40) (41) (42) (43).

3.3 Automation

As noted above, automation is a central aspect of M&V2.0. Here it is proposed that the extent of automation in an M&V workflow can be categorised according to Table 1. Here a distinction has been made between *data* automation and *analysis* automation. The former relates to the form and method of supplying input data to the M&V analysis, including the meta-data that provides the linkage between data and model. The latter relates to the M&V analysis itself and the extent to which an expert user (for example the tool vendor, the tool user, or a 3rd party) is required to choose analysis options, interpret results and to ensure that the analysis is valid and as accurate as possible (or at least sufficiently accurate for the intended purpose). Any given M&V application may have one level of data automation and a different level of analysis automation - which we denote using the short-hand notation D1A3 (i.e. data automation level 1, analysis automation level 3).

Table 1 Summary of M&V2.0 levels of autonomy

Automation level	Data automation	Analysis automation
Level 1	Input data must be supplied pre-processed (cleaned, synchronised and/or filled). User must upload data file(s) and specify all configuration options.	M&V specialist user must interpret model validity & results as well as design & initiate new analysis if required. Minimal model diagnostics and checks (for example standard metrics such as R2, CVRMSE only).
Level 2	Raw input data can be supplied. User must upload data file(s) and specify most configuration options.	Detailed model diagnostics are provided with limited suggested changes to analysis options based on default model performance. M&V specialist required to initiate (configure) and run new analysis.
Level 3	Data is obtained from an automated service (e.g., database, API, cloud service). All processing is done by the application. Minimal configuration options required (for example specify data linkages & site location).	Natural language processing of outputs & application diagnostics are reported & interpretable by a non-specialist. Suggested analysis changes provided to user for review. Analysis automatically reruns based on user selection.
Level 4	Data is obtained from an automated service. Semantic models are used to provide data interpretation & linkage. Analysis across multiple sites can be done with no specific user setup.	Application automatically self-diagnosis model performance. (e.g., non-routine events are automatically identified and accounted for. Different model options are automatically trialled if the initial model does not meet specifications.) Invalid model results cannot be provided as final outputs.

Based on the literature reviewed here, no existing M&V applications have been identified that provide full automation (D4A4). For example, Ke et al. (44) discuss a platform for automated M&V however, it requires users to upload data and choose model terms and options suggesting only D2A1 level automation. The vast majority of applications described in the review by (1) use D2 level data automation or less and A1 or A2 analysis automation. As discussed below, the current CSIRO-DCH M&V application uses D3A2 automation at present, though plans are to extend this to D4A2 shortly and D4A3 in the medium term. Moving to self-diagnostics and fully automated analysis using semantic models is the focus of current and future research.

4 DCH CSIRO M&V algorithm

4.1 Overview

An M&V2.0 algorithm (referred to here as MVApp) has been developed by CSIRO based on an extended TOWT statistical model. At present, this algorithm runs from command line (i.e., does not have a user interface) and optionally connects to the DCH (Data Clearing House) cloud platform for receiving interval meter and weather data and writing program outputs. It can also be run without a connection to the DCH platform, for example using data from a local source such as a csv file or other database.

The algorithm can be divided into a sequence of steps as summarised in Figure 1. An analysis begins with the user creating an input configuration text file (JSON formatted) which can be used to specify non-default program configurable options and settings as well as key inputs such as the start and end dates for the pre & post intervention periods (referred to as the 'baseline' and 'reporting' periods respectively). Work is currently underway to eliminate the requirement for this configuration file by using the semantic building model to provide building meta-data (for example, location and key data sources) and when default analysis options are used. It would also be possible for a UI to be created that automatically creates this configuration file based on user selections.

Once MVApp is run with a configuration file the application first fetches data from the data source and then cleans, processes, and checks the data for suitability. If the data is deemed valid, the cleaned data version is input to the baseline model training algorithm. Normalised savings are then computed, and output quantities and uncertainties are calculated. Further details are given below.

Outputs are written in a variety of formats including monthly and daily summary csv's, a Word® report, raw data files and data-stream outputs written back to the DCH platform. The complete process from initialisation to results output takes approximately 30 seconds on a desktop computer for an analysis using 2 years of half-hourly data for one site.

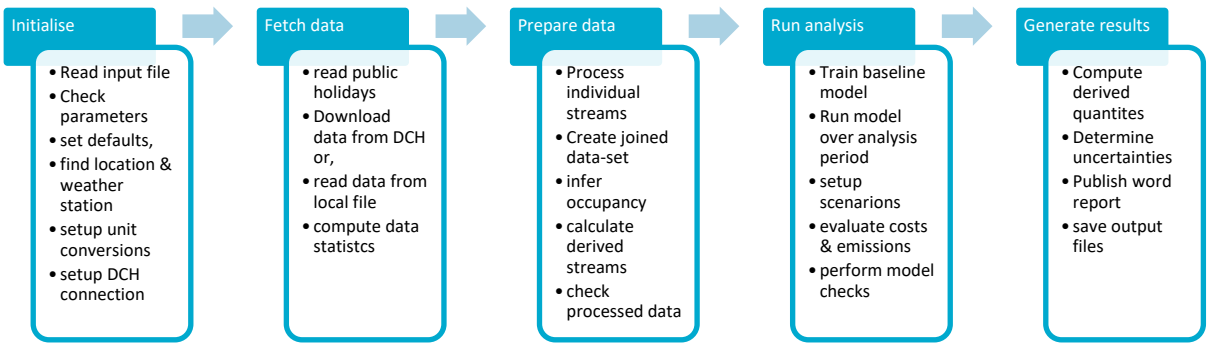

Figure 1 Overview of CSIRO M&V2.0 (MVApp) analysis sequence

4.2 Algorithm

4.2.1 Interval model

The interval model is typically used with an hourly time-interval, although both longer and shorter time-interval may also be used. The core algorithm is based on weighted piecewise continuous multi-linear regression for the overall site consumption $eload_i$ as given by Eq. 1 and described in (45).

$$eload_i = \sum_p w_{p,i} \left(\alpha_j - \beta_{0,occ} \min(T_1 - T_i, 0) + \sum_{l=1}^{n_{l,occ}} \beta_{l,occ} max(T_i - T_l, 0) + \gamma_{occ=1} h_i + \eta G_i \right)_p \quad \text{Eq. 1}$$

The baseline period is divided into N periods, each with a time duration (excluding missing data) of 30 days by default. A separate model is fit for each period *p* and weighting factors $w_{p,i}$ used to combine the predictions from each model for a given time interval *i*. The method used to determine the weighting factors is not covered here.

Continuous piecewise temperature levels are used to represent the variation of consumption with temperature after the effect of time has been excluded. As shown in Figure 2, the piecewise regression component consists of $n_{l,occ=1} + 1$ coefficients $\beta_{l,occ=1}$ for occupied time intervals and $n_{l,occ=0} + 1$ coefficients $\beta_{l,occ=0}$ for unoccupied time intervals. Adaptive temperature levels are determined based on the entire baseline period data using a 'leap-frog' method designed to ensure a minimum amount of data in each level but with relatively uniform spacing and limits applied to the maximum number of levels.

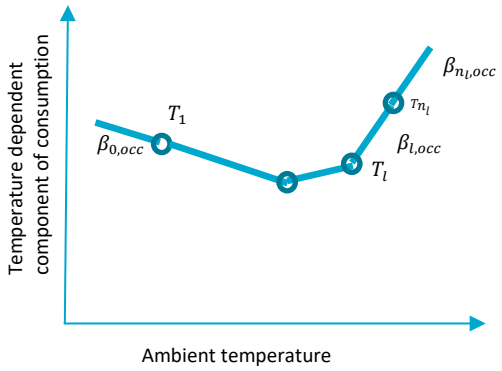

Figure 2 Schematic showing piecewise continuous temperature changepoints and associated coefficients

By default, coefficients α_j are determined for each hour interval of the week. Different options are available for the number of coefficients which is beyond the scope of this overview. The theoretical maximum number of coefficients that can be fit is 336 (24x7x2) (in the most general case each interval can be either occupied or unoccupied), however a more typical number is around 168 since each interval of the week is usually assigned as either occupied or un-occupied. Coefficients are not fit for intervals with insufficient data. In cases where an interval-occupancy combination has multiple data points, but the same interval and opposite occupancy combination has zero or 1 data point, only the former coefficient is fit and used for predicting both cases.

An optional additional coefficient $\gamma_{occ=1}$ can be fit for all occupied intervals that fall on public holidays. This applies an additional scaling factor that can be useful to account for sites with less than usual but above baseline consumption on public holidays. By default, this coefficient is included.

For the special case where generation data is unavailable and where the only onsite generation is from PV arrays, it is possible to fit the model directly to net metering data. Enabling this option includes an additional coefficient η in the model which is proportional to an irradiance value G_i for the given interval. If DCH is used for input data, global irradiance data is automatically fetched for the site location. (In general, for standard fixed mount PV panels it is sufficient to use global horizontal irradiance data even though panels may be mounted at a significant incline.) Note that, as outlined in Section 5.1, fitting of a model to net meter data, depending on the site and data, may not result in a valid model. Calculation of savings uncertainty at various resolutions (e.g., daily, monthly) is performed using the methods outlined in ASHRAE 14 (8) or optionally using cross-validation.

4.2.2 Daily model

MVApp includes a daily analysis model that fits an equation of the form given in Eq. 2 to the daily energy consumption (or net energy). Here CDD_d and HDD_d are the daily cooling and heating degree days respectively and are calculated using a dynamic (hour of the day dependent) temperature level. The coefficients *a* and *b* in Eq. 2 are varied and used to fit different baselines models with the best model chosen based on Bayes Information Criteria. G_d is the daily irradiance which is optionally used in the case where PV generation is included in the meter data and h_d and wk_d are binary flags indicating public holiday days and weekends respectively.

$$eload_d = \beta_1 h_d + \beta_2 wk_d + \beta_3 CDD_d + \beta_4 HDD_d + \beta_5 G_d \quad \text{Eq. 2}$$

$$TL = a + bsin(\pi(h-7)/12)$$

$$CDD_d = \sum \max\left(T_l - TL, 0\right), HDD_d = \sum \max\left(0, TL - T_l\right)$$

Equation 2 is fit using daily quantities that are calculated from the same interval data that is used by interval analysis model.

4.3 Data and model checks

Prior to fitting the baseline model a range of automated data checks are applied to the processed input data streams to identify potential issues. These include missing data, sufficient data, not-a-number, outlier and range checks, a periodicity check, time related checks, checks for sequences of repeated values and unique values, as well as monotonicity, trend, and autocorrelation checks. Depending on the test and computed result the application may return warnings or errors which can be configured to cause the analysis to be aborted.

Once the baseline model has been trained on the data several model checks are also performed. These include checking for daily outliers, checking for entire months that are under/over predicted, linear and seasonal trend checks on the model error, checks on the model fitting parameters and a check on the estimated cumulative saving trend. Currently these checks are configured to provide warnings to the user only. Future work may involve implementing automated processes and/or guided workflows to manage these warnings with minimal user input.

4.4 Outputs

MVApp calculates energy, cost and emissions savings from grid electricity use. Cost savings are calculated using wholesale spot prices and/or time of use energy tariffs as supplied in the configuration file and may/may not include feed-in credits. Greenhouse gas emissions can be calculated either using fixed emissions factors or, if

using the DCH as the data source, time-varying and region-specific emissions factors, and can also be configured to calculate avoided emissions (emissions reduction resulting from net export of energy to the grid).

All outputs of the program are saved for later inspection. This includes raw and cleaned data, total, monthly, daily and interval results for the baseline and analysis period, model performance metrics (such as R2, CVRMSE, NMBE), data and model check results and the baseline model itself. Monthly and overall totals can be computed either with/without extrapolation for missing data depending on the configuration options.

Detailed results are provided in both csv format, various graphs are produced and saved as image files, and a detailed Word® report is automatically generated. Program logging information which includes any warning or error messages is saved to a text file. Timeseries results can optionally be saved to DCH for later analysis. An example Word® report is provided as an appendix to this report.

4.5 Validation

The DCH MVapp has been validated using the Efficiency Valuation Organisations online EVO-portal (39) (46). This consists of 2 years of hourly electricity and temperature data from 367 buildings across North America. One year of data is supplied to code developers who run their M&V algorithm to generate baseline models and predictions for the second year of data (which is not provided to the application or developer). The portal automatically calculates model perform metrics (CVRMSE and NMBE) based on a comparison of the model predictions for the second year of data.

At the time of writing, 93 models had been submitted to the portal with summary performance results as shown in Figure 3. The DCH M&V application achieved a median CVRMSE score of 33.75% and NMBE score of 0.04% placing it in the top 5% of all models and very close to the top performing model. Note that, unlike the requirements listed in M&V standards and guidelines (see Table 7) these performance metrics are calculated on a separate testing data-set, and not on the data used to train the models. Hence, the reported CVRMSE values in particular are inherently higher than those calculated for the fitted models.

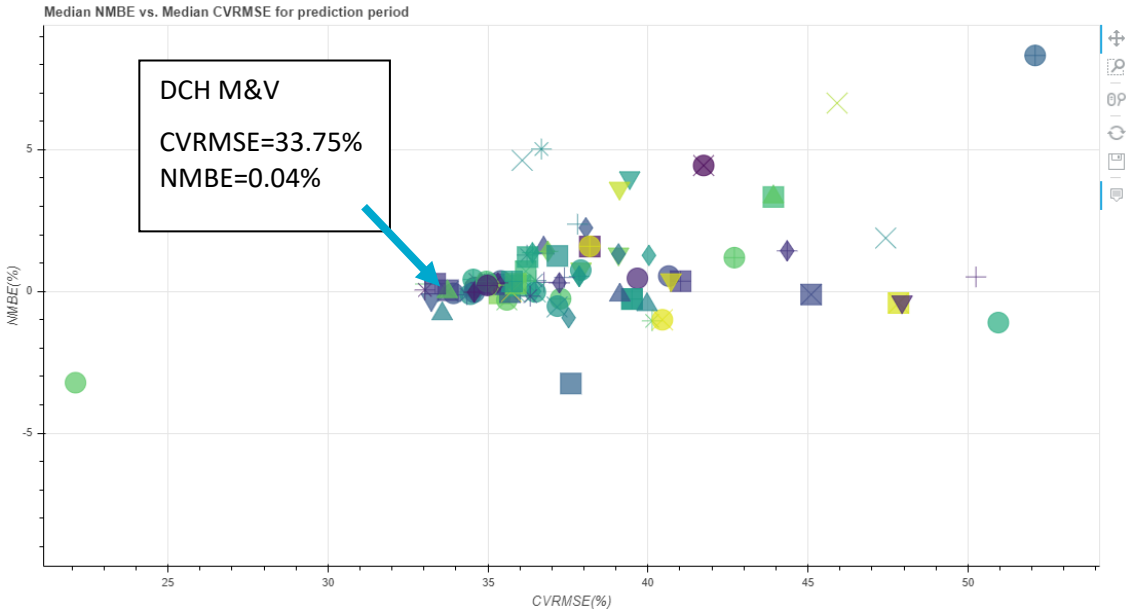

Figure 3 Screenshot of all M&V applications with results shared on the EVO-portal (as of 26th July 2023). DCH-M&V application results as indicated

5 Challenges and limitations

This section discusses in detail each of the current challenges in the path toward implementing autonomous M&V2.0 for energy efficiency applications as identified in Section 1. It includes new analysis conducted specifically for this project. Challenges specially relating to Demand Response applications of M&V2.0 are not considered here.

5.1 Net vs gross metering

Low-transaction cost is a central potential benefit of M&V2.0 and this hinges to a certain extent on the ability to use readily available high-resolution metering data. Utility billing interval meters are the most common form of metering and are generally considered to have good accuracy and reliability (2). However, if on-site generation is present, then this is also likely to be included in the electricity measured by the utility meter (which typically measure net electricity purchased from the grid). M&V on the other hand, is based on an analysis of electricity *consumption*.

References in the literature, guidelines and standards refer exclusively to M&V analysis of energy or electricity *consumption*. Section 12.7 of the most recent version of IPVMP clarifies that: "energy from these [on-site generation] systems will need to be accounted for if they impact the energy consumption, generation or costs within the measurement boundary" (2) which implies the use of additional metering to separately meter the on-site generation. However, in general the literature is not clear on this topic. For example, previous versions on IPMVP did not mention onsite generation, ASHRAE 14 makes no mention of net metering or onsite generation, the NSW PIAM&V guideline (Section 5.5.1) states that the 'addition of on-site generation' should be considered a NRE requiring submetering, though this does not apply to existing generation, and the VIC VEU (35) appears to require separate metering to calculate consumption savings and renewable energy savings for example via the requirement that the "Measurement boundary must include"... "every product co-metered with energy consuming products".

Separate metering of on-site generation is of course possible, but may require installation of one or more additional meters, semantic models to provide meter hierarchies and relationships between meters, and methods to combine and process multiple electricity data streams. Requiring new meter installations is also likely to significantly extend the duration of the M&V project since, by itself, existing utility data would be insufficient to establish a baseline.

For projects where substantial savings are expected this may not be a prohibiting factor, however for smaller sites the cost, time and effort may be unjustified. Moreover if, for example, the overall generation is *small* compared to consumption then a *sufficiently accurate* result may be obtained from analysis of the net meter data alone and hence it might be simply unnecessary to include generation metering and all that it entails. Exactly what constitutes *small* in the context of onsite generation is unknown, and no research in the open literature was found on this question. Given the lack of studies on this topic, here an analysis of the impact of running an M&V2.0 analysis on net electricity data versus consumption data was performed using data from seven commercial (office) sites.

5.1.1 Commercial site analysis

Interval and generation meter data for seven commercial sites spread across regional NSW was used for the analysis. The sites all have onsite PV generation systems with annual generation ranging from 16% to just over

70% of annual consumption. M&V analysis using the DCH MVApp was run with 1 year of baseline data and 1 year of reporting data. The same baseline and reporting periods (baseline 01/03/2021 to 28/02/2022 and reporting 01/03/2022 to 28/02/2023) were selected for each site without any assessment of the presence of non-routine events in the data. No information on the presence (or not) of energy interventions or non-routine events was available.

The presence of generation meters for each site enabled a comparison of two methods;

Method 1: MVapp applied to consumption data using both temperature and time related variables (i.e .Eq. 1 with $\eta = 0$ and Eq. 2 with $\beta_5 = 0$). Consumption was calculated from: consumption = net meter data + generation meter data, and

Method 2: MVapp applied to net metering data directly and with the radiation dependent model term activated.

Analysis was performed using both the interval model and daily models of the MVApp. To examine the effect of varying relative generation amount, additional analyses were also run with sub-hourly measured onsite generation values scaled up/down by fixed factors to simulate different sized PV generation systems.

Model performance was assessed using the commonly used metrics CVRMSE (coefficient of variation of room mean square error) and NMBE (normalised mean bias error). Given the considerable confusion around NMBE and the acceptable limits (47), we focus primarily on CVRMSE <25 as the indicator of model acceptability.

Interval model results

Interval model results are summarised in Tables 2 and 3 with sites listed in order of increasing overall generation amount as a percentage of overall consumption. Five of the seven sites had acceptable baseline models built using the consumption meter data. This decreased to only one of seven sites when the baseline model was trained using net meter data in combination with an additional irradiance term in the model. In terms of the calculated saving over the reporting period (noting that there were no known interventions applied), consumption and net meter model calculated savings overlapped at least partially for all but one site, although mean savings predictions varied considerably.

Table 2 Summary of interval model performance metrics for consumption and net meter analysis models

Site	Annual generation percentage	Consumption model (Model 1)			Net meter model (Model 2)		
		CVRMSE	NMBE	Acceptable	CVRMSE	NMBE	Acceptable
Grafton	15.9	23.2	-0.06	Y	28.5	-0.51	N
Lithgow	18.7	10.7	-0.14	Y	13.9	-0.14	Y
Dubbo	21.5	23.3	-0.11	Y	28.7	-0.06	N
Newcastle	44.1	24.6	-0.17	Y	35.3	-4.71	N
Griffith	51.8	43.8	-1.11	N	64.3	-7.77	N
Armidale	56.7	24.2	-0.21	Y	53.5	3.62	N
Moree	71.3	37.1	-0.05	N	59.0	-8.52	N

Table 3 Summary of interval model calculated reporting period savings for consumption and net meter analysis models

Site	Consumption model (Model 1)	Net meter model (Model 2)	Within bounds
Grafton	4.5% ±1.6%	4.9% ± 1.9%	Y
Lithgow	-2.2% ± 1.8%	-6.0% ± 2.5%	Y
Dubbo	-0.3% ±2.7%	3.8% ± 2.3%	Y
Newcastle	9.7% ± 3.0%	6.1% ± 3.6%	Y
Griffith	-20.3% ±7.0%	3.5% ± 7.6%	N
Armidale	4.6% ±2.4%	6.6% ± 2.9%	Y
Moree	-15.0% ± 5.8%	-5.7% ± 6.5%	Y

For each site, the relative generation amount (ratio of annual generation to annual consumption) was also varied between 0% (no generation) and 90% by scaling the generation meter data appropriately. Figure 4 compares the resultant *increase* in CVRMSE of the net meter-based model (i.e., CVRMSE_model2 – CV_RMSE_model1) versus the generation as a percentage of consumption. Symbols indicate the result corresponding to the actual generation for each site while the shaded region shows the range of values for all sites.

Given that a CVRMSE below 15 is considered a very good model, and a minimum acceptable CVRMSE of 25 is cited in several sources (8) (34), the results suggest that generation values up to 30 to 40% of consumption may be tolerable in an M&V analysis using TOWT regression models with an irradiance dependent term included. However, the consequence of this increase in CVRMSE is a reduction in the predictive power of the model and hence an increase in the minimum saving that can be identified (without the additional solar PV metering) with a given confidence level.

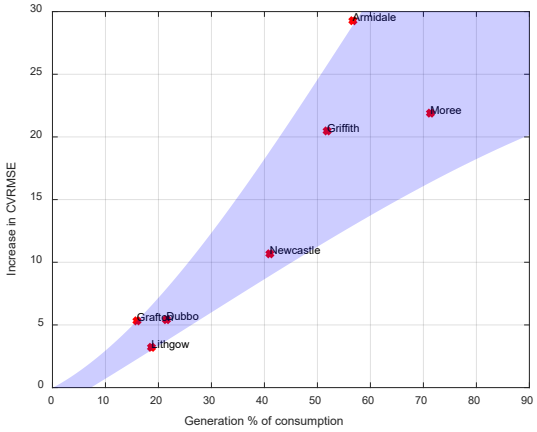

Figure 4 Increase in CVRMSE between baseline consumption interval model and net meter based interval model as a function of relative generation amount

Daily model results

Daily model performance results are summarised in Table 4. For the consumption-based model the same five sites had acceptable models as per the hourly analysis. However, in contrast to the hourly analysis approach, all five of these sites still had acceptable net meter-based models, although the CVRMSE increased in line with the increasing generation (Figure 5). In terms of estimated savings, the consumption and net based model estimates overlapped for all sites, although again there were differences in the mean estimated savings (Table 5). The

increase in CVRMSE (Figure 5) with varying generation amount for each site shows much less effect of increasing generation on model performance.

Table 4 Summary of daily model performance metrics for consumption and net meter analysis models

Site	Annual generation percentage	Consumption model (Model 1)			Net meter model (Model 2)		
		CVRMSE	R2	Acceptable	CVRMSE	R2	Acceptable
Grafton	15.9	13.8	85.3	Y	16.0	85.3	Y
Lithgow	18.7	9.7	62.3	Y	11.5	63.1	Y
Dubbo	21.5	15.1	83.3	Y	17.8	82.8	Y
Newcastle	41.0	16.7	68.8	Y	23.2	62.6	Y
Griffith	51.8	31.7	48.8	N	48.8	33.4	N
Armidale	56.7	15.7	85.3	Y	22.3	86.6	Y
Moree	71.3	31.5	62.6	N	38.7	58.5	N

Table 5 Summary of daily model calculated reporting period savings for consumption and net meter analysis models

Site	Consumption model (Model 1)	Net meter model (Model 2)	Within bounds
Grafton	6.1% ±1.7%	7.0% ± 2.0%	Y
Lithgow	-2.8% ± 2.4%	-6.3% ± 2.5%	Y
Dubbo	1.0% ±2.8%	3.8% ± 2.2%	Y
Newcastle	13.3% ± 2.7%	12.5% ± 2.4%	Y
Griffith	-8.5% ±7.9%	2.0% ± 12.3%	Y
Armidale	8.6% ±2.7%	9.5% ± 2.2%	Y
Moree	-8.0% ± 6.3%	3.1% ± 6.0%	Y

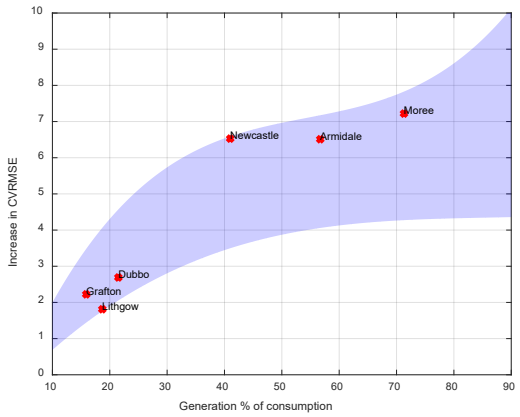

Figure 5 Increase in CVRMSE between baseline consumption daily model and net meter based daily model as a function of relative generation amount

Comparison of interval and daily models for use with net meter data

A comparison of the estimated savings using the hourly consumption, hourly net, daily consumption, and daily net models is given in Figure 6. Note that statically significant changes (either savings or energy increases) are calculated for several of the sites. However, given that no information was available on the presence (or not) of any energy interventions for these sites it is not possible to state whether the estimated savings are the result of real measures, changes to site operation or non-routine events.

Results show that using the daily based model is preferable to the interval model when analysing net meter data as the results for the daily and daily net models tend to be in closer agreement with each other than the hourly and hourly net models with each other. This is likely caused by the approximate nature of the model term used to account for onsite generation as a function of local global irradiance.

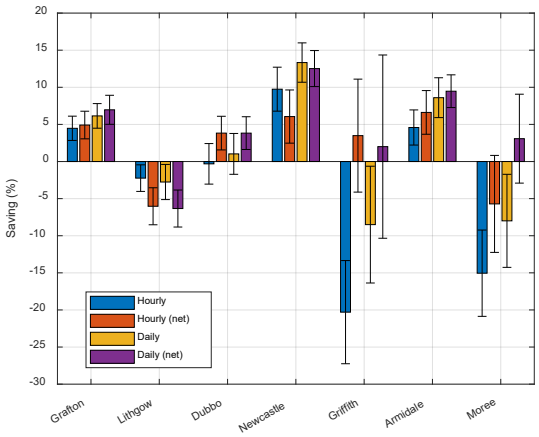

Figure 6 Comparison of estimated percentage savings and uncertainty ranges as computed using hourly, hourly net, daily and daily net models for seven sites

As shown in Figure 7 (left) for one particular location, the daily integral global horizontal irradiance is linearly correlated with the daily integral PV generation as the cumulative effects of localised shading, collector orientation, weather and other unknown effects on generation tends to be smoothed out. However, on an hourly scale (Figure 7 right) much more scatter is present. R-squared coefficients of a linear regression applied to hourly and daily data (Table 6) show this behaviour is consistent across all sites. The practical effect of this scatter is to introduce noise that makes it more difficult for the temperature and time-of-week terms in the model to be fit accurately. Inspection of hourly boxplots of the residuals of regression model fit to the hourly data (Figure 8) shows a trend of under-estimates in the morning and over-estimates in the middle of the day which is likely due to a combination of collector orientation and shading effects. Since the analysis has no access to such information for a given site (i.e., we assume it is not available) it is not possible to employ a detailed PV model. However, additional simulations were trialled assuming standard collector orientation and tilt to calculate irradiance in the plane of the collector. The resulting net meter-based model did not yield any improvement in performance (either model metrics or savings estimates).

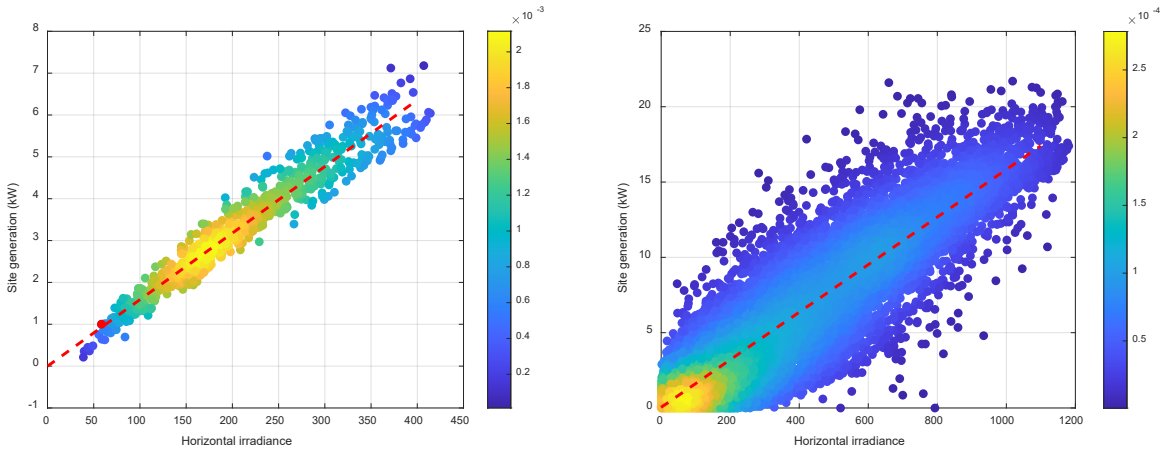

Figure 7 Correlation between daily site generation and daily horizontal irradiance (left) and hourly site generation and hourly horizontal irradiance (right) for Armidale site. Dashed lines show fitted linear regression model with zero intercept

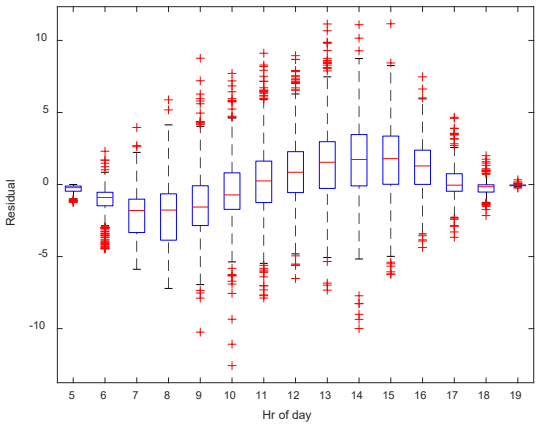

Figure 8 Boxplots of hourly irradiance – generation linear regression residual as a function of hour of the day

Table 6 R-squared values of linear regression models fitting hourly and daily site generation to hourly and daily horizontal irradiance

Site	Hourly	Daily
Grafton	0.870	0.964
Lithgow	0.860	0.900
Dubbo	0.910	0.962
Newcastle	0.830	0.950
Griffith	0.865	0.925
Armidale	0.872	0.938
Moree	0.871	0.935

5.1.2 Summary & recommendations

Although the results above are based on M&V analysis applied to a limited number of sites, they suggest that it is possible to perform M&V using net meter data only and achieve a baseline model that meets commonly acceptable performance metrics. However, it is strongly recommended that:

- If analysis using net meter data is the only option, then a model with daily or longer timescale should be used.
- The baseline model should include a term that models the influence of irradiance (assuming the onsite generation is derived from PV).
- If the overall site generation is more than approximately 30-40% of consumption, additional generation meters should be installed so that the M&V analysis can be based on consumption data.
- The reduction in model predictive power that results from using net meter data should be considered along with the desired percentage savings that are expected.
- If non-PV sources of onsite generation are present, then either additional model terms appropriate for that generation type or additional sub-metering are likely to be required.

Although not considered here, battery storage systems are a special case, and their treatment is likely to depend on how the batteries are operated at the site as well as the scope of the M&V project. For example, a small battery that never or rarely causes significant net export of power to the grid and that operates on a regular schedule of charging and discharging may be able to be ignored (i.e., incorporated into the baseline site load). On the other hand, if for example, the battery operation is subject to significant variability or results in a large quantity of exported power at certain times then additional metering is likely to be required. A further complication is whether the battery operation changes in response to the energy efficiency intervention.

5.2 Automating non-routine events

According to IPMVP (29) non-routine events (NRE) are defined as "unexpected changes in energy use within the measurement boundary resulting from changes in static factors, which are not accounted for in the energy savings calculation and are not related to the targeted energy project". Further "static factors" are defined as "those characteristics of a facility that affect Energy Consumption... that are not expected to change and were therefore not included as Independent Variables", and that "static factors should be recognised and monitored". This definition means that NRE are not simply periods where a model doesn't fit the data well (these may be classed as *potential* NRE). To be considered NRE these periods must be attributed to some unexpected but observed or known changes that occurred at the site.

Ideally NRE are avoided by anticipating such changes and planning for them (for example by including relevant model terms). Managing NRE may involve making 'non-routine adjustments' (NRA's) for example, changing the baselining period or using additional sub-models. In some cases, the decision to make *no* adjustment may be appropriate. Adjustments should only be made for NRE that are confirmed to not be part of normal operation, that are not due to the intervention being evaluated, and that have a significant impact on the estimated savings.

If identification of the cause of NRE is required before adjustments are made, how can automated processes perform this identification without access to much more information (data) relating to the site operation and equipment? As suggested by Touzani et al. (14), the answer is most likely that a 'human-in-the-loop' is still required. However, through a combination of automated analysis and prompts for user input it may be possible to make the processes relatively seamless.

As discussed by Fernandes et al. (11) and Touzani et al. (12), different types of NRE present as different characteristic patterns in the data. Further, they have corresponding NRA's that may be appropriate. For example, an automated M&V application might detect a step-change in energy use over a specific period. It may prompt the user with estimated details of the event such as the magnitude of the change, start and end dates, whether the effect has a time dependence, and the estimated effect on savings estimates. It may then suggest a certain NRA (and require the user to confirm) and/or provide a checklist for the user to run through. The details

of any such choices would presumably be automatically recorded by the M&V application and available for inspection, for example by regulatory authorities.

Such a process does not appear to have been included in any M&V2.0 applications at present, and there are no standardised approaches that can be readily implemented to automate NRA's. However, the IPMVP Guide to Non-routine Adjustments, together with the above cited references, provide a good starting point and there are a number of different analysis approaches that have been documented for identifying NRE's.

5.2.1 Summary & recommendations

- Several methods for automatic detection of *potential* NRE have been documented in the literature. However, because site information that is not available to the M&V application is required to confirm these as actual (valid) NRE, it is likely that authorisation of NRA will still require a 'human-in-the-loop'.
- NRE can be categorised based on their characteristic signature, and targeted technics applied to address them.
- Several references, including the IPMVP, provide good information which could be used to develop a guided M&V2.0 process that handles NRE's. Such a process could use a combination of statistical analysis and simplified prompts to the user or checklists to ensure that the application applies appropriate methods for the site.
- It is suggested that all such inputs provided by users be documented by the M&V application and be available for inspection.

5.3 Difficult to baseline sites

M&V analysis for a site requires determination of a baseline energy consumption model that meets certain performance criteria. These criteria vary according to the particular scheme or guideline as indicated in Table 7. The baseline model must include sufficient independent variables to characterise the variability in consumption resulting from changes to these variables, and the remaining uncertainty or unaccounted for variability in the model, after accounting for non-routine events, must be sufficiently small. For some sites, it may be difficult to develop a baseline model that meets the required performance criteria; these sites are referred to here as 'difficult to baseline sites'.

The underlying principle behind normalisation that is central to most if not all M&V programs is that a reduction in energy consumption is, by itself, insufficient to qualify for incentives. That is, energy savings are not 'counted' if they occur simply because the weather was more mild, fewer occupants were in the building, or less product was produced. For example, the NSW PIAM&V scheme requires that "the ACP must ensure that the implementation does not result in the reduction of energy consumption by reducing[2] production, service, or safety". Examples given for evidence of maintaining service levels include temperature set points meeting comfort levels, lighting levels meeting standards, and maintenance of fan flow rates (34). Similarly in Victoria the

[2] However, energy use reduction through removal of redundant equipment or excessive service levels (e.g. over-cooling in summer, or excessive lighting) is allowed according to Clause 5.4 of the Energy Saving scheme Rule: "Note: Reduced consumption of an Eligible Fuel not directly due to specific actions to improve efficiency does not qualify as a Recognised Energy Saving Activity. Mild weather, lower production, closing down part of a Site, or reducing the quality or quantity of service derived from the use of an Eligible Fuel does not qualify as a Recognised Energy Saving Activity. Reducing consumption of an Eligible Fuel where there is no negative effect on production or service levels (e.g. reduction of excessive lighting, removal of redundant installed capacity or the installation of more energy efficient equipment) is a Recognised Energy Saving Activity and is not excluded by this clause." (54)

VEU states that it is not appropriate to report savings due to "reduce[ing] greenhouse gas emissions by reducing production capacity or service levels, unless this is to correct over-servicing (such as excessive lighting or space heating)" (35). As a result, to demonstrate that the baseline M&V model can capture these variations, the *unexplained variability* must be small as quantified via the various performance criteria.

The most commonly used metric for this purpose is the Coefficient of Variation of the Root Mean Square Error (CVRMSE) which is a normalised measure of mean absolute model error. Although normalised, CVRMSE tends to be smaller for longer time interval models (i.e., daily or monthly as compared to hourly) given the greater fluctuations inherent at shorter timescales. Another criteria used by ASHRAE 14, and the only measure with a quantified recommended threshold in IPMVP, is the relative uncertainty in the estimated annual savings. However, unlike the CVRMSE, relative savings uncertainty is not solely a property of the baseline model but depends also on the post-intervention data. That is, a model with poor CVRMSE could still pass the relative savings uncertainty criteria if the energy savings were large due to aggregation of savings over many observations.

Table 7 Minimum performance criteria

Reference	Performance criteria
NSW PIAM&V (34)	CVRMSE < 25% for R2 >=0.5
	CVRMSE < 10% for R2 < 0.5
	t-statistic > 2 for each independent variable
ASHRAE 14 (8)	CVRMSE < 20% (<12 months data)
	CVRMSE < 25% (>=12 months data to 60 months)
	CVRMSE < 30% (>60 months data)
	Uncertainty < 50% of annual savings at 68% confidence
	NMBE < 0.5%[3]
VIC VEU (35)	Savings discounted if relative precision in annual savings estimate > 25% at 90% confidence
IPMVP (2)	Uncertainty < 50% of annual savings at 68% confidence
Australian Government Best Practise Guide (33)	R2 > 0.75
	CVRMSE < 25% (12 to 60 months data)
	t-statistic > 2 for each independent variable
	NMBE < 0.005%

[3]There is some confusion around this required at discussed by (47). Table 4-2 states <0.005% while Section 4.2.10 states 0.5%.

In addition to the standard model performance metrics (R2, CVRMSE) it is also important to test the standard assumptions used in the M&V model. For example, for ordinary least squares regression-based models this includes independence and normality of errors and homoscedasticity. Often, moderate violations of the standard assumptions can be permitted with minimal effect on the model estimates, however any such tests need to be fully automated to avoid requiring specialist statistical expertise.

For some sites, baseline models fit using standard independent variables such as ambient temperature, time of week, day of week and holiday status leave too much unexplained variability. For example, (15) performed automated M&V analysis on 48,000 buildings with a range of different end-uses and found that only 29% of all

office building models met standard performance criteria. There are likely multiple reasons why so many of the building models failed, including for example the presence of onsite generation, unaccounted for non-routine events and, data and measurement issues. Undoubtedly some of the buildings could not be baselined because they lacked a key independent variable that would explain significant variability. For industrial sites a production related variable may be critical. Some automated M&V tools allow for the inclusion of a generic production variable, though some intervention by the tool user is likely to be required. For other sites the variability may never be able to be plausibly explained in the context of an automated approach. This variability could be due to, for example, occupancy pattern variability which is a function of human behaviours or site operation patterns that are not able to be characterised using the simple model, for example shutdown periods aligned with school holidays or other 'non-standard' periods.

In some cases, variations caused by human behaviours may 'smooth out' in the limit of larger sites with many occupants. For others they may not due to correlated behaviour patterns (for example, many occupants depart a workplace at the same time but with an irregular pattern subject to some external driver). Indeed, the above referenced study found that the percentage of office buildings that had a valid baseline model increased to 68% specifically for large office buildings. As reported by Liang et al. (48), adding a binary occupied/not occupied dependent variable does not usually avoid this problem because; i) time variables already effectively account for occupied status, ii) a binary occupancy indicator gives no information on the number of occupants, and critically iii) the presence or not of occupants does not provide any insight into the energy use behaviours of those occupants.

Residential buildings are likely to be particularly problematic for automated M&V2.0 methods due to greater energy use variability driven by irregular occupant energy use behaviours. Perhaps partially for this reason most M&V tools are geared toward commercial and/or industrial M&V (1), and most studies in the literature have a similar focus. However, it is important to understand the likely applicability of automated M&V methods in residual applications. Thus, in the next section an analysis of the application of M&V to NMI data from 300 residential building is described.

5.3.1 Analysis of residential building baseline models – overall performance

Approximately 3 years of half-hourly electricity meter data from 300 residential dwellings across the Ausgrid network in NSW was obtained from the Solar Homes project (49). Although these homes had solar PV installed, they also had gross metering allowing the solar generation to be excluded from the M&V analysis. In addition, the controlled load circuit (if present) was also separately metered.

Automated M&V analysis was performed using the daily analysis model applied to the consumption data only for cases with and without the controlled load circuit included. Ambient temperature data was obtained from the BOM for the nearest automated weather station based on the supplied dwelling postcode. The baseline models were trained using 1 year of data (1 July 2010 to 30 June 2011) and used to predict for the following year (1 July 2011 to 30 June 2012). No known interventions occurred during the baseline and prediction periods.

Summary baseline model metrics are given in Table 8 for cases with/without the controlled load included. In both cases, only 25% of the sites met the CVRMSE criteria based on the daily analysis model. Overall metrics are marginally better when the controlled loads (CL's) were excluded. Potential non-routine events were identified in two-thirds of sites and with CL only 11.3% of sites passed all model checks.

Table 8 Summary of baseline model metrics from M&V analysis of 300 residential buildings

Metric	Percentage of all sites	
	Controlled load included	Controlled load excluded
CVRMSE < 25	25%	25%
No seasonal trend in residuals	100%	100%
No outlier days in savings	90%	89.3%
No time trend in residuals	65%	69.3%
No potential NRE	32.7%	29.7%
No over-estimated months	18.3%	13.7%
No under-estimated months	12.7%	11.7%
No seasonal trend in savings	11.3%	13.0%

Investigation of the sites with the worst performing baseline models provides further insight. Figure 9 shows the daily consumption data, cumulative distribution of consumption and daily model residuals for the site with the highest (worst) CVRMSE of 89.0. For this site, consumption displays very high irregular peaks close to 10x the mean consumption in summer and winter. This is consistent with a high air-conditioning load that is used intermittently and that is not very well correlated with ambient temperature, binary occupancy status or day of the week (i.e., is likely to be linked to occupant behaviour).

Figure 10 shows the same plots for the site with the second worst CVRMSE (87.2). For this site, there is a period of several months where consumption is near zero, most likely corresponding to the occupants being away. It is also apparent that the periods prior and post have different consumption patterns (both mean consumption and variability). Residential buildings are inherently more prone to these types of changes to consumption as the behaviour of occupants in a single dwelling is generally more variable than those in commercial sites. Using a long-time interval model such as a weekly model may alleviate these issues by averaging across behaviours that can change from day-to-day.

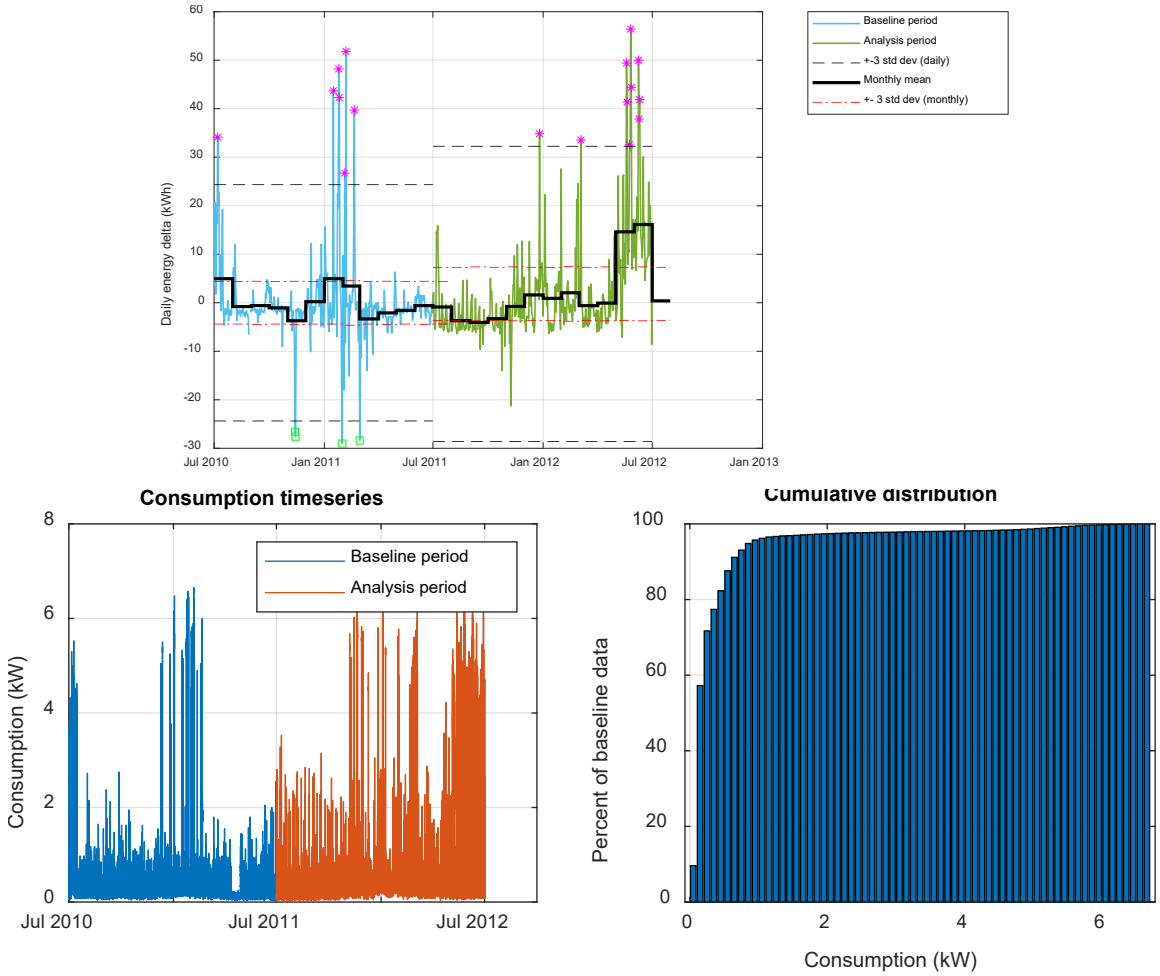

Figure 9 Daily energy residual (top): consumption (bottom left), and cumulative distribution of consumption (bottom right) for the site with the worst performing baseline model

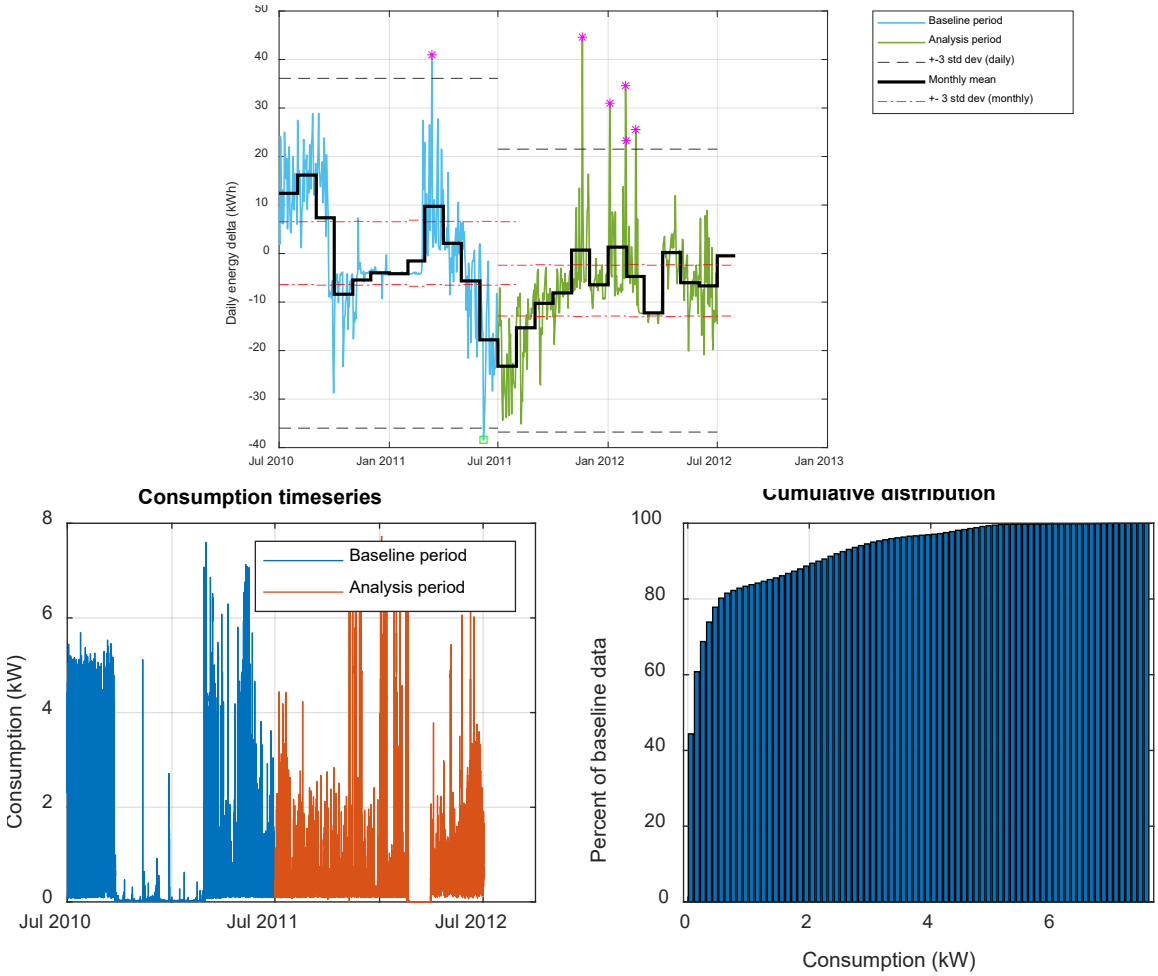

Figure 10 Daily energy residual (top): consumption (bottom left), and cumulative distribution of consumption (bottom right) for the site with the second worst performing baseline model

5.3.2 Identifying difficult to baseline sites quickly

It is important to be able to identify early if standard M&V models will not be appropriate for a particular site, both to avoid unnecessary project costs and unreasonable expectations, and also to provide the opportunity to redesign the M&V approach and potentially incorporate additional independent variable measurements.

Notwithstanding the fact that non-routine events can occur at any time, potentially invalidating a baseline model approach, it is of interest to understand the extent to which models trained with a small quantity of baseline data provide a good estimate of the uncertainty of models trained with a longer period of baseline data from the same site. For example Gallagher et al. (41) compared CVRMSE for models trained with 3, 6, 9 and 12 months of data for a single industrial site and found that the CVMRSE values calculated from the models trained with >=6 months of data were very close to those calculated from the model trained with 12 months of data.

He we expand this approach and consider a random selection of 100 of the residential sites used in the previous analysis. The interval analysis model was re-trained progressively using between 1 and 12 months of data and the resultant CVRMSE compared with that from the final model trained using 12 months of data. Here the CVRMSE was evaluated for the training data as opposed to a different testing data set, since we assume that in practice, only the training data would be available to perform the evaluation in real-time.

Results are summarised in Figure 11. Here contour lines show the fraction of the 100 buildings where the absolute difference in computed CVRMSE is less than a given amount for the model trained with a certain number of

baseline data days. For example, the red cross indicates that for 70% of sites the baseline models trained with 90 days of data have a CVRMSE that is <+5% more than the CVRMSE of the models trained with 365 days of data. Differences are skewed toward positive values corresponding to the CVRMSE generally decreasing as the amount of training data increases. These results suggest that, in general, only a relatively small amount of data is required to identify whether a site baseline model is *likely* to meet performance criteria (though of course site energy use could still substantially change at any time in the future).

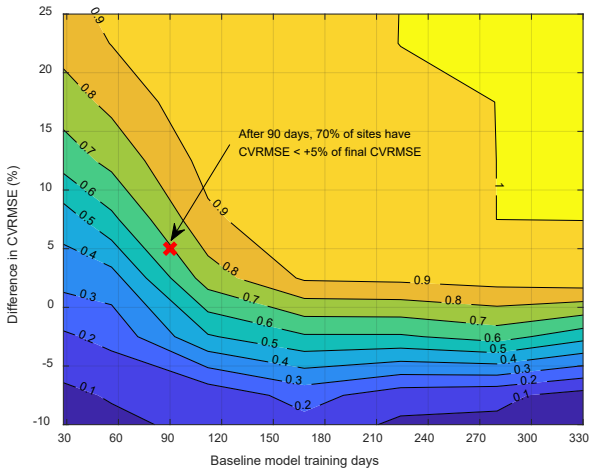

Figure 11 Difference in CVRMSE for model trained with fewer days compared to model trained with 365 days. Contours show cumulative fraction of buildings with difference below a given level

5.3.3 Summary & recommendations

- A significant portion of sites may not be able to be baselined using automated M&V tools, particularly with high resolution (i.e., hourly or daily) models. This is more likely for sites where behavioural effects are present, for example residential buildings, but may also occur for sites where other (unaccounted for) variables play an important role (for example industrial sites with production related energy use).
- Schemes designed around automated M&V workflows should enable rapid/low-cost determination of whether or not a site can be baselined (i.e., 'fail fast') to avoid unreasonable expectations and/or wasted resources.
- M&V baseline models can be re-trained periodically with smaller amounts of data and the standard performance metrics used to identify difficult to baseline sites early in the process.
- Including a production related independent variable to the standard model formulation is a simple modification that would add flexibility particularly for industrial M&V applications.

5.4 Assessing model accuracy

Calculating uncertainty in computed savings estimates is a critical part of an M&V analysis and relies on being able to quantify the uncertainty in the baseline model. Typically, model uncertainty is assumed to be much greater than the uncertainty associated with measured quantities such that the latter are ignored (7).

If the uncertainty range for the calculated savings includes zero, then it is not possible to state that there is any saving at all at the given confidence level. The importance of characterising uncertainty was underscored by Gallagher et al. (50) who applied a range of different models (linear, non-linear, machine learning) and analysis intervals (15min, hourly, daily, monthly) to industrial building M&V. They found wildly differing estimates of mean

savings, though, thanks to the relatively wide uncertainty ranges, most of the estimated uncertainty ranges overlapped.

Different methods are used to quantify model uncertainty including standard error estimates from the regression models, the ASHRAE 14 fractional saving uncertainty approach, cross-validation and boot-strapping. One of the key difficulties for hourly and daily models in particular is that the standard OLS (ordinary least squares) error estimates are biased due to autocorrelation of the residuals. That is, as sequential energy estimates are correlated, so too tend to be their errors. This issue has been explored by several authors (51) (50) (14) and is stated to result in the true uncertainty being under-estimated by standard methods. The ASHRAE 14 guideline includes a correlation factor to account for auto-correlation in the residuals, however Koran et al. (51) state that this correlation factor may over-estimate the influence of autocorrelations.

One approach to uncertainty quantification proposed by several authors is k-fold cross-validation (CV). As outlined by Travis et al. (52), this consists of dividing the baseline period into *k* 'folds' (typically 5 to 10) with the model trained on *k-1* folds of data and the remaining data used to evaluate the model error. This process is repeated *k* times and the resulting error estimates averaged. Touzani et al. (14) compared CV with the ASHRAE error formation (with adjustment for auto-correlation) and found that while both the CV and ASHRAE methods tended to under-estimated uncertainty, the ASHRAE approach was closer to the true uncertainty (the true value was within the uncertainty range 71% of the time where the expected percentage was 95% based on the chosen confidence level). Granderson and Price (53) used cross validation to assess performance of five different models across 29 sites. They report that baseline model uncertainty does not necessarily decrease with a longer training period because the building's energy behaviour changes from week to week.

Another approach is bootstrapping. As outlined by Koran et al. (51) this consists of repeatedly resampling the data (with replacement), computing the key statistic on the sampled data, and then evaluating the variance across the samples. As for CV, bootstrapping has been applied with mixed success. For example, Koran et al. found that the simpler OLS and ASHRAE error estimates were close to those from three different bootstrap-based approaches for a particular example case with real data but were significantly different for several synthetic data cases. Variations to the standard boot-strap approach were described to account for autocorrelations and independent variable relations but were more complicated to implement.

5.4.1 Summary & recommendations

- Calculation of uncertainty bounds for savings estimates in M&V analysis is essential to establish whether or not savings are likely to be real or a model artifact.
- Model accuracy does not necessarily increase with longer training data. A decreasing model accuracy with longer training period may indicate that the site's baseline energy use is changing, that there are unresolved NRE's, or that the model does not capture the site's energy use variation accurately.
- Auto-correlation of residuals is likely to have an increasing influence on computed uncertainty bounds as the base time-interval of the model reduces. Methods to account for this correlation are approximate but are important to include. Because of the danger of over-fitting and the difficulty of computing uncertainty bounds, it may be useful to compare estimates from models employing a range of base time-intervals.
- Several approaches can be used including ASHRAE formula, cross-validation and bootstrapping. Studies in the literature are mixed when it comes to the relative accuracy of the computed uncertainty bounds computed using these methods. This suggests the simpler approaches (e.g., ASHRAE method) may be preferable in the medium term.

6 Conclusion

The value proposition offered by Advanced M&V or M&V2.0 based algorithms is centred on the union of data gathering, data cleaning & processing, model construction, savings estimation, and uncertainty quantification stages of the M&V processes in a single workflow or tool that can be re-run on-demand at any stage of the process. The ability to offer additional insights from the high resolution, interval metering data is a 'nice-to-have' but is not the central benefit, and indeed in some cases more reliable savings estimates may be derived from models based on integrated or aggregated interval metering data.

Although further work is needed to automatically handle cases where data, or models deviate from the expected behaviour (for example handling of non-routine events and difficult to baseline sites) existing approaches are available and sufficient to identify when these deviations occur, even if methods have not yet been developed to automatically handle them.

The detailed design of such automated of semi-guided processes is closely linked to the requirements of the scheme, regulation or user driving the particular M&V implementation. Hence, it is suggested that the next stage of work focus on developing detailed 'user-stories' and workflows that define who will use the M&V2.0 tool, how they will use/interact with it, what inputs can be supplied and what it should produce or output. These inputs can be used to direct algorithm development efforts, which can in turn feed back into the user-centric design process in an iterative manner.

References

1. *The state of advanced measurement and verification technology and industry application.* **Granderson, J. and Fernandes, S.** s.l. : Berkeley Lab, 2017, The Electricity Journal, Vol. 30, pp. 8-16.

2. **EVO.** *International Performance Measurement and Verification Protocol: Core Concepts.* s.l. : Efficiency Valuation Organization, 2022. EV 10000 - 1:2022.

3. **IPART.** *Metered Baseline Method Guide v2.6.* Sydney : Independent Pricing and Regulatory Tribunal of NSW, 2023.

4. —. *Project Impact Assessment with Measurement and Verification Method Guide v5.0.* Sydney : Independent Pricing and Regulatory Tribunal of NSW, 2023.

5. **Common Capital.** *Calculating Savings using Measurement and Verification.* Melbourne : Victorian Department of Environment, Land, Water and Planning, 2017.

6. **Government of South Australia.** Retailer Energy Productivity Scheme. [Online] Essential Services Commission of South Australia. [Cited: 18 7 2023.] https://www.escosa.sa.gov.au/industry/reps/overview/reps.

7. **EVO.** *Uncertainty assessment for IPMVP.* s.l. : Efficiency Valuation Organization, 2019. EVO 10100 - 1:2019.

8. **ASHRAE.** *Guideline 14-2014 Measurement of energy, demand and water savings.* Atlanta : ASHRAE, 2014.

9. **Webster, L.** *IPMVP's snapshot on advanced measurement & verification.* s.l. : Efficiency Valuation Organisation, 2020.

10. *From Theory to Practice: Lessons Learned from an Advanced M&V Commercial Pilot.* **Crowe, E., Granderson, J. and Fernandes, S.** Denver : IEPEC, 2019. International Energy Program Evaluation Conference.

11. **Fernandes, S., et al.** *Detecting the undetected: Dealing with non-routine events using advanced M&V meter-based savings approaches.* s.l. : Lawrence Berkeley National Laboratory, 2020.

12. *Statistical change detection of building energy consumption: Applications to savings estimation.* **Touzani, S., et al.** s.l. : Elsevier, 2019, Energy and Buildings, Vol. 185, pp. 123-136.

13. *Application of automated measurement and verification to utility energy efficiency program data.* **Granderson, J., et al.** 1, s.l. : Elsevier, 2017, Energy and Buildings, Vol. 142.

14. *Evaluation of methods to assess the uncertainty in estimated energy savings.* **Touzani, S., et al.** s.l. : Elsevier, 2019, Energy and buildings, Vol. 193, pp. 216-225.

15. **LBNL.** *Sacramento Municipal Utility District Explores Advanced M&V Capabilities.* Berkeley : Lawrence Berkeley National Laboratory, 2018.

16. **Satchwell, A., Piette, M. and Khandekar, A.** *A national roadmap for grid-interactive efficient buildings.* s.l. : U.S. Department of Energy, 2021.

17. **Price, P., Addy, N. and Kiliccote, S.** *Predictability and persistance of demand response load shed in buildings.* s.l. : Berkeley National Laboratory, 2015.

18. **KEMA.** *Development of demand response mechanism: baseline consumption methodology.* s.l. : Australian Energy Market Operator, 2013.

19. *Assessment of model-based peak electric consumption prediction for commercial buildings.* **Granderson, J., et al.** s.l. : Elsevier, 2021, Energy and buildings, Vol. 245.

20. *Statistical analysis of baseline load models for residential buildings in the context of winter demand response.* **Poulin, A., Leduc, M. and Fournier, M.** s.l. : MDPI, 2022, Energies, Vol. 15.

21. *Data-driven baseline estimation of residential buildings for demand response.* **Park, S., et al.** s.l. : Energies, 2015, Vol. 8.

22. **AEMO.** *Wholesale demand response guidelines.* s.l. : Australian Energy Market Operator, 2021.

23. —. *Baseline eligibility compliance and metrics policy.* s.l. : Australian Energy Market Operator, 2021.

24. —. *Guide to the WDRM predictability of load (PoL) calculator.* s.l. : Australian Energy Market Operator, 2021.

25. —. *Electricity rule change proposal: Scheduled Lite.* s.l. : Australian Energy Market Operator, 2023.

26. *Spatio-temporal impacts of a utility's efficiency portfolio on the distribution grid.* **Granderson, J., et al.** s.l. : Elsevier, 2020, Energy, Vol. 212.

27. *Automated measurement and verification: performance of public doman whole-building electric baseline models.* **Granderson, J., et al.** s.l. : Elsevier, 2015, Applied Energy, Vol. 144.

28. *Decarbonization of electricity requires market-based demand flexibility.* **Golden, M., Scheer, A. and Best, C.** s.l. : Elsevier, 2019, The Electricity Journal, Vol. 32.

29. **EVO.** *IPMVP application guide on non-routine events & adjustments.* s.l. : Efficiency Valuation Organisation, 2020. EVO 10400 - 1:2020.

30. —. *Renewables application guide.* s.l. : Efficiency Valuation Organisation, 2017. EVO 10200 - 1:2017.

31. **US DOE.** *M&V guidelines: measurement and verification for performance-based contracts Version 4.0.* s.l. : Department of Energy, 2015.

32. **NREL.** *The Uniform Methods Project: methods for determining energy efficiency savings for specific measures.* Golden : National Renewable Energy Laboratory, 2018.

33. **Australasian Energy Performance Contracting Association.** *A best practise guide to measurement and verification of energy savings.* s.l. : Commonwealth of Australia, 2004.

34. **NSW Government.** *PIAM&V method application requirements for non-routine events and adjustments.* s.l. : NSW Government, 2022.

35. **Essential Services Commision.** *Measurement and verification method activity guide.* s.l. : Essential Services Commision, 2021.

36. *A review of deterministic and data-driven methods to quantify energy efficiency savings and to predict retrofitting scenarios in buildings.* **Grillone, B., et al.** s.l. : Elsevier, 2020, Renewable and Sustainable Energy Reviews, Vol. 131.

37. *A review of data-driven approaches for measurement and verification analysis of building energy retrofits.* **Alrobaie, A. and Krarti, M.** s.l. : MDPI, 2022, Energies, Vol. 15.

38. *Measurement and verification of energy conservation measures using whole-building electricity data from four identical office towers.* **Newsham, G.** s.l. : Elsevier, 2019, Applied Energy, Vol. 255.

39. *Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings.* **Granderson, J., et al.** s.l. : Elsevier, 2016, Applied Energy, Vol. 173, pp. 296-308.

40. *Gradient boosting machine for modeling the energy consumption of commercial buildings.* **Touzani, S., Granderson, J, and Fernandes, S.** s.l. : Lawrence Berkeley National Laboratory, 2018.

41. *The suitability of machine learning to minimise uncertainty in the measurement and verification of energy savings.* **Gallagher, C., et al.** s.l. : Elservier, 2018, Energy and Buildings, Vol. 158.

42. *Analysis of measurement and verification methods for energy retrofits applied to residential buildings.* **Guiterman, T. and Krarti, M.** s.l. : ASHRAE, 2011, ASHRAE Transactions, pp. 382-394.

43. *Alexa, which M&V method should I use? Data-Driven method selection for savings calculation.* **Metoyer, J., et al.** s.l. : ACEEE, 2018. ACEEE Summer study on energy efficiency in buildings.

44. *Cloud computing platform for real-time measurement and verification of energy performance.* **Ke, M., Yeh, C. and Su, C.** s.l. : Elsevier, 2017, Applied Energy, Vol. 188.

45. *Air-conditioning demand response resource assessment for Australia.* **Goldsworthy, M. and Sethuvenkatraman, S.** 8, s.l. : Taylor & Francis, 2020, Science and Technology for the Built Environment, Vol. 26.

46. **EVO.** Advanced M&V testing portal. *EVO-world.* [Online] Efficiency Valuation Organisation. [Cited: 26 7 2023.] https://mvportal.evo-world.org/.

47. *Validation of calibrated models: common errors.* **Ruiz, G. and Bandera, C.** s.l. : MDPI, 2017, Energies.

48. *Improving the accuracy of energy baseline models for commercial buildings with occupancy data.* **Liang, X., Hong, T. and Shen, G.** s.l. : Elsevier, 2016, Applied Energy, Vol. 179, pp. 247-260.

49. **Ausgrid.** Solar home electricity data. [Online] [Cited: 15 7 2023.] https://www.ausgrid.com.au/Industry/Our-Research/Data-to-share/Solar-home-electricity-data.

50. *Development and application of a machine learning supported methodology for measurement and verification (M&V) 2.0.* **Gallagher, C., et al.** s.l. : Elsevier, 2018, Energy and Buildings, Vol. 167, pp. 8-22.

51. *A comparison of approaches to estimating the time-aggregated uncertainty of savings estimated from meter data.* **Koran, B., et al.** Baltimore : s.n., 2017. International Energy Program Evaluation Conference.

52. **Travis, W., Price, P. and Sohn, M.** *Uncertainty estimation improves energy measurement and verification procedures.* s.l. : Lawrence Berkeley National Laboratory, 2014.

53. **Granderson, J. and Price, P.** *Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models.* s.l. : Lawrence Berkerley Laborartory, 2013.

54. **NSW Government.** Energy Savings Scheme Amendment No.1 Rule 2023. [Online] [Cited: 18 8 2023.] https://www.energy.nsw.gov.au/sites/default/files/2023-01/202301-Energy-Savings-Scheme-Amendment-No.1-Rule-2023.pdf.

55. *Statistical change detection of building energy consumption: applications to savings estimation.* **Touzani, S., et al.** s.l. : Elsevier, 2019, Energy and Buildings, Vol. 185, pp. 123-136.

As Australia's national science agency and innovation catalyst, CSIRO is solving the greatest challenges through innovative science and technology.

CSIRO. Unlocking a better future for everyone.

Contact us

1300 363 400
+61 3 9545 2176
csiro.au/contact
csiro.au

For further information

Energy
Mark Goldsworthy
+61 2 4960 6112
mark.goldsworthy@csiro.au
csiro.au/energy