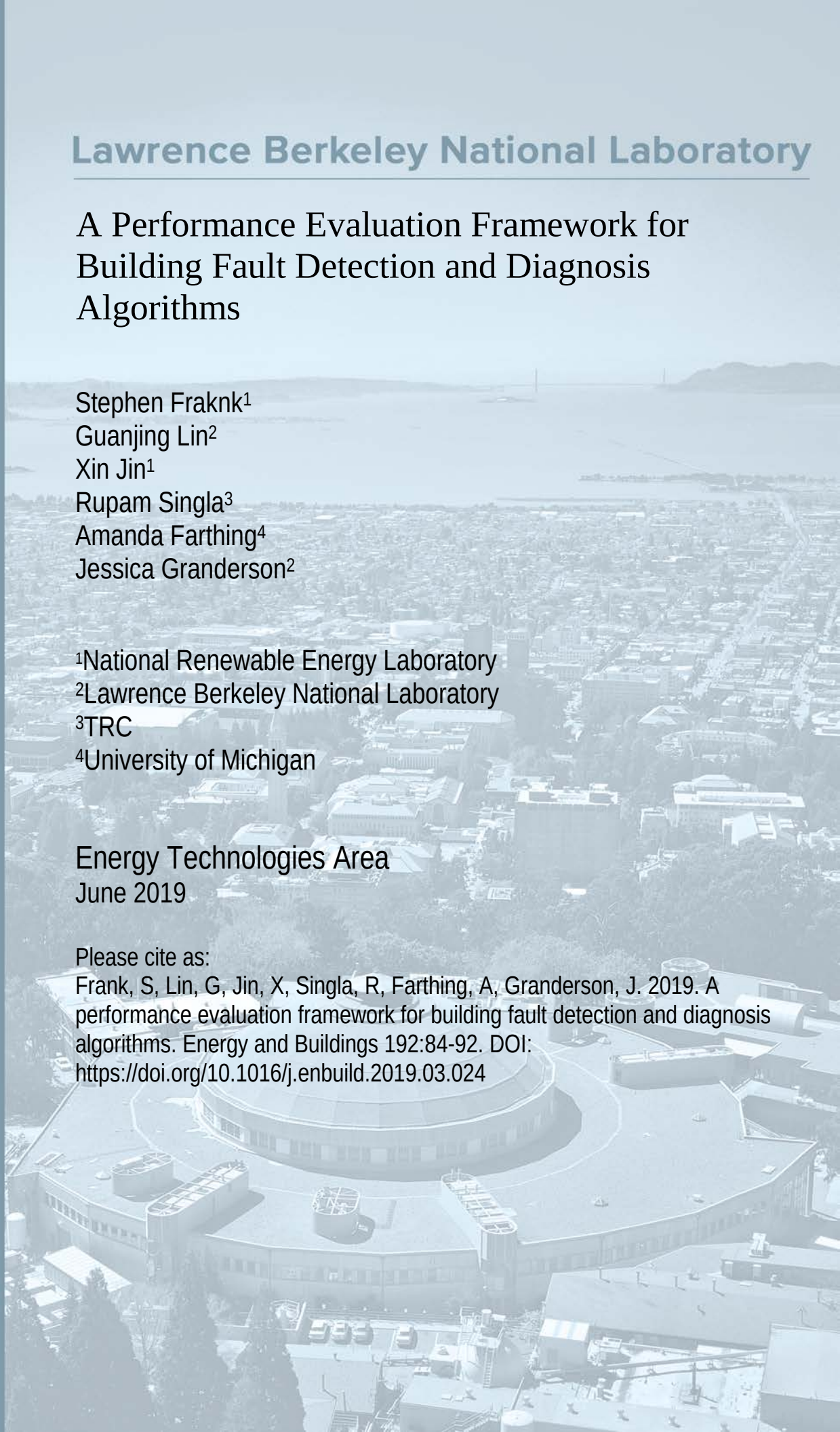


Lawrence Berkeley National Laboratory

A Performance Evaluation Framework for Building Fault Detection and Diagnosis Algorithms

Stephen Fraknk[1]
Guanjing Lin[2]
Xin Jin[1]
Rupam Singla[3]
Amanda Farthing[4]
Jessica Granderson[2]

[1]National Renewable Energy Laboratory
[2]Lawrence Berkeley National Laboratory
[3]TRC
[4]University of Michigan

Energy Technologies Area
June 2019

Please cite as:
Frank, S, Lin, G, Jin, X, Singla, R, Farthing, A, Granderson, J. 2019. A performance evaluation framework for building fault detection and diagnosis algorithms. Energy and Buildings 192:84-92. DOI: https://doi.org/10.1016/j.enbuild.2019.03.024

DISCLAIMER

COPYRIGHT NOTICE

A Performance Evaluation Framework for Building Fault Detection and Diagnosis Algorithms☆

Stephen Frank[a,*], Guanjing Lin[b], Xin Jin[a], Rupam Singla[c], Amanda Farthing[d], Jessica Granderson[b]

[a] *National Renewable Energy Laboratory, Golden, CO, USA*
[b] *Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[c] *TRC, Oakland, CA, USA*
[d] *University of Michigan, Ann Arbor, MI, USA*

Abstract

Fault detection and diagnosis (FDD) algorithms for building systems and equipment represent one of the most active areas of research and commercial product development in the buildings industry. However, far more effort has gone into developing these algorithms than into assessing their performance. As a result, considerable uncertainties remain regarding the accuracy and effectiveness of both research-grade FDD algorithms and commercial products—a state of affairs that has hindered the broad adoption of FDD tools. This article presents a general, systematic framework for evaluating the performance of FDD algorithms. The article focuses on understanding the possible answers to two key questions: in the context of FDD algorithm evaluation, what defines a fault and what defines an evaluation input sample? The answers to these questions, together with appropriate performance metrics, may be used to fully specify evaluation procedures for FDD algorithms.

Keywords: Fault detection and diagnosis, performance evaluation, algorithm testing, benchmarking, building systems, building energy performance

2010 MSC: 62H30, 62P30

Declaration of Interests: None

1. Introduction

Faults and operational inefficiencies are pervasive in today's commercial buildings [1–3]. Fault detection and diagnosis (FDD) tools use building operational data to identify the presence of faults and isolate their root causes. Widespread adoption of such tools and correction of the faults they identify would deliver an

☆This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308, and by Lawrence Berkeley National Laboratory, operated for the DOE under Contract No. DE-AC02-05CH11231. Funding was provided by the DOE Assistant Secretary for Energy Efficiency and Renewable Energy Building Technologies Office Emerging Technologies Program. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government.

*Corresponding author
Email address: Stephen.Frank@nrel.gov (Stephen Frank)

estimated 5%–15% energy savings across the commercial buildings sector [1, 4]. In the United States, this opportunity represents 260–790 TWh (0.9–2.7 quadrillion Btu) of primary energy, or approximately a 2% reduction in national primary energy consumption [5, 6].

Fault detection is a process of detecting faulty behavior and fault diagnosis is a process of isolating the cause(s) of the fault after it has been detected. Fault detection and diagnosis are sometimes performed separately but are often combined in a single step. In the last three decades, the development of automated fault detection and diagnosis (AFDD) methods for building heating, ventilation, and air conditioning (HVAC) and control systems has been an area of active research. Two International Energy Agency (IEA) Annex Reports [7, 8] and literature reviews by Katipamula and Brambley [9, 10], Katipamula [2], and Kim and Katipamula [11] are the major review publications in the HVAC FDD area.

Kim and Katipamula [11] indicate that since 2004, more than 100 FDD research studies associated with building systems have been published. A great diversity of techniques are used for FDD, including physical models [12, 13], black box [14, 15], grey box [16, 17], and rule-based approaches [18, 19]. Commercial AFDD software products represent one of the fastest growing and most competitive market segments in technologies for building analytics. There are dozens of AFDD products for buildings now available in the United States, and new products continue to enter the market [20, 21]. However, considerable debate continues and uncertainties remain regarding the accuracy and effectiveness of both research-grade FDD algorithms and commercial AFDD products—a state of affairs that has hindered the broad adoption of AFDD tools.

Far more effort has gone into developing FDD algorithms than into assessing their performance. Indeed, there is no generally accepted standard for evaluating FDD algorithms. There is an urgent need to develop a broadly applicable evaluation procedure for existing and next-generation FDD tools. Such a procedure would provide a trusted, standard method for validation and comparison of FDD tools at all stages of development, from early-stage research to mature commercial products. Given the wide variety of FDD use cases and competing techniques, establishing a standard evaluation methodology is a daunting challenge [22, 23]. Significant progress has been made in establishing FDD test procedures and metrics within both the buildings sector [24, 25] and other industries [26, 27]. Nevertheless, existing approaches to evaluation differ significantly with respect to specific evaluation parameters within a given general methodology and how these choices impact evaluation results.

Therefore, this article describes a general, systematic framework for evaluating the performance of FDD algorithms that leverages and unifies prior work in FDD evaluation and incorporates insights from interviews with industry experts. Section 2 provides a brief summary of relevant prior work. Section 3 then outlines the process required to evaluate an FDD algorithm. Sections 4 and 5 examine two critical questions that must be answered to apply this evaluation process:

1. What defines a fault?
2. What defines an evaluation input sample?

Section 6 provides a brief introduction to FDD evaluation outcomes in the context of performance metrics. Finally, Section 7 discusses these findings in light of key considerations for FDD algorithm performance evaluation and Section 8 concludes with recommendations and suggested areas of future research.

2. Background

To assess the state of the art in FDD evaluation, we reviewed articles, book chapters, and technical reports related to FDD evaluation in five industries: buildings, aerospace, power systems, manufacturing, and process control. In the buildings sector, IEA Annex 34 technical report [8] provides a broad overview of early development and evaluation of FDD algorithms for HVAC systems and equipment. In the report, House et al. [28] notes the need for systematic performance evaluation of FDD algorithms. The report presents examples of several such evaluations, including detailed descriptions of the experimental procedures. However, the report does not provide a similarly comprehensive description of the evaluation framework or performance metrics. Although some FDD research contemporaneous with the report does provide detailed analysis of algorithm performance [29, 30], the evaluation methods and results are not presented in a way that facilitates comparison of results among disparate evaluation efforts.

Building on the Annex 34 work, Reddy [24, 31] and Yuill and Braun [23, 25, 32, 33] have contributed significantly to the development of FDD evaluation methodologies for chillers and unitary equipment, respectively. Reddy [24] describes FDD algorithm performance evaluation as one component of a broader evaluation methodology that examines FDD tools' performance, cost, ease of implementation, ease of use, data requirements, training requirements, and applicability to the needs of a particular site or customer. The author catalogs possible raw evaluation outcomes (see Section 6) and associated performance metrics.

Yuill and Braun [25] incorporate the evaluation outcomes described in [24] into a general FDD evaluation approach that includes an evaluation workflow, a description of evaluation metrics, and a discussion of establishing ground truth by means of defining a fault impact threshold (see Section 4.3). This general methodology is expanded in [32] and forms the foundation for the present work.

In the power systems sector, Kurtoglu et al. [26] present an FDD evaluation workflow that largely parallels that of Yuill and Braun [25], but with greater emphasis on temporal performance metrics (see Section 6). SAE Aerospace Recommended Practice ARP5783 [27] provides a highly detailed methodology for evaluating aircraft fault detection tools. Literature in other industries focuses largely on mathematical treatments of proposed FDD performance metrics [34, 35].

Shortcomings common (although not universal) in the literature reviewed include:

- Inconsistent, conflicting, or unclear explanation of the method(s) for assigning ground truth in scenarios used for FDD algorithm evaluation
- Lack of clear or rigorous definition of input samples used for FDD algorithm evaluation
- Lack of rigorous mathematical definitions for performance metrics reported
- No formal treatment of the substantial differences in evaluation approach found in the existing literature.

The present work addresses these topics.

3. Methodology

The objective of the research was to develop a general and practical performance evaluation framework for FDD algorithms by synthesizing prior research with industry domain expertise. The elements of the framework are drawn from the technical literature and from interviews conducted with six FDD experts in the buildings industry. Our intended audience is the buildings industry; however, the principles outlined are broadly applicable and inform FDD evaluation methodologies for other industries.

3.1. Problem Statement

The purpose of an FDD algorithm is to determine whether building systems and equipment are operating improperly (fault detection) and, in the case of abnormal or improper operation, to isolate the root cause (fault diagnosis). The purpose of FDD performance evaluation is to quantify how well an FDD algorithm performs these two tasks. Achieving a credible outcome from FDD performance evaluation requires adherence to a clear and well-designed evaluation procedure. The purpose of the general evaluation framework presented in this article is to provide a rigorous foundation upon which such FDD evaluation procedures may be constructed. The framework is therefore descriptive rather than prescriptive; we outline the process required to evaluate an FDD algorithm and we document the choices faced by an FDD evaluator.

3.2. General Performance Evaluation Framework

With the procedure of Yuill and Braun [25] as a starting point, Figure 1 presents a general FDD performance evaluation framework consisting of six components or steps:

1. Determine a set of **input scenarios**, which define the driving conditions, fault types, and fault intensities (fault severity with respect to measurable quantities).
2. Create a set of **input samples** drawn from the input scenarios, each of which is a test data set for which the performance evaluation will produce a single outcome.

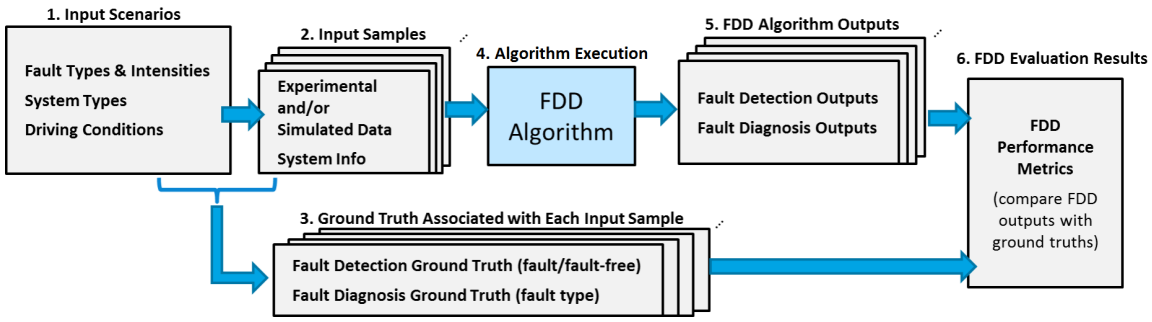


Figure 1: FDD performance evaluation framework (expanded and generalized from Yuill and Braun [25, Figure 1])

3. Assign **ground truth** information associated with each input sample.
4. **Execute the FDD algorithm** that is being evaluated for each input sample. The FDD algorithm receives input samples and produces fault detection and fault diagnosis outputs.
5. Retrieve FDD algorithm **fault detection and fault diagnosis outputs**.
6. Evaluate FDD **performance metrics**. First, raw outcomes are generated by comparing the FDD algorithm output and the ground truth information for each sample. Then, the raw outcomes are aggregated to produce performance metrics.

Steps 1, 2, 4, and 5 are original to the evaluation procedure presented by Yuill and Braun [25], while steps 3 and 6 are novel.

3.2.1. Input Scenarios

Each input scenario defines a test case consisting of one or more input samples. Input scenarios may specify [24, 25]:

- Building types and characteristics (age, size, use patterns, etc.)
- Equipment types
- Faults types, intensities, and prevalence
- Environmental conditions
- Data available to the FDD algorithm (*e.g.*, from sensors, meters, or a control system)
- Cost data (if applicable for calculating performance metrics).

3.2.2. Input Samples

Input samples are drawn from the input scenarios that make up the AFDD evaluation data set. Each input sample is a collection of data for which the AFDD performance evaluation should produce a single

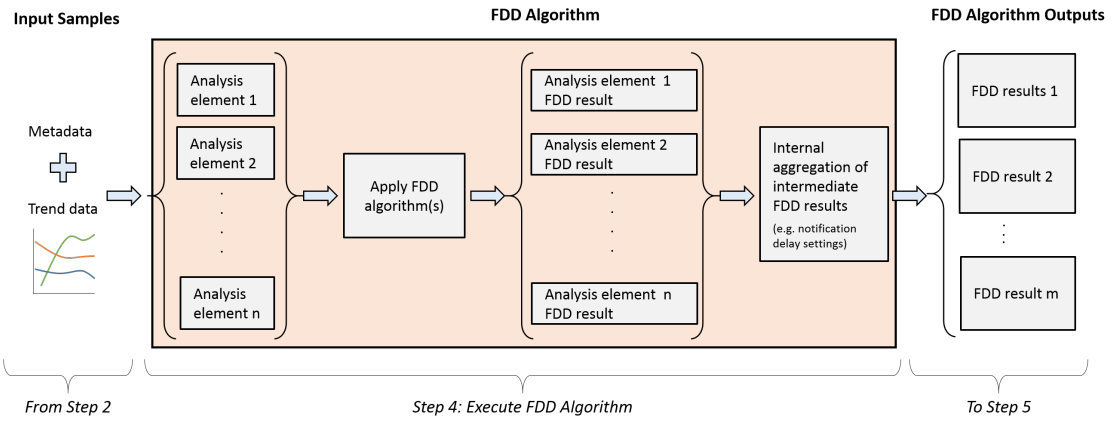


Figure 2: Workflow for executing an FDD algorithm during performance evaluation, depicting Step 4 from Figure 1, with connections to Steps 2 and 5

result (6). Input samples may include system information (metadata) and time series trend data from building sensors and control systems.

3.2.3. Ground Truth

In order to evaluate whether the output of an AFDD algorithm is correct for a given input sample, it is first necessary to establish the state of the system represented by that sample: faulted or unfaulted, and, if faulted, which fault cause(s) are present. In this step, each input sample generated in Step 2 is assigned a ground truth state.

3.2.4. Algorithm Execution

In this step, the FDD algorithm is first initialized and then executed for each input sample. Initialization may include input of metadata specific to the selected input scenario(s), supervised learning using a training data set (input samples labeled with ground truth), or tuning (parameter adjustment) to adjust the algorithm's sensitivity.

Figure 2 illustrates the workflow by which an FDD algorithm converts input samples (Step 2) into algorithm outputs (Step 5). Within the FDD algorithm, the input sample is preprocessed into one or several analysis elements required. (For example, Ferretti et al. [36] preprocess 1-minute interval data into an hourly average). The FDD algorithm analyzes each analysis element, yielding intermediate FDD results. The FDD algorithm then aggregates these intermediate results across all the analysis elements to produce detection and diagnosis results that are displayed to the user (Step 5).

3.2.5. Algorithm Outputs

For each input sample, the FDD algorithm is expected to produce a detection result that indicates whether a fault is present as well as a diagnosis result that presents further information about the precise

nature or root cause of the fault. Together, the detection and diagnosis results yield a single output for use in Step 6.

3.2.6. Evaluation Results and Performance Metrics

Evaluation results are generated by comparing the FDD algorithm's output for each sample (Step 5) with the ground truth data (Step 3), producing a set of raw evaluation outcomes. These raw outcomes are then aggregated to produce one or more FDD performance metrics (Step 6).

4. Definition of a Fault

The presence of a fault may be—and has been—defined in many ways. The existing literature and commercial FDD tools use three general methods or categories of fault definition: condition-based, behavior-based, or outcome-based.

As an introductory example, consider an air handling unit (AHU) with its cooling coil valve stuck open, causing chilled water to leak through the coil. First, examine the case in which the unit is experiencing a call for heating. The unit's faulted state may be defined by the unit's condition (the chilled water valve is stuck open), behavior (the unit is simultaneously heating and cooling), or outcome (the unit's chilled water consumption is greater than expected). If, however, the same unit were cooling rather than heating, it would still be considered faulted under the condition-based definition (the valve is still stuck), but not under the behavior-based definition (it is no longer simultaneously heating and cooling). The unit's state under the outcome-based definition would be determined by the amount of chilled water flow through the stuck valve compared to an expected level of chilled water consumption.

Although rarely identified explicitly, these three categories of fault definition are used consistently in disparate fields, including aerospace, industrial process control, power systems, and buildings. With respect to building HVAC systems, Wen and Regnier [37] distinguish between the condition-based and behavior-based categories while Yuill and Braun [25, 32] describe the outcome-based category. Here, we extend the prior research by formally defining and comparing the three categories.

4.1. Condition-Based

The condition-based definition of a fault is the presence of an improper or undesired physical *condition* in a system or piece of equipment. Examples of condition-based fault definitions include stuck valves, fouled coils, and broken actuators. In the case of control systems, the definition may be extended to encompass an error in the underlying control code. Although the faulty condition may (and typically will) cause improper or undesired system or equipment operation, the presence or absence of such operation does not define the presence or absence of the fault. Rather, the system is faulted so long as the faulty condition is present, regardless of whether its behavior is presently exhibiting symptoms of the fault.

Many existing articles on FDD evaluation use exclusively condition-based ground truth. Examples can be found in the aerospace [26], defense [38], power systems [39], water treatment [35], and buildings industries [36, 40, 41]. Among articles that use different categories of fault definition for different faults, condition-based definitions are also common, for example, Morgan et al. [42].

4.2. Behavior-Based

The behavior-based definition of a fault is the presence of improper or undesired *behavior* during the operation of a system or piece of equipment. Examples of behavior-based fault definitions include simultaneous heating and cooling and short cycling. Typically, the faulty behavior is caused by some underlying faulty condition; Wen and Regnier [37] observe that many faults can be described in terms of either symptoms (behavior) or sources (underlying conditions). However, the key difference between the condition-based and behavior-based fault definitions is the treatment of the case when a fault condition is physically present but the system or equipment is not symptomatic: a condition-based definition still considers the system faulted, but a behavior-based definition does not.

Faulty behavior is typically defined with respect to rules—logical statements that dictate expected behavior. Alternatively, faulty behavior may be defined using observability criteria; for instance, the results of a hypothesis test that the observed sensor readings differ statistically from normal operation. Analysis of fault observability (detectability) is widely used in chemical and industrial process monitoring [43, 44].

A few articles describe mixes of faults, of which some have a behavior-based ground-truth definition: diesel engine overheating [42], reduced condenser and evaporator water flow rates for chillers [31], and failure to maintain air handling unit temperature and pressure set points [37]. Regardless of the ground truth definition, use of equipment behavior as the primary fault detection criteria is common in FDD algorithms, particularly rule-based algorithms that leverage indirect sensor readings [24, 25, 36, 45].

4.3. Outcome-Based

The outcome-based definition of a fault is a state in which a quantifiable *outcome* or performance metric for a system or piece of equipment deviates from a correct or reference outcome, termed the expected outcome. Examples of outcome-based fault definitions include increased hot or chilled water consumption (compared to an expected value), reduced coefficient of performance (compared to an expected or rated value), and zone temperature outside of comfort bounds. Although there is significant overlap between behavior-based and outcome-based fault definitions, the key feature of an outcome-based definition is the presence of an expected, or baseline, outcome against which the system or equipment performance is compared.

Use of an outcome-based fault definition is common in manufacturing and industrial process control, in which the key criterion is whether the output of the production process conforms to expected metrics or

tolerances [34, 46]. In the buildings industry, [25, 32] have proposed that ground truth samples for unitary equipment faults be classified as faulted or unfaulted according to their fault impact ratio (FIR), which is the ratio between the measured and baseline value of some metric of interest,

$$\mathrm{FIR} = \frac{\mathrm{Value}_{\mathrm{faulted}} - \mathrm{Value}_{\mathrm{unfaulted}}}{\mathrm{Value}_{\mathrm{unfaulted}}}. \quad (1)$$

Aside from the process control industry, only a few articles surveyed used an outcome-based detection method within the FDD algorithm. Frank et al. [47] use deviation of building energy consumption outside of normal bounds as the fault detection criteria. This approach is similar to energy monitoring tools that flag abnormal energy consumption in monthly utility bills, for example, Reichmuth and Turner [48].

5. Definition of an Input Sample

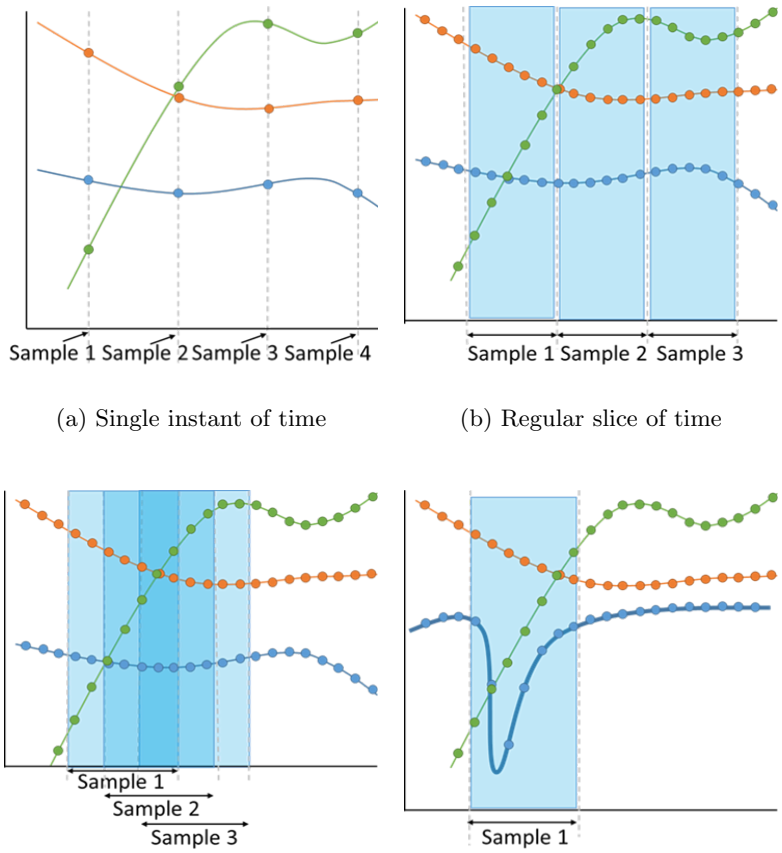
AFDD performance evaluation requires a data library consisting of a large set of input samples, which the AFDD algorithm will process to produce raw outcomes for evaluation. There are several ways to define an input sample (Figure 3). The existing academic literature uses two common methods: a single instant of time and a regular slice of time.

5.1. Single Instant of Time

An input sample defined as a single instant of time (Figure 3a) consists of a single set of simultaneous measurements of the selected system variables, representing a snapshot of system parameters under a certain condition. This type of input sample is well-suited for use with continuous processes and has been used in diverse contexts, including for aerospace applications [27], diesel engines [42], wastewater treatment [35], chillers [22], and air conditioning equipment [25, 40].

5.2. Regular Slice of Time

An input sample defined as a regular slice of time (Figure 3b) contains multiple measurements of the selected system variables recorded within a fixed time window (for example, one day or one week). In the academic literature, time slices are typically on a repeating cycle (for example, every hour on the hour) and measurements within the time slice are recorded at a regular interval (for example, each minute). Use of this type of input sample is also common in the academic literature [26, 36, 39, 41, 45, 49]. In some evaluation approaches (for example, Zhao et al. [45]), the fault is imposed for the full duration of the time slice. In other cases (for example, Ferretti et al. [36]), the fault is imposed for only a portion of the time slice but the entire sample is nevertheless considered to represent a fault.



(a) Single instant of time

(b) Regular slice of time

(c) Rolling time horizon

(d) Event

Figure 3: Various ways to define an input sample for FDD algorithm evaluation

5.3. Other Definitions for Input Samples

Other, less common definitions for input samples include rolling time horizons, event-based windows, and hybrid windows that combine nonconsecutive measurements or combine concepts from the single instant in time and regular slice of time definitions. The rolling time horizon definition for an input sample (Figure 3c) is similar to a regular slice of time (Figure 3b), but the time window shifts through time at a fixed interval of less than the window width (for example, 60-minute windows centered on each minute of the day). Event-based input samples define a sample as a set of measurements taken within a window of time immediately before, during, and/or after a triggering event. An event may be a large change in a monitored variable (Figure 3d) or an external action, such as takeoff of an aircraft [38, 50] or insertion of a fault condition [26]. Use of rolling time horizon-based or event-based input samples for evaluation is uncommon in the academic literature, and the few available literature examples of event-based samples are all outside of the buildings domain. However, some commercial AFDD algorithms use these types to determine AFDD outputs.

The three papers mentioned above also illustrate hybrid definitions of an input sample. To evaluate FDD algorithms for aircraft engines, DePold et al. [38] and Simon et al. [50] use a hybrid sample consisting of two sets of nonconsecutive steady-state measurements recorded during two separate events: takeoff and cruise. Kurtoglu et al. [26] combine event-based and single instant in time definitions for input samples. The evaluation samples consist of variable-length time series data collected after a fault is inserted in an electrical power system (an event). The authors compute temporal performance metrics with respect to single instances of time within this time series but use the AFDD algorithm outputs for the final instant of time within the event window to compute static metrics.

5.4. An Illustrative Example

Consider again the example of a stuck AHU chilled water valve. The input sample definitions provided above are illustrated by a few typical rules that commercial AFDD software might use to detect this fault:

- **Single instant of time:** A simple rule to detect a stuck valve might sample and compare the valve command and status at a regular interval (for instance, every 15 minutes) and label any difference as a fault. One result is reported per sample.
- **Regular slice of time:** A more sophisticated version of the rule might sample and compare the valve command and status multiple times each hour, reporting a fault only if the number of times that the values differ exceeds a pre-determined threshold. One result is reported per time period.
- **Rolling time horizon:** A third possibility is an algorithm that examines valve status and reports a fault if it has not changed for a predetermined amount of time, for example, 24 hours. The time threshold (in this case, 24 hours) represents the length of the rolling time horizon.

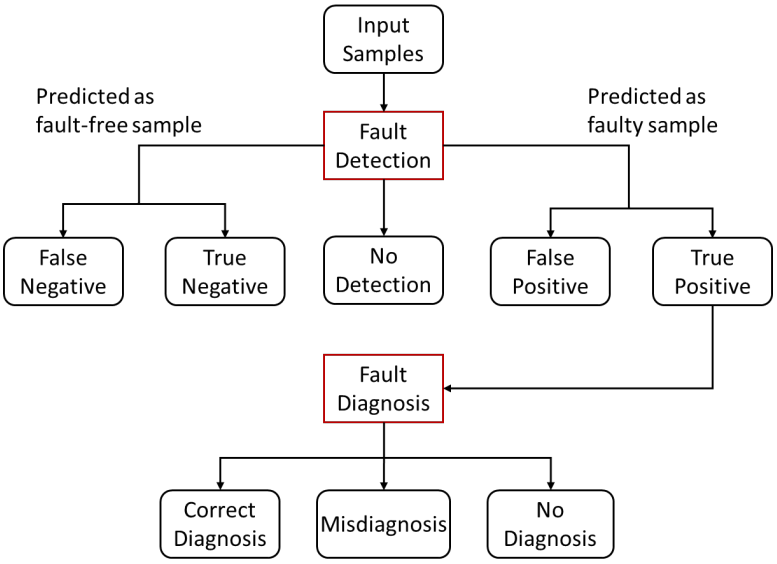


Figure 4: Classification of fault detection and diagnosis outcomes during algorithm evaluation. (Adapted from Reddy [24, Figure 1])

6. Evaluation Outcomes

FDD performance metrics are abundant in the literature [24–26], and most of them are quantitative measures. Existing AFDD performance metrics may be divided into two categories: temporal and static [26]. Temporal metrics quantify an FDD algorithm's evolving response to a time-varying fault signal, while static metrics quantify an FDD algorithm's performance with respect to a collection of samples independent of their ordering in time. As discussed in Section 7.1.1, the raw evaluation outcomes used to compute these metrics are strongly influenced by the choice of fault and input sample definitions.

Most static performance metrics are computed using the same basic set of possible algorithm outcomes. Conceptually, an FDD algorithm labels a sample as faulty or fault-free (detection), and, if faulty, describes the possible cause(s) of the fault (diagnosis). The algorithm may also fail to provide an output for either the detection stage or the diagnosis stage. Combining these possibilities for algorithm output with possible ground truth states yields five possible outcomes for fault detection and three for fault diagnosis (Figure 4):

False positive refers to the case in which the ground truth indicates a fault-free state but the algorithm reports the presence of a fault. Also known as a false alarm or Type I error,

False negative refers to the case in which the ground truth indicates a fault exists but the algorithm reports a fault-free state. Also known as missed detection or Type II error.

True positive refers to the case in which the ground truth indicates a fault exists and the algorithm correctly reports the presence of the fault.

True negative refers to the case in which the ground truth indicates a fault-free state and the algorithm correctly reports a fault-free state.

No detection refers to the case in which the algorithm cannot be applied (for example, due to insufficient data) or the algorithm gives no response because of excessive uncertainty.

Correct diagnosis refers to a true positive case in which the predicted fault type (diagnosed cause) reported by the algorithm matches the true fault type.

Misdiagnosis refers to a true positive case in which the predicted fault type does not match the true fault type.

No diagnosis refers to a true positive case in which the algorithm does not or cannot provide a predicted fault type, because, for example, of excessive uncertainty.

The most commonly used performance metrics comprise the rate of these outcomes across the input samples, such as the false positive rate, false negative rate, and so on. For example, the true positive rate is the proportion of positive fault cases that are correctly identified as such. For a more comprehensive discussion of performance metrics including conceptual illustrations, full mathematical definitions, and a survey of technically advanced metrics, refer to [51].

7. Discussion

In order to ground the review presented in this article in the actual practice of FDD algorithm developers, vendors, implementers, and end users, the authors interviewed six domain experts with deep knowledge of the building analytics industry: three in the commercial sector and three in the academic sector. This section presents the result of these interviews, followed by a discussion of the impact of evaluation procedure choices on evaluation outcomes and on data set generation. Additional methodology concerning these expert interviews is documented in [51].

7.1. Summary of Industry Expert Opinion

All six industry experts agreed that both commercially available and research AFDD algorithms can be found that leverage all three fault definitions for fault detection. Experts were split on the question of what fault definition to use in a ground truth data set intended for FDD algorithm evaluation. All experts interviewed were extremely hesitant to select a single approach, citing the need for more context. Nearly all experts noted that condition-based definitions are more widely used and more appropriate for fault diagnosis, even when the detection algorithm is behavior-based or outcome-based. Experts noted that behavior-based and outcome-based fault definitions have little diagnostic power. However, experts disagreed as to whether algorithms should be penalized for differences in the fault definitions used for detection and diagnosis.

Within a given FDD algorithm, an input sample may be preprocessed into one or several analysis elements required by the algorithm. Most experts stated that they are familiar with at least one algorithm that uses each of the four ways to define an analysis: a single instant of time, a regular slice of time, a rolling time horizon, and an event. Experts noted that algorithms typically use one output for each analysis element. When multiple analysis elements are used, these outputs may require aggregation to yield a single outcome for the input sample. All experts agreed that some form of notification delay setting commonly exists in FDD algorithms, especially in commercially available AFDD tools. The delay setting may be based on fault duration or number of fault appearances counted from intermediate AFDD results. Most experts recommended using a "regular slice of time" (time window) of one day or longer for evaluation samples, as this length is well-aligned with the design and typical use of commercially available AFDD products for buildings. The exception was for handheld diagostic devices, for which "single instant of time" is a better choice for evaluation samples.

7.1.1. Impact of Evaluation Design Choices on Evaluation Outcomes

The evaluation design choices made for fault and input sample definitions have direct effects on FDD evaluation outcomes. In general, use of a condition-based fault definition results in the largest number of samples being classified as faulted in the ground truth data, while use of an outcome-based definition results in the smallest number of faulted samples. Therefore, all else being equal (including the samples in the evaluation data set), using condition-based ground truth will result in fewer false alarms and more missed detections, while outcome-based ground truth will result in more false alarms and fewer missed detections. Because systems and equipment may exhibit some fault symptoms (adverse behaviors) without significantly altering performance outcomes, using behavior-based ground truth is likely to yield evaluation results that fall somewhere between the results for the other two definitions. These trade-offs are apparent in the literature [25, 45].

One key way that the definition of an input sample affects evaluation outcomes is by defining the number of cases counted in the evaluation, which is important for ratio-based metrics. For example, if the evaluator uses a single instant of time sample definition for evaluating algorithm A and a regular slice of time (one-hour) sample definition for evaluating algorithm B, then the false alarm rates of the two algorithms cannot be fairly compared side by side as the referencing point differs due to the inconsistent input sample definition. In short, for fair comparison, the definition of input sample should be consistent across all the FDD algorithm candidates involved in an evaluation. Furthermore, as confirmed by industry experts, algorithms differ in reporting timescale. As a result, regardless of the input sample definition selected, there will be instances in which FDD algorithms generate outputs at a different timescale from the input sample. The FDD evaluator should clearly document how this mismatch is handled. Zhao et al. [45] provide an example of good practice for such documentation.

7.1.2. Considerations for Data Set Generation

To generate a data set for FDD evaluation, ground truth must be assigned to each input sample. Because fault impact varies, the evaluator must establish severity thresholds that distinguish between faulted and unfaulted samples. These thresholds should be consistent with the ground truth fault definition method that the evaluator has elected to use. Methods to define thresholds include:

- **Condition-based ground truth:** Yuill and Braun [25] propose the term *fault intensity* (FI), which is defined for each fault in terms of measurable numeric quantities related to the physical condition of the system or its control parameters. FI may be binary (*e.g.*, power failure) or continuous (*e.g.*, refrigerant 15% undercharged). For each fault, the evaluator should document the range of FI values that are considered sufficiently severe to include as faults in the data set.

- **Behavior-based ground truth:** the evaluator should define and document either a set of rules for expected behavior, violation of which establishes a fault, or a statistical significance test for fault observability that establishes when a fault is symptomatic. In the former case, the rules are similar to rules used in rule-based AFDD algorithms: they typically take the form of if/then statements describing expected system actions and may include tunable numeric thresholds.

- **Outcome-based ground truth:** the evaluator should first define the performance metrics (outcomes) of interest. For each outcome, the evaluator must establish and document both a baseline (expected) value (possibly different for each input sample) and the FIR that defines a fault. The requirement for a baseline complicates generation of ground truth. Yuill and Braun [25] discuss the relative merits of various methods for obtaining the baseline.

Evaluation data may be supplied from simulation, laboratory experiments, or field measurements from a real building. Each approach has advantages and disadvantages. The closer the evaluation procedure can adhere to the realism of a field study, the greater the credibility, but the more difficult it is to obtain and sufficiently screen the data. It is important to recognize that all data sets make implicit assumptions about fault prevalence, and these assumptions affect computed performance metrics.

The input sample definition should also be considered when selecting a data set generation approach, because input sample definition constrains the available approaches for generating data and determines the efforts required to process the raw data. The following are key considerations for various input sample types:

- **Single instant of time type of input sample:** It is a snapshot of system operation conditions. Thus, it is usually desirable that the measurements be taken when the system is at a steady state. The steady-state requirement means that the laboratory or model should have the capability to control the operation conditions at a desired value throughout the data generation period. Steady-state operating conditions are hard to find in field data.

- **Regular slice of time type of input sample:** Longer time durations require more laboratory time, which may not be feasible for experiments due to resource constraints. In this case, simulation or building field data may be better data sources.

- **Other types of input sample** (for example, rolling window horizon and event): If a more esoteric type of input sample is selected, considerable computing or programming efforts may be required to convert the raw data to the needed structure.

8. Conclusion

This article proposes a general FDD performance evaluation framework and documents the design decisions required to implement the framework. Two key decisions that are required are the definition of a fault and the definition of an input sample for evaluation. A fault can be defined by the condition or state of a physical system, by a system's undesired or improper behavior, or by a quantitative outcome's deviation from an expected value or range. The choice of fault definition determines the ground truth classification of evaluation input samples and, by extension, affects the values of the metrics computed from FDD outcomes associated with those samples.

In the existing literature, input samples for FDD evaluation are usually defined as a single instant in time (a set of simultaneous measurements) or a regular, repeating slice of time. Commercial FDD tools may also use rolling time horizons or event-based windows. The definition of an input sample has implications for evaluation data set generation, mapping FDD outputs to performance evaluation results, and comparison of FDD algorithms.

8.1. Best Practices

The proposed FDD performance evaluation framework accommodates many options for specific evaluation parameters. This article provides examples of these options and design decisions from the FDD literature for buildings and other industries. Regardless of the specific options chosen, it is critical to clearly disclose and fully document all aspects of the performance evaluation for it to be credible and replicable. Documentation should address the fault and sample definitions employed; relevant metric definitions and mathematical expressions; the scenarios used; and all relevant assumptions about fault prevalence, cost, etc. Additionally, "apples-to-apples" comparison of the performance of AFDD algorithms requires (i) that the algorithms be tested using consistent fault, input sample, and performance metric definitions; and (ii) that they be tested using the same evaluation data set (the same scenarios, input samples, and ground truth). If different data sets must be used (for instance, if evaluators are working independently with access to diverse data sets), then efforts should be made to align the samples statistically (*e.g.*, for similar fault prevalence and severity). These efforts should be clearly documented.

Although there is no single choice of evaluation parameters that will universally be perceived as ideal, the findings from this work indicate some consensus for design of FDD evaluation procedures. Condition-based fault definitions are commonly used in the literature for both algorithm development and as ground truth in FDD performance evaluation. Subject matter experts also noted that condition-based ground truth is the most widely employed and best aligned with diagnosis. In contrast, behavior-based approaches are relatively less frequently used for ground truth in the literature, while outcome-based approaches can present challenges for experimentally generated data sets and data sets drawn from field studies. Taken together, these findings suggest that a condition-based approach to ground truth definition represents the most practical near-term choice.

For input sample definition, regular daily time slices are well-suited for evaluating typical FDD algorithms because many such tools provide results that building operators review daily or weekly. For handheld diagnostic tools, which are often used to perform "spot checks," the best input sample definition is a single point in time. In the case of metrics, false positive rate, false negative rate, and correct diagnosis rate are the most common and therefore lend themselves to ease of interpretation across a broad audience.

8.2. Recommended Future Work

Further research can support the evolution of the proposed general FDD performance evaluation framework into a set of standard, trusted evaluation procedures. To this end, the authors recommend further investigation into user and stakeholder expectations for FDD algorithm performance and comparative analysis, development of publicly available fault performance evaluation data sets that facilitate independent comparison of FDD algorithms, and implementation of case studies that compare the effect of evaluation design choices on evaluation outcomes. Together, these will enhance the industry's understanding of the trade-offs inherent in FDD performance evaluation and the desired form and content of outcomes. High priority longer-term efforts include research to estimate fault prevalence, impact, and cost, as well as the quantification of the nonenergy costs and benefits of acting on FDD algorithm outputs, whether accurate or inaccurate.

Acknowledgement

The authors thank Marina Sofos and Amy Jiron of the U.S. Department of Energy (DOE) Building Technologies Office for their support of this work. In addition, the authors thank the members of the DOE AFDD project technical advisory group for their reviews and feedback and Kim Trenbath of NREL for her assistance with article preparation.

References

[1] K. W. Roth, D. Westphalen, M. Y. Feng, P. Llana, L. Quartararo, Energy Impact of Commercial Buildings Controls and Performance Diagnostics: Market Characterization, Energy Impact of Building Faults and Energy Savings Potential, Technical Report D0180, TIAX LLC, Cambridge, MA, 2005.

[2] S. Katipamula, Improving Commercial Building Operations Thru Building Re-Tuning: Meta-Analysis, 2015.

[3] Y. Yu, D. Yuill, A. Behfar, Fault Detection and Diagnostics (FDD) Methods for Supermarkets - Phase 1, Technical Report 1615-RP, ASHRAE, Omaha, NE, 2017.

[4] M. R. Brambley, P. Haves, S. C. McDonald, P. Torcellini, D. G. Hansen, D. Holmberg, K. W. Roth, Advanced Sensors and Controls for Building Applications: Market Assessment and Potential R&D Pathways, Technical Report PNNL-15149, Pacific Northwest National Laboratory, Richland, WA, 2005.

[5] U.S. Energy Information Administration (EIA), Commercial Building Energy Consumption Survey (CBECS), `www.eia.gov/consumption/commercial/`, 2012.

[6] U.S. Energy Information Administration (EIA), Annual Energy Outlook 2018, `https://www.eia.gov/aeo`, 2018.

[7] J. Hyvärinen, K. Satu (Eds.), Building Optimization and Fault Diagnosis Source Book, Technical Research Centre of Finland, Finland, ISBN 978-952-5004-10-6, oCLC: 246254321, 1996.

[8] A. Dexter, J. Pakanen (Eds.), Demonstrating Automated Fault Detection and Diagnosis Methods in Real Buildings, Technical Research Centre of Finland, Finland, 2001.

[9] S. Katipamula, M. R. Brambley, Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part I, HVAC&R Research 11 (1) (2005) 3–25, ISSN 1078-9669, doi:10.1080/10789669.2005.10391123.

[10] S. Katipamula, M. R. Brambley, Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems—A Review, Part II, HVAC&R Research 11 (2) (2005) 169–187, ISSN 1078-9669, doi:10.1080/10789669.2005.10391133.

[11] W. Kim, S. Katipamula, A Review of Fault Detection and Diagnostics Methods for Building Systems, Science and Technology for the Built Environment 24 (1) (2018) 3–21, ISSN 2374-4731, doi:10.1080/23744731.2017.1318008.

[12] M. Bonvini, M. D. Sohn, J. Granderson, M. Wetter, M. A. Piette, Robust On-Line Fault Detection Diagnosis for HVAC Components Based on Nonlinear State Estimation Techniques, Applied Energy 124 (2014) 156–166, ISSN 0306-2619, doi:10.1016/j.apenergy.2014.03.009.

[13] T. Muller, N. Rehault, T. Rist, A Qualitative Modeling Approach for Fault Detection and Diagnosis on HVAC Systems, in: Proceedings of the 13th International Conference for Enhanced Building Operations, Montreal, Canada, 2013.

[14] D. Jacob, S. Dietz, S. Komhard, C. Neumann, S. Herkel, Black-Box Models for Fault Detection and Performance Monitoring of Buildings, Journal of Building Performance Simulation 3 (1) (2010) 53–62, ISSN 1940-1493, doi:10.1080/19401490903414454.

[15] S. Wang, Q. Zhou, F. Xiao, A System-Level Fault Detection and Diagnosis Strategy for HVAC Systems Involving Sensor Faults, Energy and Buildings 42 (4) (2010) 477–490, ISSN 0378-7788, doi:10.1016/j.enbuild.2009.10.017.

[16] B. Sun, P. B. Luh, Q. S. Jia, Z. O'Neill, F. Song, Building Energy Doctors: An SPC and Kalman Filter-Based Method for System-Level Fault Detection in HVAC Systems, IEEE Transactions on Automation Science and Engineering 11 (1) (2014) 215–229, ISSN 1545-5955, doi:10.1109/TASE.2012.2226155.

[17] D. Zogg, E. Shafai, H. P. Geering, Fault Diagnosis for Heat Pumps with Parameter Identification and Clustering, Control Engineering Practice 14 (12) (2006) 1435–1444, ISSN 0967-0661, doi:10.1016/j.conengprac.2005.11.002.

[18] K. Bruton, P. Raftery, P. O'Donovan, N. Aughney, M. M. Keane, D. T. J. O'Sullivan, Development and Alpha Testing of a Cloud Based Automated Fault Detection and Diagnosis Tool for Air Handling Units, Automation in Construction 39 (2014) 70–83, ISSN 0926-5805, doi:10.1016/j.autcon.2013.12.006.

[19] J. M. House, H. Vaezi-Nejad, J. M. Whitcomb, An Expert Rule Set for Fault Detection in Air-Handling Units, ASHRAE Transactions 107.

[20] J. Granderson, R. Singla, E. Mayhorn, P. Ehrlich, D. Vrabie, S. Frank, Characterization and Survey of Automated Fault Detection and Diagnostic Tools, Technical Report LBNL-2001075, Lawrence Berkeley National Laboratory, Berkeley, CA, 2017.

[21] U.S. Department of Energy (DOE), Find a Product or Service, `https://smart-energy-analytics.org/product-service`, 2018.

[22] T. A. Reddy, J. Braun, S. Bendapudi, A. Singhal, J. Seem, Evaluation and Assessment of Fault Detection and Diagnostic Methods for Centrifugal Chillers - Phase II, Technical Report 1275-RP, ASHRAE, Philadelphia, PA, 2006.

[23] D. P. Yuill, J. E. Braun, Effect of the Distribution of Faults and Operating Conditions on AFDD Performance Evaluations, Applied Thermal Engineering 106 (2016) 1329–1336, ISSN 1359-4311, doi:10.1016/j.applthermaleng.2016.06.149.

[24] T. A. Reddy, Formulation of a Generic Methodology for Assessing FDD Methods and Its Specific Adoption to Large Chillers, ASHRAE Transactions 113 (2007) 334–342.

[25] D. P. Yuill, J. E. Braun, Evaluating the Performance of Fault Detection and Diagnostics Protocols Applied to Air-Cooled Unitary Air-Conditioning Equipment, HVAC&R Research 19 (7) (2013) 882–891, ISSN 1078-9669, doi:10.1080/10789669.2013.808135.

[26] T. Kurtoglu, O. J. Mengshoel, S. Poll, A Framework for Systematic Benchmarking of Monitoring and Diagnostic Systems, in: 2008 International Conference on Prognostics and Health Management, 1–13, doi:10.1109/PHM.2008.4711454, 2008.

[27] S. Aerospace, Health and Usage Monitoring Metrics: Monitoring the Monitor, 2008.

[28] J. M. House, J. E. Braun, T. M. Rossi, G. E. Kelly, Evaluaton of FDD Tools, in: A. Dexter, J. Pakanen (Eds.), Demonstrating Automated Fault Detection and Diagnosis Methods in Real Buildings, Technical Research Centre of Finland, Finland, 319–357, 2001.

[29] T. M. Rossi, J. E. Braun, A Statistical, Rule-Based Fault Detection and Diagnostic Method for Vapor Compression Air Conditioners, HVAC&R Research 3 (1) (1997) 19–37, ISSN 1078-9669, doi:10.1080/10789669.1997.10391359.

[30] S. Katipamula, R. G. Pratt, D. P. Chassin, Z. T. Taylor, K. Gowri, M. R. Brambley, Automated Fault Detection and Diagnostics for Outdoor-Air Ventilation Systems and Economizers: Methodology and Results from Field Testing, ASHRAE Transactions 105 (1).

[31] T. A. Reddy, Application of a Generic Evaluation Methodology to Assess Four Different Chiller FDD Methods (RP-1275), HVAC&R Research 13 (5) (2007) 711–729, ISSN 1078-9669, doi:10.1080/10789669.2007.10390982.

[32] D. Yuill, J. Braun, Methodology for Evaluating FDD Protocols Applied to Unitary Systems, in: B. L. Capehart, M. R. Brambley (Eds.), Automated Diagnostics and Analytics for Buildings, Fairmont Press, Lilburn, GA, 1 edition edn., ISBN 978-1-4987-0611-7, 491–517, 2014.

[33] D. P. Yuill, J. E. Braun, A Figure of Merit for Overall Performance and Value of AFDD Tools, International Journal of Refrigeration 74 (2017) 651–661, ISSN 0140-7007, doi:10.1016/j.ijrefrig.2016.11.015.

[34] J. F. MacGregor, T. Kourti, Statistical Process Control of Multivariate Processes, Control Engineering Practice 3 (3) (1995) 403–414, ISSN 0967-0661, doi:10.1016/0967-0661(95)00014-L.

[35] L. Corominas, K. Villez, D. Aguado, L. Rieger, C. Rosén, P. A. Vanrolleghem, Performance Evaluation of Fault Detection Methods for Wastewater Treatment Processes, Biotechnology and Bioengineering 108 (2) (2011) 333–344, ISSN 1097-0290, doi:10.1002/bit.22953.

[36] N. M. Ferretti, M. A. Galler, S. T. Bushby, D. Choinière, Evaluating the Performance of Diagnostic Agent for Building Operation (DABO) and HVAC-Cx Tools Using the Virtual Cybernetic Building Testbed, Science and Technology for the Built Environment 21 (8) (2015) 1154–1164, ISSN 2374-4731, doi:10.1080/23744731.2015.1077670.

[37] J. Wen, A. Regnier, AHU AFDD, in: B. L. Capehart, M. R. Brambley (Eds.), Automated Diagnostics and Analytics for Buildings, Fairmont Press, Lilburn, GA, 1 edition edn., ISBN 978-1-4987-0611-7, 467–489, 2014.

[38] H. DePold, J. Siegel, J. Hull, Metrics for Evaluating the Accuracy of Diagnostic Fault Detection Systems, ASME Turbo

Expo: Power for Land, Sea, and Air, Volume 2: Turbo Expo 2004 (2004) 835–841doi:10.1115/GT2004-54144.

[39] J. Cusidó, L. Romeral, J. A. Ortega, J. A. Rosero, A. G. Espinosa, Fault Detection in Induction Machines Using Power Spectral Density in Wavelet Decomposition, IEEE Transactions on Industrial Electronics 55 (2) (2008) 633–643, ISSN 0278-0046, doi:10.1109/TIE.2007.911960.

[40] S. Gouw, R. Faramarzi, Is This My Fault? A Laboratory Investigation of FDD on a Residential HVAC Split System, in: 2014 ACEEE Summer Study on Energy Efficiency in Buildings, vol. 1, ACEEE, Pacific Grove, CA, 84–95, 2014.

[41] T. Mulumba, A. Afshari, K. Yan, W. Shen, L. K. Norford, Robust Model-Based Fault Diagnosis for Air Handling Units, Energy and Buildings 86 (2015) 698–707, ISSN 0378-7788, doi:10.1016/j.enbuild.2014.10.069.

[42] I. Morgan, H. Liu, B. Tormos, A. Sala, Detection and Diagnosis of Incipient Faults in Heavy-Duty Diesel Engines, IEEE Transactions on Industrial Electronics 57 (10) (2010) 3522–3532, ISSN 0278-0046, doi:10.1109/TIE.2009.2038337.

[43] H. H. Yue, S. J. Qin, Reconstruction-Based Fault Identification Using a Combined Index, Industrial & Engineering Chemistry Research 40 (20) (2001) 4403–4414, ISSN 0888-5885, doi:10.1021/ie000141+.

[44] S. Joe Qin, Statistical Process Monitoring: Basics and Beyond, Journal of Chemometrics 17 (8–9) (2003) 480–502, ISSN 0886-9383, 1099-128X, doi:10.1002/cem.800.

[45] Y. Zhao, J. Wen, F. Xiao, X. Yang, S. Wang, Diagnostic Bayesian Networks for Diagnosing Air Handling Units Faults – Part I: Faults in Dampers, Fans, Filters and Sensors, Applied Thermal Engineering 111 (2017) 1272–1286, ISSN 1359-4311, doi:10.1016/j.applthermaleng.2015.09.121.

[46] G. Taguchi, S. Chowdhury, Y. Wu, Taguchi's Quality Engineering Handbook, Wiley [u.a.], Hoboken, NJ, ISBN 978-0-471-41334-9, oCLC: 728091434, 2005.

[47] S. Frank, M. Heaney, X. Jin, J. Robertson, H. Cheung, R. Elmore, G. Henze, Hybrid Model-Based and Data-Driven Fault Detection and Diagnostics for Commercial Buildings, in: 2016 ACEEE Summer Study on Energy Efficiency in Buildings, ACEEE, Pacific Grove, CA, 2016.

[48] H. Reichmuth, C. Turner, A Tool for Efficient First Views of Commercial Building Energy Performance, in: 2010 ACEEE Summer Study on Energy Efficiency in Buildings, vol. 3, ACEEE, Pacific Grove, CA, 325–338, 2010.

[49] Q. Jiang, X. Yan, W. Zhao, Fault Detection and Diagnosis in Chemical Processes Using Sensitive Principal Component Analysis, Industrial & Engineering Chemistry Research 52 (4) (2013) 1635–1644, ISSN 0888-5885, doi:10.1021/ie3017016.

[50] D. L. Simon, J. Bird, C. Davison, A. Volponi, R. E. Iverson, Benchmarking Gas Path Diagnostic Methods: A Public Approach, in: AASME Turbo Expo 2008: Power for Land, Sea, and Air, 325–336, doi:10.1115/GT2008-51360, 2008.

[51] S. Frank, G. Lin, X. Jin, R. Singla, A. Farthing, L. Zhang, J. Granderson, Metrics and Methods to Assess Building Fault Detection and Diagnosis Tools, Technical Report, National Renewable Energy Laboratory, Golden, CO, In Press.