Fault Diagnosis of HVAC Air-Handling Systems Considering Fault Propagation Impacts Among Components

Ying Yan, *Student Member, IEEE*, Peter B. Luh, *Fellow, IEEE*, and Krishna R. Pattipati, *Fellow, IEEE*

***Abstract*— In a heating, ventilation, and air conditioning system, an air-handling system is a key module. Its components (e.g., air handling unit, air-mixing box, and fans), linked through airflows, condition air to a desired temperature and/or humidity based on comfort or controlled environment requirements. Identifying failure modes and estimating their severities allow maintenance crews to know which faults have occurred, how critical they are, and be guided in the repair process to improve the system availability. The problem of fault detection and diagnosis in air-handling systems is complex because of fault propagation across components, and high false alarm rates caused by uncertainties in system and measurement dynamics. In this paper, to capture fault propagation impacts in an efficient manner, dynamic hidden Markov models are developed to identify failure modes, since they contain state transition matrices depending on other components and do not generate joint states. To filter out false alarms, "coupled statistical process control" techniques are developed by using state transitions matrices representing coupling among components. Experimental results show that the method can effectively diagnose faults with high-diagnosis accuracy.**

***Note to Practitioners*—Faults in heating, ventilation, and air conditioning air handling units (AHU) may cause high energy consumption and discomfort to occupants. Fault diagnosis in AHU is challenging since: 1) effects of faults propagate across components connected by airflows and 2) measurement noises may cause high false alarm rates. In this paper, a novel fault diagnosis method is established to identify failure modes and fault severities. This method explicitly considers the fault coupling among components. To reduce false alarm rates, new statistical process control techniques are developed to filter out false alarms. Experimental results show that our method can effectively diagnose faults with high diagnosis accuracy.**

***Index Terms*— Air handling system, air handling unit (AHU), cooling coil, coupled statistical process control (SPC), damper, dynamic hidden Markov model (HMM), failure modes, fan, fault diagnosis, fault propagation, fault severities, variable air volume.**

I. Introduction

APPROXIMATELY 50% energy consumption of a commercial building is associated with heating, ventilation, and air conditioning system (HVAC) [1]. In HVAC systems, faults may cause high-energy consumption and make occupants feel uncomfortable. Thus, accurate diagnosis of faults in HVAC systems is critical. As a key module of an HVAC system, air-handling systems are used to condition and circulate air in rooms. These systems can be generally classified into two types, including constant air volume (CAV) and variable air volume (VAV). Unlike a CAV system that supplies a constant airflow at a variable temperature, a VAV system provides a varied airflow at a constant temperature. Various air-handling systems may contain different components, e.g., a mixing box, spray humidifiers and thermal wheels. To bound the scope of the problem, a simple VAV air-handling system consisting of an air-mixing box, an air handling unit (AHU), fans and ducts is considered, as shown in Fig. 1. In the mixing box, the recirculation air damper and the outdoor air (OA) damper are used to mix air in a desired proportion. The mixed air is then delivered to the AHU consisting of filters, a heating coil and a cooling coil. The coils are used to condition the air via heat exchange. The supply fan delivers the air to VAV boxes. The return fan delivers air to exhaust air (EA) damper and the air-mixing box. Before repairing faults, it is important to identify: 1) failure modes and 2) fault severities or failure conditions. The former allows maintenance crews to know which faults occurred and their locations. The latter helps guide maintenance crews to recognize how severe faults are or conditions of faults, e.g., damper stuck positions.

To diagnose faults, component health conditions (e.g., normal and faulty conditions) are required to identify failure modes and fault severities, but they cannot be directly measured. Given models, health conditions of components can be estimated based on measurements with noise, which may cause high false alarm rates. In existing works, to filter

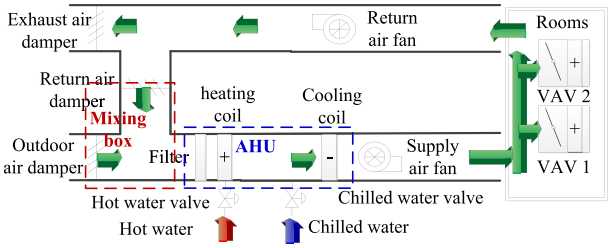

Fig. 1. Structure of a specific air-handling system considered.

Manuscript received November 24, 2016; accepted February 5, 2017. Date of publication March 10, 2017; date of current version April 5, 2017. This paper was recommended for publication by Editor J. Wen upon evaluation of the reviewers' comments. This work was supported by the National Science Foundation under Grant CCF-1331850.

The authors are with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: ying.yan@uconn.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2017.2669892

out false alarms, prior models of noise were established based on physical knowledge that may not be available. Moreover, components are linked through airflows, and may need to satisfy set points. A fault in a component may cause increase of load in others to ensure that set points are still satisfied. Thus, these components are more likely to break down, leading to fault propagation. Capturing the impacts improves diagnosis performance as discussed in our numerical results, e.g., F-measure representing diagnosis accuracy is increased by 2.83%–4.79%. However, they are rarely considered in existing works. Additionally, identification of fault severities is important but is rarely investigated in existing work. Addressing these issues is difficult since: 1) modeling failure modes and fault severities as well as their fault propagation impacts is complex, and requires high-computational efforts and 2) measurement noises may cause high false alarm rates.

In this paper, faults in an air-handling system under the cooling mode are considered. Based on accepted practice in HVAC systems [2], [3], 16 faults are considered in this paper. For the EA damper, stuck closed/open are considered. For the OA damper, stuck closed and leakage are considered. For the cooling coil, four faults including: 1) tube fouling; 2) dust on fins; 3) stuck closed valve; and 4) stuck open valve are considered. For ducts, leaking before/after the supply fan are considered. For supply/return fans, three faults, including: 1) complete failure; 2) running at a fixed speed; and 3) decrease in fan efficiency, are considered. Some of these faults, e.g., damper stuck closed, occurs suddenly, are considered as sudden faults; others, e.g., tube fouling, become worse gradually, are considered as gradual faults. Each failure mode has fault severities representing how severe they are. Some faults, including: 1) tube fouling; 2) dust on fins; 3) decrease of fan efficiency; and 4) duct leakage, are reflected by multiple parameters. Thus, fault severities are identified based on decrease/increase percentages of these parameters. Other faults do not have severities thus their failure conditions, e.g., damper stuck angle, are identified based on measurements. In Section II, typical fault diagnosis methods are classified and reviewed.

To identify failure modes, models are required to estimate health conditions. These models need to capture behaviors of components as well as coupling among components. Physics-based models capture component behaviors, but they are developed for individual components, and the coupling is rarely considered. Hidden Markov model (HMM) is a statistic Markov model, which represents relationships between hidden states and observations statistically. To capture the coupling, coupled HMMs are usually used [4], but they generate many joint states, leading to high-computational requirement. Since the fault propagation takes a long time, it is not necessary to capture it precisely by coupled HMMs. To capture the coupling in an efficient way, in Section IV, dynamic HMMs are developed to capture the impacts since they: 1) contain different state transition matrices depending on other components and 2) do not generate joint states. Additionally, capturing fault severities in dynamic HMMs generates many discrete states with low resolution. To address this issue, given identified failure modes, model parameters are estimated using a Kalman filter (KF) or an unscented particle filter (UPF). Fault severities are represented in terms of increase/decrease percentages of the estimates with high resolution.

In Section V, to estimate parameters and states of dynamic HMMs, EM algorithm is commonly used [5]. However, it computes log-likelihood functions with associated singular problems. Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm that does not calculate the functions, and hence singular problems are avoided. A Gibbs sampler is developed for dynamic HMMs, while considering fault dependencies among components. Statistical process control (SPC) is a method to monitor and control a process. To filter out false alarms, SPC can be used to check whether components are faulty or estimated as a failure falsely by comparing estimates with their control limits. However, it was usually developed for individual components and fault propagations were not captured [6]. To address this issue, different state transition matrices representing fault propagation are used to derive control limits.

In Section VI, data from a simulated small building and ASHRAE project 1312-RP are used to test our fault diagnosis method. Experimental results show that this method can diagnose faults with high F-measure scores.

II. Literature Review

To diagnose faults in HVAC systems, many methods have been developed, and are generally categorized into three groups, including: 1) quantitative model-based; 2) qualitative model-based; and 3) process history based [7].

A. Quantitative Model-Based Methods

These methods use explicit mathematical models, including physics-based and gray-box models, to represent behaviors of components. Physics-based models express systems in terms of mathematical functions based on physical knowledge. Gray-box models are usually simplified forms of physics-based models, and part of physical knowledge is retained in these models. Given input measurements, outputs are estimated and compared with measured outputs to identify failure modes. Variables (e.g., temperatures or the static air pressure) may be considered as outputs [8]. Control signals under faulty conditions can also be considered as outputs, since they are different from those under normal operation to compensate for effects of faults [9]. Additionally, parameters related to faults, e.g., leakage rate and UA value, can be considered as outputs [6], [10], [11]. To estimate these parameters, multiple methods were developed. In [10], a normalized least mean-square metric was used to estimate model parameters. In [6] and [11], KF and extended KF were used to estimate parameters of physics-based models to identify sudden and gradual faults. Several papers related residuals to fault severities, but they did find appropriate measures to identify the severities [12], [13]. Quantitative models are accurate and capture system behaviors under both normal and faulty conditions. However, they are hard to develop and may require variables which are not available. Additionally, these models

are usually developed for individual components, and fault propagation among components are rarely captured.

B. Qualitative Model-Based Methods

Qualitative models are also developed based on physical principles. Unlike quantitative models, they represent qualitative cause-effect relationships among faults and observed effects to infer faults. For example, rule-based methods employ physical knowledge and individual experience to generate multiple if-then-else rules for fault diagnosis, and are widely used in HVAC systems [15]–[17]. In [15], rules were developed and implemented in a decision tree to diagnose sudden and gradual faults in outdoor-air ventilation and economizer operation. Based on this method, a tool known as the outdoor-air economizer diagnostician was implemented. In [16], a rule set was developed to diagnose faults in AHUs under different operating modes, and was implemented in a software called AHU performance assessment rules [17]. Based on physical knowledge, a decision tree was developed to diagnose sudden and gradual faults of an AHU using data from the ASHRAE project 1312-RP [18]. Failure modes and failure conditions, e.g., the fan fixed speed, were considered together and identified simultaneously. These methods have the virtue of explanatory capability for fault inference. However, it is difficult to ensure that all rules are applicable for different systems and for all operating conditions. Developing rules requires expertise and knowledge. Additionally, qualitative methods do not model fault behaviors using mathematical functions, thus cannot identify fault severities with high resolution.

C. Process History Methods

In these data-driven methods, relationships between measured inputs and outputs are represented by black-box models. Unlike physics-based models, black-box models are established based only on data without regard to physical principles. Multiple methods, e.g., principal component analysis (PCA), artificial neural networks (ANNs), support vector machine (SVM) and HMMs, have been developed to diagnose faults in HVAC systems. For instance, two PCA models were developed based on AHU measurements related to heat balance and pressure-flow balance to diagnose sensor faults [19]. In [20], a wavelet-PCA method was developed to diagnose sudden and gradual faults of an AHU by removing the influence of weather conditions. As black-box models, ANNs are good at classifying conditions of components by learning from training data. This method is widely used for HVAC systems [21]–[23]. In [24], SVM techniques were applied to model parameters representing AHU states to classify faults. This method was compared with others based on F-measure that measure the quality of diagnosis.

HMMs are black-box models, and represent state evolution and relationships between states and measurements statistically. To diagnose faults, HMMs are trained based on normal and faulty data. Log-likelihood values obtained from these models will be compared with each other, and the HMM with the largest value is the most likely model [25]. By using this method, HMMs for all failure modes are required, leading to high-computational requirements. To avoid establishing many HMMs, multiple methods treated component health conditions as HMM states and then use HMM inference techniques to identify faults. For instance, to diagnose faults of power distribution systems, engines, etc., coupled factorial HMM was used to estimate system states based on measurements to infer faults [4]. To estimate parameters of HMMs, multiple methods were developed, e.g., EM algorithm [5] and Gibbs sampler [26]. In the EM algorithm, log-likelihood function $\log(f)$ is calculated and will diverge to minus infinity if f approximates to 0, leading to singularities [27]. Gibbs sampling is an MCMC algorithm. Unlike the EM algorithm, Gibbs sampler does not calculate the log-likelihood function, thus convergence to singularities is avoided [26]. To estimate HMM states, measurement noise may cause high false alarm rates. To filter out false alarms, *a priori* model of measurement uncertainties in HMMs was established based on physical knowledge [28]. However, the prior information of uncertainties may not be available. Additionally, SPC rules, e.g., X-chart, were widely used to filter out false alarms [6], [29]. They are applicable to individual components, but not for coupled ones. Process history methods do not require an understanding of physical systems, and are thus well suited to problems where data are plentiful. However, the models are specific to the system for which data are available and may not carry over to other ones. Moreover, fault severities are represented by discrete states which have low resolution.

In existing works, there are two voids, which are planned to fill. These include: 1) identifications of fault severities were rarely investigated and 2) fault propagation effects among components were rarely considered.

III. Method Framework and Models of Components

In Section III-A, the framework of our fault diagnosis method is presented. In Section III-B, HMMs of EA/OA dampers are established. In Section III-C, a physics-based model and a HMM of a cooling coil are described. In Section III-D, the HMM of a duct is established. In Section III-E, physics-based models and HMMs of supply/return fans are presented. In Section III-F, fault propagation impacts among components are analyzed. Based on the analysis, Section III-G shows how to establish dynamic HMMs.

A. Framework of the Fault Diagnosis Method

To diagnose faults, models capturing transitions among different conditions and fault impacts among components are required. Future states of components depend solely on present states, and thus state sequences possess the Markov property. To capture the property, filtering techniques (e.g., KF and UPF) and HMMs can be used. Usually, filtering methods are built on physics-based models, which may require input variables that cannot be measured. Additionally, a fault in a component may increase load in other components, which are more likely to break down. However, most physics-based models are developed for individual components, and thus

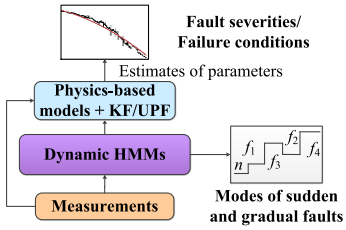

Fig. 2. Framework of our fault diagnosis method.

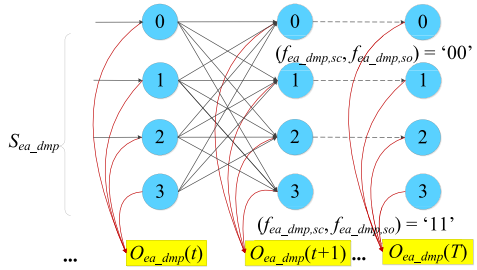

Fig. 3. HMM of the EA damper.

rarely capture fault impacts. Compared to them, HMMs have no strict requirements and can still be used even if multiple measurements are not available. To capture fault propagation, coupled HMM is used [4], but they generate many joint states and require high-computational effort. Since the fault propagation takes long time, it is not necessary to capture it precisely by coupled HMMs. Dynamic HMMs are used since they have different state transition matrices depending on other components to capture fault propagation, and do generate joint states. As shown in Fig. 2, dynamic HMMs are used to identify failure modes.

If dynamic HMMs are also used to identify fault severities, modeling both failure modes and fault severities together could generate many states, resulting in high-computational requirements. Additionally, estimates of dynamic HMMs are discrete, thus have low resolution. To overcome these difficulties, KF/UPF are used to estimate parameters of physics-based models to identify fault severities as shown in Fig. 2. Failure modes and fault severities are identified in succession, thus fewer states are generated and high-computational requirements are avoided. Additionally, estimates obtained by filters are continuous, and thus have high resolution. Failure conditions are identified by measuring corresponding variables directly. In the next set of sections, dynamic HMMs are established for identification of failure modes. Since faults including: 1) tube fouling; 2) dust on fins; 3) decrease in fan efficiency; and 4) duct leakage have fault-related parameters that represent fault severities, the physics-based models containing these parameters will be formalized for estimating fault severities.

B. HMMs of EA/OA Dampers

A damper is used to stop or regulate airflow inside a duct or air-handling equipment by adjusting the damper blade angle. EA damper could be stuck open or stuck closed suddenly. For the open case, exhaust air increases, thus recirculating air $\dot{m}_{a,\mathrm{rn}}$ decreases, leading to an increase of outdoor airflow rate $\dot{m}_{a,\mathrm{oa}}$ to maintain the supply airflow rate $\dot{m}_{a,\mathrm{sup}}$. Supply fan speed φ_{sf} and return fan speed φ_{rf} are forced to increase to raise the pressure to suck in more air. On the contrary, for the closed case, exhaust air decreases, leading to a decrease of these variables. Damper blade angles are measured to identify failure conditions.

1) HMMs of Dampers: For the EA damper, two sudden failure modes are considered. Thus, the HMM has $2^2 = 4$ states denoted by $s_{\mathrm{ea\_dmp}} = 0$, 1, 2, and 3 as shown in Fig. 3.

In this figure, the stuck closed damper fault is denoted by $f_{\mathrm{ea\_dmp,sc}}$ and the damper stuck open is denoted by $f_{\mathrm{ea\_dmp,so}}$, where $(f_{\mathrm{ea\_dmp,sc}}, f_{\mathrm{ea\_dmp,so}}) =$ '01' means that the damper is stuck closed. The state evolution is determined by a 4×4 state transition matrix. The HMM observations at time t is denoted by $O_{\mathrm{ea\_dmp}}(t)$, which are extracted from fault indicators to capture distinguishable faulty information. Usually, measured variables related to faults are considered as fault indicators. However, they may significantly change with loads even though no faults occur. Compared to them: 1) residuals between model outputs and measurements or those between outputs of different models and 2) differences between measurements and their set-points change a little bit under the normal condition, but significantly under faulty conditions. Thus, they are also considered as fault indicators. Thus, the fault indicator matrix is

$$X_{\mathrm{ea\_dmp}} = \begin{bmatrix} \dot{m}^1_{a,\mathrm{sup}} & \dot{m}^1_{a,\mathrm{oa}} & \dot{m}^1_{a,rn} & \varphi^1_{\mathrm{sf}} & \varphi^1_{\mathrm{rf}} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \dot{m}^K_{a,\mathrm{sup}} & \dot{m}^K_{a,\mathrm{oa}} & \dot{m}^K_{a,rn} & \varphi^K_{\mathrm{sf}} & \varphi^K_{\mathrm{rf}} \end{bmatrix} \quad (1)$$

where K is the length of the state sequence. These fault indicators contain measurement noises. Denote $X_{\mathrm{ea\_dmp}}$ by y and their true values by u, then

$$y_t = u_t + w_t \quad (2)$$

where w_t is the measurement noise at time t. The measurement noise is assumed to follow a multivariate normal distribution $N(u_t, \Sigma)$. To judge this statement, the Chi-square goodness of fit test is used to test the hypothesis that data comes from a normal distribution [30]. Thus, the emission probability function that represents the probability density of observations given its true values is

$$b_t(y_t, u_t) = |2\pi\,\Sigma|^{-1/2} \exp[-(y_t - u_t)' \Sigma^{-1} (y_t - u_t)]. \quad (3)$$

These fault indicators have different units, and differences between them are large. To avoid $X_{\mathrm{ea\_dmp}}$ becoming a singular matrix, they are normalized. These indicators are correlated, and contain redundant information. To remove redundancy, PCA is used to project this variable matrix into a reduced space. If m vectors represent more than 95% correlated information, they are regarded as adequate set of "principal components" to reflect original spaces, and regarded as observations [17]. For $X_{\mathrm{ea\_dmp}}$, the first three principal components capture 97.831% of variation in the observed data, and thus are

selected as observation $O_{\text{ea\_dmp}}$. EA damper has four observation distributions corresponding to its states. The HMM of the OA damper is developed in a similar manner.

C. Physics-Based Model and HMM of a Cooling Coil

A cooling coil is a coiled arrangement of tubes for heat transfer between chilled water and air. Chilled water flows through tubes, and air passes through fins. For the cooling coil, two sudden faults, including: 1) valve stuck fully/partially closed $f_{\text{cc,vlv\_sc}}$ and 2) valve stuck fully/partially open $f_{\text{cc,vlv\_so}}$, and two gradual faults: 1) tube fouling $f_{\text{cc,tube}}$ and 2) dust on fins $f_{\text{cc,fin}}$ are considered.

1) Physics-Based Model of a Cooling Coil: The heat transfer coefficient U_{cc} of the cooling coil is the reciprocal of the thermal resistance representing how the cooling coil resists a heat flow. The thermal resistance consists of four parts, including: 1) air-side thermal resistance $1/\alpha_{a,e}\tau$; 2) thermal resistance caused by dust $1/R_f$; 3) thermal resistance of the tube wall; and 4) water-side thermal resistance $1/\alpha_w$ [31]

$$\text{UA}_{\text{cc}} = \left[\frac{1}{\alpha_{a,e}\tau} + R_f + \frac{\delta_{\text{tube}}}{\lambda_{\text{tube}}} + \frac{1}{\alpha_w}\right]^{-1} (A_{\text{tube,out}} + A_{\text{fin}}) \quad (4)$$

with

$$\alpha_{a,e} = \alpha_a[1 - A_{\text{fin}}(1 - \eta_{\text{fin}})/(A_{\text{tube,out}} + A_{\text{fin}})], \text{ [32]} \quad (5)$$

$$\text{and } \alpha_w = A \cdot \dot{v}_{\text{chw}}^{0.4} \cdot \psi^{0.4}/d_{\text{tube,in}}^{0.6} \quad (6)$$

where A_{fin} and $A_{\text{tube,out}}$ are the fin surface area and the tube outside surface area, respectively; parameter $\tau = (A_{\text{fin}} + A_{\text{tube,out}})/A_{\text{tube,in}}$, and $A_{\text{tube,in}}$ is the tube inside surface area; parameter δ_{tube} is the tube thickness; parameter λ_{tube} is the tube thermal conductivity; parameter η_{fin} is the fin efficiency; variable $\dot{v}_{\text{chw}}$, is the chilled water volumetric flow rate; variable ψ is the heat flux; and $d_{\text{tube,in}}$ is the inside tube diameter. Additionally, the heat transfer coefficient can be calculated based on air mass flow rate and logarithmic average of the temperature difference (LMTD)

$$\text{UA}_{\text{cc}} = \dot{m}_{a,\text{sup}} \cdot (E_{a,\text{mix}} - E_{a,\text{dis}})/\text{LMTD} \quad (7)$$

$$\text{with LMTD} = (\Delta T_{\text{sup}} - \Delta T_{\text{rn}})/(\ln \Delta T_{\text{sup}} - \ln \Delta T_{\text{rn}}) \quad (8)$$

where $\Delta T_{\text{sup}} = T_{a,\text{dis}} - T_{\text{chw,sup}}$, $\Delta T_{\text{rn}} = T_{a,\text{mix}} - T_{\text{chw,rn}}$. Variables $T_{a,\text{mix}}$ and $T_{a,\text{dis}}$ are mixed air and discharge air temperatures; variables $T_{\text{chw,sup}}$ and $T_{\text{chw,rn}}$ are supply and return chilled water temperatures. Given (4) and(7), the gray-box model of a cooling coil is obtained as

$$\frac{\dot{m}_{a,\text{mix}} \cdot (E_{a,\text{mix}} - E_{a,\text{dis}})}{\text{LMTD}} = \left[\frac{1}{\alpha_{a,e}\tau} + \frac{\delta_{\text{tube}}}{\lambda_{\text{tube}}} + R_f + \frac{1}{\alpha_w}\right]^{-1} \times (A_{\text{tube,out}} + A_{\text{fin}}). \quad (9)$$

The right-side of (9) depends on geometrical parameters, e.g., $d_{\text{tube,in}}$ and A_{fin}. Parameter $d_{\text{tube,in}}$ gradually decreases with tube fouling. A certain decrease in the tube diameter represents the degree of fault severity. Similarly, parameter A_{fin} decreases with dust accumulation on fins, and its decrease by a certain percentage represents the fault severity. If these parameters are set to normal values, (9) will be violated under faulty conditions. Thus, the residual between the left-side and the right-side is considered as a fault indicator. Based on discussion in [3], chilled water flow rate $\dot{m}_{\text{chw}}$, the difference between the enthalpy of mixed air $E_{a,\text{mix}}$ and the enthalpy of supply air $E_{a,\text{sup}}$, $\dot{m}_{\text{a,sup}}$, and φ_{sf} are related to faults and are selected as fault indicators. Valve openings are measured to identify failure conditions of valve faults. Additionally, faults could result in: 1) the supply air temperature cannot follow its set-point and 2) the zone temperature cannot satisfy its set-point. Therefore, these differences are also considered.

2) HMM of a Cooling Coil: Four failure modes are considered. Thus, the HMM has $2^4 = 16$ states denoted by $s_{\text{cc}} = 0, \ldots, 15$, and has a 16×16 state transition matrix. As discussed above, the fault indicator matrix is shown in (10) shown at the bottom of this page, where variable R_{Kcc} is the residual between the left-side and the right-side of (9); variable $\Delta T_{\text{spt,sup}}$ is the difference between the supply air temperature and its set-point; variable $\Delta T_{\text{spt,zone}}$ is the difference between the zone temperature and its set-point. The first five principal components capture 96.349% of variability in data, and thus are used as observations.

D. HMM of a Duct

Ducts are used to deliver and remove air. In ducts, air may leak gradually before the supply fan and is denoted by $f_{\text{duct,lb}}$. It may cause a decrease in the outdoor airflow rate as well as the supply fan speed. On the contrary, the leakage after the supply fan $f_{\text{duct,la}}$ may cause an increase in the outdoor airflow rate and an increase in the supply fan speed to compensate this situation. As two faults are considered, the duct has $2^2 = 4$ states denoted by $s_{\text{duct}} = 0$, 1, 2, and 3, and a 4×4 state transition matrix. Impacts of these faults were analyzed in [3]. Based on the analysis, the fault indicator matrix is

$$X_{\text{duct}} = \begin{bmatrix} \varphi_{\text{sf}}^1 & \varphi_{\text{rf}}^1 & \dot{m}_{a,\text{sup}}^1 & \dot{m}_{a,\text{oa}}^1 & \dot{m}_{a,\text{rn}}^1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \varphi_{\text{sf}}^K & \varphi_{\text{rf}}^K & \dot{m}_{a,\text{sup}}^K & \dot{m}_{a,\text{oa}}^K & \dot{m}_{a,\text{rn}}^K \end{bmatrix}. \quad (11)$$

The first four principal components capture 95.867% of variation in data, and are thus used as observations.

E. Physics-Based Models and HMMs of Supply/Return Fans

In an air-handling system, fans deliver conditioned air to rooms or remove air to outside. In the fan, the rotating blades increase the pressure of air by consuming electricity, and move air against resistances. Usually, the supply fan and the return

$$X_{\text{cc}} = \begin{bmatrix} T_{\text{chw,rn}}^1 - T_{\text{chw,sup}}^1 & E_{a,\text{mix}}^1 - E_{a,\text{dis}}^1 & m_{a,\text{sup}}^1 & \dot{m}_{\text{chw}}^1 & R_{\text{Kcc}}^1 & \Delta T_{\text{spt,sup}}^1 & \Delta T_{\text{spt,zone}}^1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ T_{\text{chw,rn}}^K - T_{\text{chw,sup}}^K & E_{a,\text{mix}}^K - E_{a,\text{dis}}^K & m_{a,\text{sup}}^K & \dot{m}_{\text{chw}}^K & R_{\text{Kcc}}^K & \Delta T_{\text{spt,sup}}^K & \Delta T_{\text{spt,zone}}^K \end{bmatrix} \quad (10)$$

fan are the same type, thus the former is used as an example. Two sudden faults, including: 1) complete failure $f_{\text{sf,cf}}$ and 2) running at a fixed speed $f_{\text{sf,fs}}$, and a gradually decrease in supply fan efficiency $f_{\text{sf,eff}}$ are considered.

1) Physics-Based Model of a Fan: Decrease in fan efficiency e_{sf} may cause an increase in fan power Q_{sf} to maintain the supply air mass flow rate $\dot{m}_{a,\text{sup}}$. Additionally, the fan pressure rise $\Delta P_{r,\text{sf}}$ can be increased by consuming more electricity. Thus, a physics-based model of a fan was developed as in [33]. The model is

$$Q_{\text{sf}} = \dot{m}_{a,\text{sup}} \cdot \Delta P_{r,\text{sf}} / (\rho_a e_{\text{sf}}) \tag{12}$$

where ρ_a is the air density. Fan faults, such as wear in a bearing, cause decrease in fan efficiency e_{sf}, which is treated as a fault severity parameter. To compensate for a decrease in e_{sf}, more electricity is consumed to maintain the supply airflow rate. Complete failure and running at a fixed speed cause a decrease in supply airflow, where fan speed is measured to identify failure conditions.

2) HMMs of Fans: Three failure modes of the supply fan are considered. Thus, the HMM has $2^3 = 8$ states, which are denoted by $s_{\text{sf}} = 0, \ldots, 7$. The fault indicator matrix is

$$X_{\text{sf}} = \begin{bmatrix} \varphi^1_{\text{sf}} & \varphi^1_{\text{rf}} & \dot{m}^1_{a,\text{sup}} & \dot{m}^1_{a,\text{rn}} & \Delta T^1_{\text{spt,sup}} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \varphi^K_{\text{sf}} & \varphi^K_{\text{rf}} & \dot{m}^K_{a,\text{sup}} & \dot{m}^K_{a,\text{rn}} & \Delta T^K_{\text{spt,sup}} \end{bmatrix} \tag{13}$$

where variable $\Delta T_{\text{spt,sup}}$ is the difference between the supply air temperature and its set-point. The first three components capture 98.052% of variation in data, and are used as observed features. Similarly, the HMM of the return fan is established.

F. Fault Impacts Among Components

Components of an air-handling system are linked through airflows and may need to satisfy certain set-points. If a component has a fault, loads in other components may increase to ensure that set-points are still satisfied. Thus, these components are more likely to break down. For instance, suppose the cooling coil valve is stuck at fully closed. To compensate for this situation, the supply fan should increase its speed and causing the bearing to wear, leading to a decrease in fan efficiency. Because of increase in supply airflow rate, duct leakage may become worse. To establish accurate models, it is important to capture these fault propagation effects. Additionally, fault propagation effects between two adjacent components are usually larger than those between nonadjacent ones, and such spatial dependencies are considered in our models.

G. Dynamic HMMs of Components

To capture fault impacts among adjacent components, coupled HMMs are usually used. However, they generate many joint states, which grow rapidly with increase in the number of failure modes. To address this issue, dynamic HMMs are developed, since they have different state transition matrices conditioned on other components to capture fault impacts among components without generating joint states. Here, duct is used as an example to show how to establish a dynamic

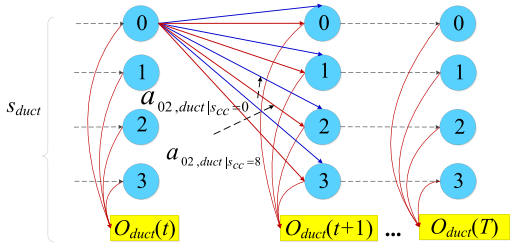

Fig. 4. Dynamic HMM of the duct.

HMM. The duct is linked with the cooling coil. The set of HMM parameters associated with duct is denoted by λ_{duct}. The dynamic HMM associated with the duct has 16 state transition matrices corresponding to the 16 states of the cooling coil, and are denoted by $P_{\text{duct},i}$, $i = 1, \ldots, 16$. It has $64 = 16 \times 4$ observation distributions corresponding to states of the duct and of the cooling coil, and are denoted by $N(\mu_{\text{duct},k}, \Sigma_{\text{duct},k})$, $k = 1, \ldots, 64$. To judge whether parameters and states estimated match observations, the joint probability $P(O_{\text{duct}}, s_{\text{duct}}|\lambda_{\text{duct}}, s_{\text{cc}})$ is calculated using forward recursions of the HMM inference algorithms. The duct observation at time t, $O_{\text{duct}}(t)$, depends on both s_{cc} and s_{duct} at time $t-1$. The probability of the observation sequence, $O_{\text{duct},1}$, Oduct, $2, \ldots O_{\text{duct},t}$ $(t \le T)$ and $s_{\text{duct}} = k$ given the model parameter λ_{duct} and $s_{\text{cc}} = j$ can be derived from [25]

$$\alpha_{\text{duct},1|s_{\text{cc}}=j}(k) = \pi_{\text{duct},k|s_{\text{cc}}=j} L_{\text{duct},k}(O_{\text{duct},1|s_{\text{cc}}=j}) \quad 1 \le k \le M \tag{14}$$

and

$$\alpha_{\text{duct},t+1|s_{\text{cc}}=j}(k) = \left[\sum_{i=1}^{M} \alpha_{\text{duct},t|s_{\text{cc}}=j}(i) a_{ik|s_{\text{cc}}=j}\right] \times L_{\text{duct},k}(O_{\text{duct},t+1|s_{\text{cc}}=j}) \quad 1 \le t \le T-1, \quad 1 \le k \le M \tag{15}$$

where $L_{\text{duct},k} = P(O_{\text{duct}}|\mu_{\text{duct},k}, \Sigma_{\text{duct},k})$ is the probability of observation O_{duct} given the mean vector $\mu_{\text{duct},k}$ and covariance matrix $\Sigma_{\text{duct},k}$ for state k; the parameter $a_{ik|s_{\text{cc}}=j}$ is the probability of transiting from state i to state k given that the cooling coil is in state j; and M is the observation dimension of the duct.

Valve getting stuck closed ($s_{\text{cc}} = 2$) causes an increase in the supply airflow rate, and duct leakage ($s_{\text{duct}} = 2$) before the supply fan is likely to become worse. In other words, the duct is more likely to transit to the faulty condition if the cooling coil is under the faulty condition, as shown in Fig. 4.

To represent this relationship, an equation is shown as

$$a_{02,\text{duct}|s_{\text{cc}}=0} < a_{02,\text{duct}|s_{\text{cc}}=2} \tag{16}$$

where the probability of transiting from $s_{\text{duct}} = 0$ to $s_{\text{duct}} = 2$ given $s_{\text{cc}} = 0$ is denoted by $a_{02,\text{duct}|s_{\text{cc}}=0}$. Similarly, the duct is more likely to be normal given $s_{\text{cc}} = 0$ than $s_{\text{cc}} = 2$, so that

$$a_{00,\text{duct}|s_{\text{cc}}=0} > a_{00,\text{duct}|s_{\text{cc}}=2}. \tag{17}$$

Similarly, dynamic HMMs of other components are established.

IV. Fault Diagnosis Method

In Section IV-A, a Gibbs sampling algorithm is developed for dynamic HMMs to estimate parameters and states to identify failure modes. In Section IV-B, a new coupled-SPC method is developed to filter out false estimates, while considering fault propagation impacts. In Section IV-C, KF/UPF are used to estimate parameters of physics-based models to identify fault severities.

A. Gibbs Sampler to Estimate Parameters and States of Dynamic HMMs

A Gibbs sampler is developed for dynamic HMMs. To estimate parameters, prior distributions are assumed, and posterior distributions are derived based on prior ones by using maximum-*a posteriori* probability estimation. The Gibbs sampler amounts to alternating between updating parameters given observations and the hidden Markov state sequence, and updating the state sequence conditional on observations and parameters interactively until the iteration is converged [34]. For the sake of convenience, a cooling coil is used as an example to show this process.

1) Update Parameters Given States and Observations: To obtain the multivariate distribution of parameters, the prior distributions are required [34]. For the initial state vector, each element represents the probability of starting from a certain state and their summation is one. Dirichlet distribution represents the probabilities of K events, and the summation of these probabilities is also one. Therefore, the initial vector is assumed to follow a Dirichlet distribution:

$$\pi_{\mathrm{cc}} \sim \mathrm{Dir}(\alpha_{\mathrm{cc},1}, \ldots, \alpha_{\mathrm{cc},N}) \tag{18}$$

where $\alpha_{\mathrm{cc},i}$, $i = 1, \ldots, N$ $(N = 16)$ are distribution parameters and are usually assumed to be one [34]. Based on the prior distributions, posterior conditional distributions of parameters are updated via Bayes' rule

$$p(\pi_{\mathrm{cc}}|O_{\mathrm{cc}}) \propto \left[\prod_{i=1}^{N} f(O_{i,\mathrm{cc}}|\pi_{\mathrm{cc}})\right] p(\pi_{\mathrm{cc}}) \tag{19}$$

where $p(\pi_{\mathrm{cc}})$ is the prior distribution of parameter π_{cc}; and $f(O_{i,\mathrm{cc}}|\pi_{\mathrm{cc}})$ is the probability of an observation of the cooling coil $O_{i,\mathrm{cc}}$ when the parameter is π_{cc}. The posterior distribution $p(\pi_{\mathrm{cc}}|O_{\mathrm{cc}})$ is

$$\pi_{\mathrm{cc}}|\ldots \sim \mathrm{Dir}(n_1 + \alpha_{\mathrm{cc},1}, \ldots, n_N + \alpha_{\mathrm{cc},N}) \tag{20}$$

where parameter n_i is the number of visits to state i at time 1. Similarly, given prior distribution in [34], the posterior distributions of state transition matrices are

$$\begin{aligned}&(a_{\mathrm{cc},i1}, \ldots, a_{\mathrm{cc,iN}})|_{s_{\mathrm{sf}}=j}\\ &\sim \mathrm{Dir}(n_{\mathrm{cc},i1}|_{s_{\mathrm{sf}},=j} + \beta_{\mathrm{cc},1}, \ldots, n_{\mathrm{cc,iN}}|_{s_{\mathrm{sf}}=j} + \beta_{\mathrm{cc},N})\end{aligned} \tag{21}$$

where $n_{\mathrm{cc,mn}}|_{s_{\mathrm{sf}},=j}$ is the number of transitions from state m to n in the state sequence, while the supply fan is in state j. The observation mean corresponding to $s_{\mathrm{cc}} = i$ and $s_{\mathrm{sf}} = j$ is given as

$$\mu_i|_{s_{\mathrm{sf}}=j} \ldots \sim N\left(\bar{\mu}_i|_{s_{\mathrm{sf}}=j}, \overline{\sum}_i|_{s_{\mathrm{sf}}=j}\right) \tag{22}$$

Algorithm 1: Backwards Messages of Dynamic HMMs

Input: transition potentials P, emission potentials L
Output: HMM backwards message F
  for $t = 1, 2, \ldots, T$ **do**
    $L_{t,i} \leftarrow \exp(-(o_t - \mu_{i,s_{\mathrm{sf},t}})\Sigma^{-1}_{i,s_{\mathrm{sf},t}}(o_t - \mu_{i,s_{\mathrm{sf},t}}))$
  return L
  $B_{\mathrm{T},} :\leftarrow 1$
  for $t = T - 1, T - 2, \ldots, 1$ **do**
    $B_{t,i} \leftarrow \sum_{j=1}^{N} P_{i,s_{\mathrm{sf},t}} B_{t+1,j} L_{t+1,j}$
  return B

with

$$\bar{\mu}_i|_{s_{\mathrm{sf}}=j} = \bar{\sum}_i \left(n_i \sum_i^{-1} \bar{o}_i + \sum_{\mu}^{-1} \xi\right)|_{s_{\mathrm{sf}}=j} \tag{23}$$

$$\bar{o}_i|_{s_{\mathrm{sf}}=j} = \frac{1}{n_i} \sum_{k=1}^{n_i} o_k|_{s_{\mathrm{sf}}=j} \tag{24}$$

$$\overline{\sum}_i^{-1}|_{s_{\mathrm{sf}}=j} = \left(\sum_{\mu}^{-1} + n_i \sum_i^{-1}\right)^{-1}|_{s_{\mathrm{sf}}=j}. \tag{25}$$

Based on the prior distribution [34], the posterior distribution of the observation covariance matrix corresponding to $s_{\mathrm{cc}} = i$ and $s_{\mathrm{sf}} = j$ is the inverse Wishart distribution

$$\sum_i |_{s_{\mathrm{sf}}=j}, \ldots \sim \mathrm{IW}\left(V_{ij}^{-1} + Q_{ij}, m_{ij} + n_{ij}\right) \text{ with} \tag{26}$$

$$\begin{aligned}Q_{ij} = &\sum_{k:s_k=i} (O_k|_{s_{\mathrm{sf}}=j} - \mu_i|_{s_{\mathrm{sf}}=j})\\ &\times (O_k|_{s_{\mathrm{sf}}=j} - \mu_i|_{s_{\mathrm{sf}}=j})'.\end{aligned} \tag{27}$$

The posterior distribution of the scale matrix V is

$$V_{ij}|\cdots \sim \mathrm{IW}\left(G_{ij}^{-1} + \sum_{ij}^{-1}, m_{ij} + h_{ij}\right). \tag{28}$$

Samplings are randomly generated from posterior distributions, and regarded as estimates of parameters. Similarly, distributions of other components are obtained.

2) Update States Given Parameters and Observations: Given the posterior distributions and observations, state sequences of components are simulated. A backward recursion-forward sampling method was developed to simulate the state sequence [35]. In this method, the backward message corresponding to the probability of observing remaining observations given the state at each time is calculated by backward recursion to simulate the state sequence. The method developed in [35] does not consider fault impacts among components, and is extended to dynamic HMMs. To compute the backward message, the emission potential L representing probabilities that observations belong to different distributions is required. To capture the impacts, both means of observations of the two components are used to calculate L. The backward message of $s_{\mathrm{cc}} = i$ at time t, $B_{t,I}$, is calculated in Algorithm 1.

Given backward messages, probabilities of ending up in any states are calculated. Based on these probabilities, samples of

states are generated. Similarly, this Gibbs sampler is used to estimate states for other components.

B. Identify Fault Severities and Failure Conditions

Some faults, including: 1) tube fouling; 2) dust on fins; 3) decrease in fan efficiency; and 4) duct leakage are related to parameters that decrease or increase with faults. The fault severities are expressed in terms of decrease or increase in percentages of these parameters from their nominal values. Since they cannot be directly measured, dynamic filtering methods are used to estimate them based on measurements. The cooling coil is used here as an example. In the model of cooling coils, parameters $d_{\text{tube,in}}$ and A_{fin} decrease with tube fouling and dust on fins, and thus are considered as parametric states. The state evolution is

$$x_{\text{cc},t+1} = x_{\text{cc},t} + \upsilon_{\text{cc},t} \tag{29}$$

where $x_{\text{cc},t} = [d_{\text{tube,in},t} \quad A_{\text{fin},t}]^T$; process noise $\upsilon_{\text{cc},t} = [\upsilon_{\text{tube},t}\upsilon_{\text{fin},t}]^T$ is assumed to be normally distributed, and $\upsilon_{\text{tube},t}$ and $\upsilon_{\text{fin},t}$ are assumed to be uncorrelated. To represent the relationship between states and measurements, the measurement equation is derived from the physics-based model shown in (9)

$$z_{\text{cc},t} = h_{\text{cc}}(x_{\text{cc},t}) + w_{\text{cc},t} \text{ with} \tag{30}$$

$$z_{\text{cc},t} = \dot{m}_{a,\text{mix}} \cdot (E_{a,\text{mix}} - E_{a,\text{dis}})/\text{LMTD and} \tag{31}$$

$$h_{\text{cc}}(x_{\text{cc},t}) = \left[\frac{1}{\alpha_{a,e}\tau} + R_f + \frac{\delta_{\text{tube}}}{\lambda_{\text{tube}}} + \frac{1}{\alpha_w(d_{\text{tube,in}})}\right]^{-1} \times (A_{\text{tube,out}} + A_{\text{fin}}) \tag{32}$$

where $w_{\text{cc},t}$ is the measurement noise that is normally distributed. To estimate the states of this nonlinear model, UPF is used, since it is good at dealing with nonlinearities [36]. Since unscented KF adopts heavier tailed distributions, it is used in UPF to generate new particles around previous particles. Weights of these particles are calculated based on their likelihoods. Estimates of parameters are approximated by these particles per their weights. Similarly, to identify fault severities of fans, the fan efficiency e_{sf} is estimated. Since the fan model is a linear function with respect to e_{sf}, KF is used. To identify fault severities of duct leakage, estimates of fault-related parameters, e.g., the size of leaks, are required, but are not evident in the existing physics-based models. To address this issue, fault impacts between ducts and the supply fan are taken advantage of. If the duct leakage occurs before the supply fan, less electricity is required to maintain the airflow rate, and thus is equivalent to an increase in e_{sf}. Similarly, duct leakage after the supply fan is equivalent to a decrease in e_{sf}. Thus, the parameter e_{sf} is estimated to identify fault severities of duct leakage.

Other faults including: 1) stuck closed/open damper; 2) stuck closed/open valve; and 3) running at a fixed speed also have severities. For instance, the stuck closed/damper may be completely stuck, or move a little by following control with a lag. The former is more severe than the latter. However, these fault severities are not implemented in simulation data and real data that were used to test our method. Thus, only their failure conditions (but not their fault severities) are identified. Measurements of damper blade angles, valve openings and fan speed are used for identifying these failures.

C. Coupled-SPC to Filter False Alarms

A coupled-SPC is developed to filter out false alarms while capturing fault propagation impacts among components. In this method, a transition among normal and faulty conditions is detected if n back-to-back estimates fall outside their control limits [8]. If less than n points fall outside control limits, they are regarded as false alarms. The parameter n is set to ensure that the false alarm rate is below 0.01 while obtaining high detection rate.

To show the procedure for selecting n, the cooling coil is used as an example. The event that state estimates fall outside their control limits at time t is denoted by B_t. The probability of this event is given by

$$P(B_{t-n+1}, \ldots, B_t | s_{\text{sf},t-1}) = P(B_{t-n+1}) \prod_{j=t-n+2}^{t} P(B_j | B_{j-1}, s_{\text{sf},t-1}) \tag{33}$$

where Markovian property is used and where the estimate of the supply fan state at time t is denoted by $s_{\text{sf},t}$. The required probability $P(B_{t-n+1}|s_{\text{sf},t-1})$ is calculated via

$$P(B_{t-n+1}|s_{\text{sf},t-1}) = 1 - \sum_{\text{LCL}^w_{t-n+1}}^{\text{UCL}^w_{t-n+1}} P(s_{\text{cc},t-n+1} = int_{t-n+1} | s_{\text{sf},t-1}) \tag{34}$$

where the state estimate of the cooling coil at time t is denoted by $s_{\text{sf},t}$, and lower control limit (LCL) and upper control limit (UCL) are the lower bound and the upper bound of control limits, respectively. Integers within the control limits UCL and LCL at time t are denoted by int_t. To derive LCL and UCL, estimates of component states are assumed to follow discrete normal distributions [37]. Thus, the upper bound and low bound of control limits are

$$[\text{LCL}, \text{UCL}] = [\bar{\mu}_t - 2\bar{\sigma}_t, \bar{\mu}_t + 2\bar{\sigma}_t] \tag{35}$$

where $\bar{\mu}_t$ is the mean of previous estimates, and $\bar{\sigma}_t$ is the average standard deviation at time t. Given (34) and $P(s_{\text{cc},t-1})$, the probability $P(B_{t-n+1})$ can be calculated as

$$P(B_{t-n+1}) = \sum_{i=1}^{8} P(B_{t-n+1}|s_{\text{sf},t-1}) \cdot P(s_{\text{sf},t-1} = i). \tag{36}$$

Similarly, the probability $P(B_t|B_{t-1})$ is given by

$$\begin{aligned} &P(B_t|B_{t-1}, s_{\text{sf},t-1}) \\ &= 1 - P(\bar{B}_t|B_{t-1}, s_{\text{sf},t-1}) \\ &= 1 - \sum_{\text{LCL}^w_t}^{\text{UCL}^w_t} \\ &\times \frac{\left[1 - \sum_{\text{LCL}^w_{t-1}}^{\text{UCL}^w_{t-1}} \frac{P(s_{\text{cc},t}=\text{int}_t | s_{\text{cc},t-1}=\text{int}_{t-1}, s_{\text{fan},t-1}) \cdot P(s_{\text{cc},t-1}=\text{int}_{t-1})}{P(s_{\text{cc},t}=\text{int}_t)}\right]}{\left[1 - \sum_{\text{LCL}^w_{t-1}}^{\text{UCL}^w_{t-1}} P(s_{\text{cc},t-1} = \text{int}_{t-1})\right] / P(s_{\text{cc},t} = \text{int}_t)}. \end{aligned} \tag{37}$$

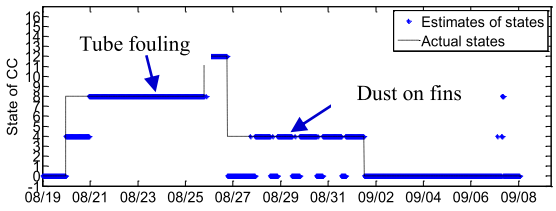

Fig. 5. State estimates of the cooling coil by using dynamic HMM.

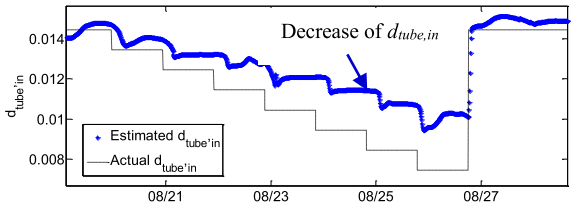

Fig. 6. Estimates of the tube inside diameter.

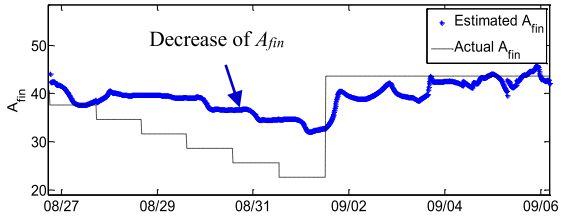

Fig. 7. Estimates of the fin surface area.

Here, the probability $P(s_{\mathrm{cc},t} = \mathrm{int}_t | s_{\mathrm{cc},t-1} = \mathrm{int}_{t-1}, s_{\mathrm{sf},t-1})$ represents the relationship between the cooling coil and the supply fan, and is obtained from the state transition matrices of dynamic HMMs. Based on (36) and (37), (33) can be calculated to determine n given the false alarm rate.

V. Experimental Results

Our fault diagnosis method is implemented in MATLAB and was run on a laptop with Intel Core i7-6920HQ 2.9 GHz processor and 32 GB of memory. To test our method, both simulation data and real data are used. In Example 1, a small building with two rooms and a VAV AHU is simulated via two simulation packages, including DesignBuilder [38] and EnergyPlus [39]. By using DesignBuilder, the building and HVAC structures were established visually. The rough simulation model was then imported in EnergyPlus to select appropriate component models and change parameters to simulate faults. Example 1 shows that: 1) failure modes are identified by dynamic HMMs; 2) dynamic HMMs perform better than HMMs; and 3) fault severities are identified by estimating parameters related to faults. In Example 2, the method is evaluated using data from ASHRAE project 1312-RP [3]. This example illustrates that: 1) failure modes and failure conditions are identified; 2) false alarms are filtered out by our coupled-SPC rule; and 3) our method performs better than existing methods.

Example 1: The simple building has two 95.517 m^3 rooms. In the building, tube diameter $d_{\mathrm{tube,in}}$ is set to be 0.01445 m; the outside surface area of fins is 43.59555 m^2; the fan efficiency is 0.7. Two gradual faults of cooling coils are considered, including: 1) tube fouling simulated from 8/20 to 8/26 and 2) dust on fins simulated from 8/26 to 9/1. Additionally, a decrease in fan efficiency is simulated from 9/2 to 9/7. The data are divided into two groups. The first group (2/3 of the data) is used for training and the other one is used for testing.

A. Identify Gradual Failure Modes and Fault Severities of Cooling Coil

By using the dynamic HMM, states of the cooling coil are estimated, as shown in Fig. 5. In this figure, actual states are marked by black dashed lines, and estimates of states are marked by blue stars. To make the figure clear to read, control limits obtained by coupled-SPC are not shown. From 8/20 to 8/26, the state estimates are "8" corresponding to $(f_{\mathrm{cc,tube}}, f_{\mathrm{cc,fin}}, f_{\mathrm{cc\_vlv,sc}}, f_{\mathrm{cc\_vlv,so}}) =$ "1000," which means that tube fouling occurs but other faults do not occur. Between 8/27 and 9/1, there are multiple actual states falsely estimated as "0" (normal), but their actual states are "4" (Dust on fins). This case is called false positive (FP) for the normal condition. Additionally, during 9/7, multiple actual states are "0," but they are estimated as "4" falsely. This case is called false negative (FN). These false estimates are caused by measurement noise. To measure fault diagnosis accuracy based on them, F-measure is used [22]

$$F\text{-measure} = \Sigma F_b / N_f \text{ with} \tag{38}$$

$$F_b = 2 \cdot \mathrm{precision} \cdot \mathrm{recall} / (\mathrm{precision} + \mathrm{recall}) \tag{39}$$

$$\mathrm{precision} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \text{ and}$$

$$\mathrm{recall} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{40}$$

where N_f is the number of failure modes; variable *TP* is true positive and *TN* is true negative, which represent true classifications. The higher the F-measure is, the better the diagnosis performance is. If the relationship between the cooling coil and the fan is considered, F-measures of $f_{\mathrm{cc,tube}}$ and $f_{\mathrm{cc,fin}}$ calculated based on (38) are 0.9695 and 0.9595.

To identify fault severities of the cooling coil, UPF is used to estimate tube inside diameter $d_{\mathrm{tube,in}}$ based on (29)–(32), and estimates are shown in Fig. 6.

In this figure, the actual tube inside diameter is marked by black dashed lines. Its estimates are represented by blue stars. The estimates gradually decrease with faults. Similarly, estimates of the fin surface area A_{fin} are estimated as shown in Fig. 7, and gradually decrease with faults. Thus, given these estimates, severities of the two faults are identified.

B. Dynamic HMMs Perform Better Than HMMs

If the HMM is used without considering the fault impacts among components, states of the cooling coil are estimated, as shown in Fig. 8. The F-measures of $f_{\mathrm{cc,tube}}$ and $f_{\mathrm{cc,fin}}$ are 0.9428 and 0.9156, which are smaller than those obtained by using dynamic HMMs (0.9695 and 0.9595). Thus, dynamic HMMs perform better than HMMs, since they represent state

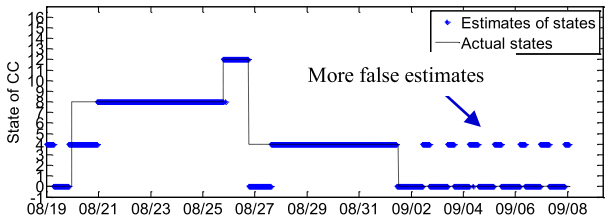

Fig. 8. State estimates of the cooling coil by using HMM.

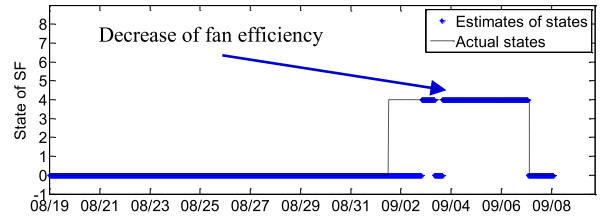

Fig. 9. Fan estimates by dynamic HMM (decreases 2.9% per day).

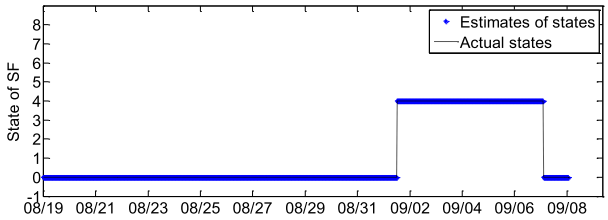

Fig. 10. Fan estimates by dynamic HMM (decreases 7.1% per day).

transitions more accurately than HMMs by considering fault propagation.

C. Identify Gradual Failure Mode and Fault Severities of Supply Fan

Decrease in supply fan efficiency is simulated from 9/2 to 9/7. The fan efficiency was reduced by 2.9% per day. The fan states are estimated using the dynamic HMM, as shown in Fig. 9. Since the degradation rate is small, there is insignificant difference between the normal condition and the faulty condition when the fault just occurs; however, the fault is detected after 1 day by the presented method. When the degradation rate is increased to 7.1%, fan states are estimated, as shown in Fig. 10. The fault is immediately detected with F-measure $= 1$ due to moderate decrease in fan efficiency.

As shown in Fig. 11, efficiency e_{sf} decreases by 7.1% per day. The estimate decreases with the fan fault, and it determines the severity of supply fan fault.

Example 2: In ASHRAE project 1312-RP, AHU-A, and AHU-B were calibrated to be identical [3]. AHU-B is fault-free, and multiple faults were implemented in AHU-A during spring, summer, and winter. This paper focuses on cooling mode, thus summer data from 8/19 to 9/8 are used, and 2/3 of the data is used for training and the 1/3 of the data is used for testing. Detailed description of data is in [3].

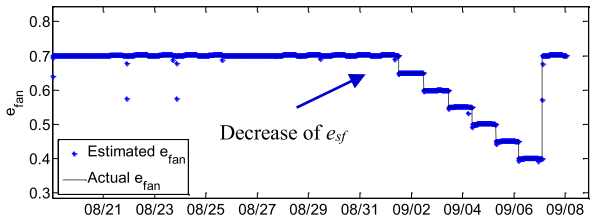

Fig. 11. Estimates of the supply fan efficiency by KF.

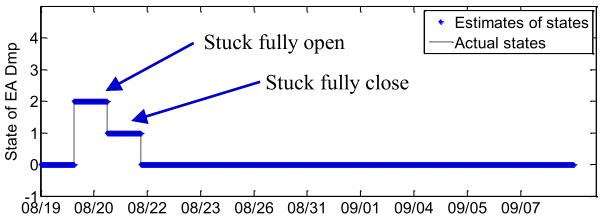

Fig. 12. State estimates of the EA damper.

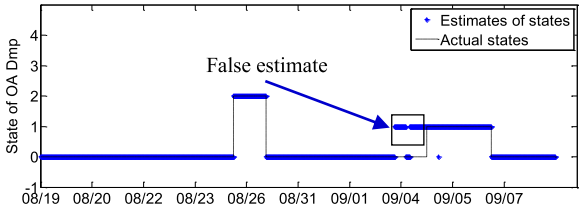

Fig. 13. State estimates of the OA damper.

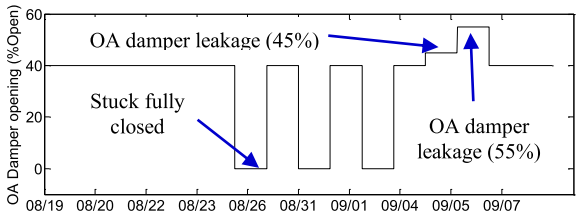

Fig. 14. OA damper opening (%).

D. Diagnose Faults of EA Damper

The damper was stuck fully open and closed on 8/20 and 8/21. By using the dynamic HMM, states of the EA damper are estimated for isolating the two failure modes, as shown in Fig. 12. F-measures of these two failure modes are calculated as 0.9981 and 1.

E. Diagnose Faults of OA Damper

Similarly, by using the dynamic HMM, states of the OA damper are estimated, as shown in Fig. 13. Based on the estimates, stuck closed damper and leakage are diagnosed with F-measures equaling 0.9992 and 0.9083. Their failure conditions are identified based on OA damper opening, as shown in Fig. 14.

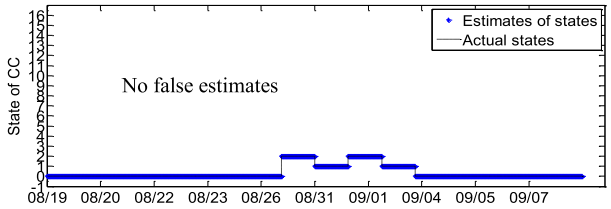

Fig. 15. State estimates of the cooling coil.

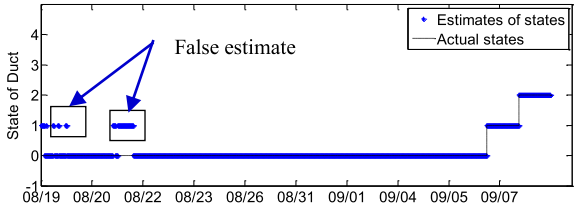

Fig. 16. Estimates of duct states.

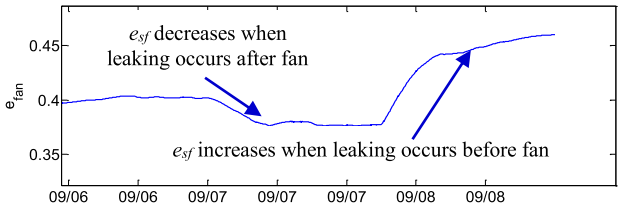

Fig. 17. Estimates of supply fan efficiency by KF.

F. Diagnose Faults of Cooling Coil

States of the cooling coil are estimated by using the dynamic HMM as shown in Fig. 15. F-measures of valve stuck open and valve stuck closed are both 1. Failure conditions are identified based on the cooling coil valve opening.

G. Diagnose Faults of Ducts

Duct states are estimated by using the dynamic HMM to identify leakage before/after the supply fan, as shown in Fig. 16. F-measures of the two failure modes are 0.9984 and 0.8550, respectively.

As discussed in Section IV-B, supply fan efficiency e_{sf} reflects the degree of duct leakage. As shown in Fig. 17, the fan efficiency decreases when the leakage occurs after the fan and increases when the leakage occurs before the fan. Thus, the decrease/increase percentage of the fan efficiency determines the fault severities of ducts.

H. Diagnose Faults of Return Fan

Fan states are estimated by using the dynamic HMM, as shown in Fig. 18.

F-measures of running at a fixed speed and complete failure are 1 and 0.9997, respectively. Their failure conditions are identified based on measurements of the fan speed.

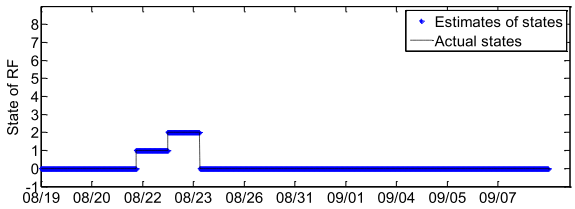

Fig. 18. State estimates of the return fan.

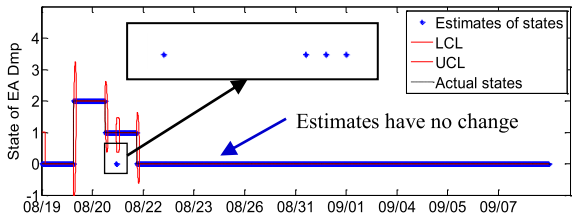

Fig. 19. False estimates of EA damper are filtered out.

I. Filter Out False Alarms

By using our coupled-SPC rule, as shown in Fig. 19, LCLs and UCLs are obtained based on previous estimates, and are marked by red lines. Most of the time, estimates have no change, and their standard deviations are 0. Thus, LCL, UCL, and estimates have the same value and appear as one. Components transit among normal and faulty conditions, and the transitions are detected if n back-to-back points fall outside their control limits. In the figure, there are one and three estimates that fall outside their control limits in sequence. Based on (33), the parameter n is calculated as six. Thus, these estimates are deemed false and are filtered out.

J. Comparison Between Our Method and Others

Many methods, e.g., SVM, decision tree, Bayesian network and ANN, have been developed to diagnose faults in AHU systems. To evaluate our method, the F-measure of each failure mode is compared with those of other methods based on ASHRAE project 1312-RP data as shown in Table I.

In this table, it can be found that the average F-measure of our method is better than the SVM [22] and the decision tree [16]. Additionally, as discussed in [22], the F-measure of their method is 0.923 that is significantly better than other methods, e.g., LibSVM, Naïve Bayes, radial basis function network, Bayesian network, NN and random forest decision tree. Therefore, the performance of our method is also better than these methods. Since our problem is not NP-hard, the computational time is not considered. Our method performs better because: 1) dynamic HMMs use different state transitions matrices depending on other components, thus more relevant knowledge on dependent fault propagation is considered; 2) residuals of different physics-based models are considered as HMM observations, and they are directly related to faults, thereby reflecting the fault impacts better than variables; and 3) coupled-SPC is developed to filter out false alarms. Additionally, compared with other methods, our method not only identifies failure modes, but also fault

TABLE I
F-MEASURES OF EACH FAILURE MODE OBTAINED BY OUR METHOD AND OTHERS

	Our method	SVM [22]	Decision tree [16]
$f_{ea\_dmp,sc}$	1	0.923	0.9
$f_{ea\_dmp,so}$	0.9981	0.923	NA
$f_{oa\_dmp,sc}$	0.9992	NA	NA
$f_{oa\_dmp,leak}$	0.9083	0.994	NA
f_{cc,vlv_sc}	1	0.928	1
f_{cc,vlv_so}	1	0.785	0.98
$f_{duct,lb}$	0.9984	0.981	1
$f_{duct,la}$	0.8550	1	NA
$f_{rf,cf}$	0.9997	0.887	1
$f_{rf,fs}$	1	0.795	1
Average	0.9759	0.923	0.97

severities. Thus, it helps maintenance crews to know which faults have occurred, how critical they are, and be guided in the repair process to improve the HVAC system availability.

VI. CONCLUSION

This paper presented a method for integrating dynamic HMMs, KF/UPF and coupled-SPC to identify failure modes and fault severities of air handling systems, while considering fault propagation impacts among components. Contributions of our work are: 1) developing a method to capture the fault propagation impacts across components to identify the failure modes; 2) opening an effective way to measure and identify fault severities of devices; and 3) deriving a coupled-SPC to filter out false alarms, while capturing the cross-component impacts. For a few coupled components, our method performs well. However, if the method is concerned with many coupled components, high-computational effort is required. In the future, our method will be extended to faults of many coupled components, e.g., VAV boxes. Additionally, sensor faults will be investigated.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] K. Bruton, D. Coakley, P. Raftery, D. O. Cusack, M. M. Keane, and D. T. J. O'Sullivan, "Comparative analysis of the AHU InFO fault detection and diagnostic expert tool for AHUs with APAR," *Energy Efficiency*, vol. 8, no. 8, pp. 299–322, 2015.

[2] *Building Optimization and Fault Diagnosis Source Book*, IEA ANNEX, Tech. Res. Centre Finland, VTT Building Technol., Finland, 1996.

[3] S. Li and J. Wen, "Description of fault test in summer of 2007," ASHRAE, Tech. Rep. 026, Sep. 2007.

[4] A. Kodali, K. Pattipati, and S. Singh, "Coupled factorial hidden Markov models (CFHMM) for diagnosing multiple and coupled faults," *IEEE Trans. Syst., Man, Cybernetics, Syst.*, vol. 43, no. 3, pp. 522–534, May 2013.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–285, Feb. 1989.

[6] B. Sun, P. B. Luh, Q. S. Jia, Z. O'Neill, and F. Song, "Building energy doctors: An SPC and Kalman filter-based method for system-level fault detection in HVAC systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 1, pp. 215–229, Jan. 2014.

[7] S. Katipamula and M. R. Brambley, "Methods for fault detection, diagnostics, and prognostics for building systems—A review, Part I," *Int. J. HVAC&R Res.*, vol. 11, no. 1, pp. 3–25, 2005.

[8] L. K. Norford, J. A. Wright, R. A. Buswell, D. Luo, C. J. Klaassen, and A. Suby, "Demonstration of fault detection and diagnosis methods for air-handling units," *Int. J. HVAC&R Res.*, vol. 8, no. 1, pp. 41–71, 2002.

[9] T. I. Salsbury and R. C. Diamond, "Fault detection in HVAC systems using model-based feedforward control," *Energy Buildings*, vol. 33, no. 4, pp. 403–415, 2001.

[10] P. Haves, T. Salsbury, and J. A. Wright, "Condition monitoring in HVAC subsystems using first principles models," *ASHRAE Trans.*, vol. 102, no. 1, pp. 519–527, 1996.

[11] H. Yoshida, T. Iwami, H. Yuzawa, and M. Suzuki, "Typical faults of air-conditioning systems and fault detection by ARX model and extended Kalman filt," *ASHRAE Trans.*, vol. 102, no. 1, pp. 557–564, 1996.

[12] J. E. Seem and J. M. House, "Evaluation of an AHU fault detection scheme based on finite state machine sequencing control," Johnson Controls Inc. Iowa Energy Center, Iowa, USA, Tech. Rep., Jan. 2007.

[13] M. R. Brambley *et al.*, "Final project report: Self-correcting controls for VAV system faults Filter/Fan/Coil and VAV box sections," U.S. Dept. Energy, Pacific Northwest Nat. Lab., Oak Ridge, TN, USA, Tech. Rep. PNNL-20452, May 2011.

[14] G. M. Kaler, "Expert system predicts service," *Heating, Piping, Air Conditioning*, vol. 60, no. 11, pp. 99–101, 1988.

[15] S. Katipamula, R. G. Pratt, D. P. Chassin, Z. T. Taylor, K. Gowri, and M. R. Brambley, "Automated fault detection and diagnostics for outdoor-air ventilation systems and economizers: Methodology and results from field testing," *ASHRAE Trans.*, vol. 105, no. 1, pp. 1–13, 1999.

[16] J. House, M. Vaezi-Nejad, and J. M. Whitcomb, "An expert rule set for fault detection in air-handling units," *ASHRAE Trans.*, vol. 107, no. 1, pp. 858–871, 2001.

[17] J. Schein, S. T. Bushby, N. S. Castro, and J. M. House, "A rule-based fault detection method for air handling units," *Energy Buildings*, vol. 38, no. 12, pp. 1485–1492, 2006.

[18] R. Yan, Z. Ma, T. Zhao, and G. Kokogiannakis, "A decision tree based data-driven diagnostic strategy for air handling units," *Energy Buildings*, vol. 133, pp. 37–45, Dec. 2016. [Online]. Available: http://dx.doi.org/doi:10.1016/j.enbuild.2016.09.039

[19] S. Wang and F. Xiao, "AHU sensor fault diagnosis using principal component analysis method," *Energy Buildings*, vol. 36, no. 2, pp. 147–160, 2004.

[20] S. Li and J. Wen, "A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform," *Energy Buildings*, vol. 68, pp. 63–71, Jan. 2014.

[21] H. C. Peitsman and V. Bakker, "Application of black-box models to HVAC systems for fault detection," *ASHRAE Trans.*, vol. 102, no. 1, pp. 628–640, 1996.

[22] Z. Du, B. Fan, X. Jin, and J. Chi, "Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis," *Building Environ.*, vol. 73, pp. 1–11, Mar. 2014.

[23] W. Y. Lee, J. M. House, C. Park, and G. E. Kelly, "Fault diagnosis of an air-handling unit using artificial neural networks," *ASHRAE Trans.*, vol. 102, no. 1, pp. 540–549, 1996.

[24] T. Mulumba, A. Afshari, K. Yan, W. Shen, and L. K. Norford, "Robust model-based fault diagnosis for air handling units," *Energy Buildings*, vol. 86, pp. 698–707, Jan. 2015.

[25] H. M. Ertunc, K. A. Loparo, and H. Ocak, "Tool wear condition monitoring in drilling operations using hidden Markov models (HMMs)," *Int. J. Mach. Tools Manuf.*, vol. 41, no. 9, pp. 1363–1384, 2001.

[26] I. Rezek, P. Sykacek, and S. J. Roberts, "A comparison of Bayesian and maximum likelihood learning of coupled hidden Markov models," *IEE Proc. Sci. Technol. Meas.*, vol. 147, no. 6, pp. 345–350, 2000.

[27] Y. H. Chen and M. R. Gupta, "EM demystified: An expectation-maximization tutorial," Dept. Elect. Eng., Univ. Washington, Seattle, WA, USA, Tech. Rep. UWEETR-2010-0002, 2010.

[28] J. Gronbaek, H.-P. Schwefel, A. Ceccarelli, and A. Bondavalli, "Improving robustness of network fault diagnosis to uncertainty in observations," in *Proc. 9th IEEE Int. Symp. Netw. Comput. Appl.*, Cambridge, MA, USA, Jul. 2010, pp. 229–232.
[29] P. Liu, D. Wang, and Z. Xu, "A method for the fault prediction of printing press based on statistical process control of registration accuracy," *J. Inf. Comput. Sci.*, vol. 10, no. 17, pp. 5579–5587, 2013.
[30] P. E. Pfeifer. (Oct. 21, 2008). *Chi-Square Goodness-of-Fit Test, Darden Case UVA-QA-0692*. [Online]. Available: http://ssrn.com/abstract=1284265
[31] Q. S. Yan, W. X. Shi, and C. Q. Tian, *Refrigeration Technology for Air Conditioning*. China, Beijing: Construction Industry Press, 2010.
[32] T. K. Hong and R. L. Webb, "Calculation of fin efficiency for wet and dry fins," *Int. J. HVAC&R Res.*, vol. 2, pp. 27–41, 1996.
[33] J. Wen and S. Li, "RP-1312–Tools for evaluating fault detection and diagnostic methods for air-handling units," ASHRAE, Tech. Rep., 2012.
[34] T. Ryden, "EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective," *Bayesian Anal.*, vol. 3, no. 4, pp. 659–688, 2008.
[35] M. J. Johnson, "Bayesian time series models and scalable inference," M.S. thesis, Dept. Elect. Eng., MIT, Cambridge, MA, USA, 2014.
[36] R. Dermerwe, A. Doucet, N. Freitas, and E. Wan, "The unscented particle filter," Eng. Dept., Cambridge Univ., Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR 380, 2000.
[37] D. Roy, "The discrete normal distribution," *Commun. Statist.-Theory Methods*, vol. 32, no. 10, pp. 1871–1883, 2003.
[38] DesignBuilder Software. (Oct. 2009) *DesignBuilder 2.1 User-Manual*. [Online]. Available: http://www.designbuildersoftware.com/docs/designbuilder/DesignBuilder_2.1_Users-Manual_Ltr.pdf
[39] (Mar. 2016). *EnergyPlus Engineering Reference*. [Online]. Available: https://energyplus.net/sites/all/modules/custom/nrel_custom/pdfs/pdfs_v8.5.0/EngineeringReference.pdf

Ying Yan (S'12) received the B.S. degree in automation and the M.S. degree in control theory and control engineering from Southeast University, Nanjing, China, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of Connecticut, Storrs, CT, USA.

His current research interests include fault detection, diagnosis and prognosis of heating, and ventilation and air conditioning systems.

Peter B. Luh (S'77–M'80–SM'91–F'95) received the B.S. degree from National Taiwan University, Taipei, Taiwan, the M.S. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the Ph.D. degree from Harvard University, Cambridge.

Since 1980, he has been with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT, USA, where he is currently the SNET Professor of Communications and Information Technologies. He is also a member of the Chair Professors Group, Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing, China. His current research interests include smart, green and safe buildings; smart grid, electricity markets, and effective renewable integration to the grid; and intelligent manufacturing systems.

Dr. Luh was a recipient of the 2013 Pioneer Award of the IEEE Robotics and Automation Society for his pioneering contributions to the development of near-optimal and efficient planning, scheduling, and coordination methodologies for manufacturing and power systems. He was the founding Editor-in-Chief of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, and the Editor-in-Chief of the IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION.

Krishna R. Pattipati (S'77–M'80–SM'91–F'95) received the B.Tech. (with highest Hons.) degree in electrical engineering from the Indian Institute of Technology Kharagpur, Kharagpur, India, in 1975, and the M.S. and Ph.D. degrees in systems engineering from the University of Connecticut (UCONN), Storrs, CT, USA, in 1977 and 1980, respectively.

From 1980 to 1986, he was with Alphatech, Inc., Burlington, MA, USA. He has been with the Department of Electrical and Computer Engineering, UCONN, where he is currently the Board of Trustees Distinguished Professor and the UTC Chair Professor of Systems Engineering. He is Co-Founder of Qualtech Systems, Inc., Rocky Hill, CT, USA, a firm specializing in advanced integrated diagnostics software tools (TEAMS, TEAMS-RT, TEAMS-RDS, and TEAMATE), and serves on the Board of Aptima, Inc., Woburn, MA, USA. His current research interests include proactive decision support, uncertainty quantification, smart manufacturing, autonomy, knowledge representation, and optimization-based learning and inference. A common theme among these applications is that they are characterized by a great deal of uncertainty, complexity, and computational intractability.

Dr. Pattipati was a co-recipient of the Andrew P. Sage Award for the Best SMC Transactions Paper for 1999, the Barry Carlton Award for the Best AES Transactions Paper for 2000, the 2002 and 2008 NASA Space Act Awards for "A Comprehensive Toolset for Model-based Health Monitoring and Diagnosis," and "Real-Time Update of Fault-Test Dependencies of Dynamic Systems: A Comprehensive Toolset for Model-Based Health Monitoring and Diagnostics," the 2003 AAUP Research Excellence Award at UCONN, and the Centennial Key to the Future Award. He was selected by the IEEE Systems, Man, and Cybernetics Society as the Outstanding Young Engineer of 1984. He has served as the Editor-in-Chief of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART B from 1998 to 2001. He is an elected Fellow of the Connecticut Academy of Science and Engineering.