International Journal of Heat and Mass Transfer 57 (2013) 698–707

Contents lists available at SciVerse ScienceDirect

International Journal of Heat and Mass Transfer

journal homepage: www.elsevier.com/locate/ijhmt

Error estimation in POD-based dynamic reduced-order thermal modeling of data centers

Rajat Ghosh, Yogendra Joshi *

The George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0405, United States

ARTICLE INFO

Article history:
Received 28 May 2012
Received in revised form 25 September 2012
Accepted 3 October 2012

Keywords:
Reduced-order model
Transient data center
Proper orthogonal decomposition
Extrapolation
Analytical error

ABSTRACT

A proper orthogonal decomposition (POD)-based reduced-order modeling framework that predicts transient air temperatures in an air-cooled data center is developed. The framework is applied on an initial temperature data set acquired by measurements at discrete time instants. The subsequent data analysis predicts air temperatures for times both inside and outside of the discrete time domain. The predicted temperature data are compared with corresponding experimental observations, and the prediction error is analyzed. An alternative analytical approach is developed for determining the error, and an iteration-based optimization procedure is developed to calibrate the analytical error against the POD-based modeling error. The calibrated analytical error is added to the corresponding POD-predicted temperature data to obtain reliable new temperature data.

1. Introduction

With proliferation of cloud-computing-based e-commerce services, data centers are emerging as prominent economic hubs. Data centers, consuming as much as 2% of the total electricity produced in the United States, are posing a serious engineering challenge in the form of sustainable design [1]. Rapidly increasing numbers of data centers are facing the problem with increasing power demands in constrained capacity because most of these data centers were designed for high performance and uninterrupted availability with little attention to undesirable heat removal. With increasing power density in data centers (~O(10 GW)) [1], benchmarking studies reveal that the cooling infrastructure in a data center consumes as much as 20–50% of the total facility energy usage, and the life-cycle cost of cooling is becoming comparable to that of information technology (IT) hardware for commodity computing [2]. The energy efficiency of data centers can be significantly improved by implementing the accepted "best practices" [3] and the data center infrastructure management (DCIM)-based technologies [4] which streamlines the dynamic energy usage during various transient events in data centers. Common events in data centers related to dynamic energy usage are transients that stem from time-varying IT heat loads, equipment upgrades, and changes in set-points of computer room air conditioning (CRAC) units. For the optimization of dynamic allocation of cooling resources in a data center, a particularly useful capability is a model-based real-time thermal control framework [5] which is presently stymied by the lack of an efficient modeling algorithm.

Most of the state-of-the-art temperature modeling techniques for data centers rely on computational fluid dynamics and heat transfer (CFD/HT)-based numerical simulations [6]. However, since the turbulent multi-scale convective heat transfer inside a data center facility involves millions of degrees of freedom (DOFs) with several decades of length scales and time scales [7], the CFD/HT-based modeling is computationally inefficient for a model-based dynamic thermal control framework. Alternatively, an emerging modeling trend includes physics-based reduced-order models, which attempt to efficiently capture dominant energy components. Mathematical techniques for filtering dominant components are non-linear Volterra theory [8], harmonic balance approximation [9], and proper orthogonal decomposition (POD) [10]. Non-linear Volterra theory models a non-linear system as an infinite sum of multi-dimensional convolution integrals in an increasing order. On the other hand, the harmonic balance approximation relies on frequency domain techniques to model a non-linear system. Lastly, POD, a model reduction technique particularly useful for resolving stochastic processes such as turbulent convective heat transfer in an air-cool data center, depends upon finding optimal basis: the N optimal basis capture more energy than any other N basis. The POD-based technique is widely published in the literature for modeling convective air temperature field in data centers [11]. Additional techniques involving simplifications in physical description include potential flow modeling [12] and thermal zone-based modeling [13]. Although several reduced-order modeling frameworks as discussed above exist for a turbulent convective thermal

* Corresponding author.
E-mail address: yogendra.joshi@me.gatech.edu (Y. Joshi).

0017-9310/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijheatmasstransfer.2012.10.013

Nomenclature

(x, y, z)	spatial co-ordinates
t	time
$T_i(x, y, z; t_i)$	transient temperature data with time as the parameter
$T_0(x, y, z)$	parameter-independent component of temperature data
$T_i^*(x, y, z; t_i)$	parameter-dependent component of temperature data
C.E.P	captured energy percentage
b	POD coefficients
λ	eigenvalue
k	the number of principle components
e_{int}	interpolation error bound
c_0	optimization constant for interpolation
e_{ext}	extrapolation error bound
t_i	extrapolation start time
t_f	extrapolation end time
α	thermal diffusivity
$\vec{u}$	velocity field
M.A.R.	modeling accuracy requirement
E.H.	extrapolation horizon
$\dot{q}$	volumetric heat generation rate
E_H	Eddy diffusivity for heat transfer

system like an air-cooled data center, a lack of error bounds stymies the model-based control of data center cooling resources.

Using a POD-based framework, this paper studies the transient convective air temperature field ensuing an experimentally-simulated CRAC start-up in a data center laboratory. From the measured temperature data, the POD-based analysis is performed on an ensemble of temperatures collected at discrete times within the measurement domain. While the POD/interpolation framework is validated for its predictive capability within the ensemble time domain, the POD/extrapolation framework is validated for the time instants greater than the maximum value of the time domain at which the temperature ensemble is constructed. The paper analyzes the prediction error between the experimental data and the POD-model predictions. In addition, an analytical methodology is developed for determining the error incurred in the POD/interpolation and POD/extrapolation frameworks. The analytical error has been formulated using the analysis outlined in [14]. In an intermediate step, the approach developed in [14] uses the finite element method which involves the discretization of governing differential equations. In contrast, this paper utilizes experimental data measured by a grid-based data acquisition system. The experimental discrete data are utilized with pertinent time and length scales, and an analysis similar to [14] is used to derive the analytical error. The analytical error is fully determined by using an iteration-based optimization procedure which involves the calibration of the analytical error against the prediction error. The calibrated analytical error can be added to the POD prediction to derive a priori temperature predictions.

2. POD-based dynamic model for transient air temperatures

Fig. 1 describes the POD algorithm used in the present study. First, a temperature ensemble, with time as the parameter, $T_i(x, y, z; t_i) \in \mathrm{R}^{m\times n}$ is generated from physical experiments. The temperature ensemble consists of n observations of the m-dimensional transient temperature field, taken at n different time instants, $t_i (i = 1, 2, \ldots, n)$. Then, the time-independent part of the temperature ensemble, T_0, is determined by taking the row-wise average of T_i. The time-dependent ensemble is obtained by subtracting the time-averaged temperature, T_0, from every observation in T_i.

$$T_0(x, y, z) = \frac{\sum_{i=1}^{n} T_i(x, y, z; t_i)}{n}. \tag{1}$$

$$T_i^*(x, y, z; t_i) = T_i(x, y, z; t_i) - T_0(x, y, z). \tag{2}$$

The time-dependent part of the temperature ensemble, T_i^*, is modeled as a sum of product of POD modes, $\psi^{m\times n}$, and POD coefficients, $b^{n\times n}$. POD modes represent basis functions of the transient component of the temperature ensemble. They are determined by solving a constrained optimization problem for maximizing its projection $\langle |T^*, \psi|^2 \rangle$ with a constraint $\|\psi\|^2 = 1$. The corresponding functional for this constrained variational problem is:

$$J(\psi) = \langle |T^*, \psi|^2 \rangle - \lambda(\|\psi\|^2 - 1). \tag{3}$$

A necessary condition for the optimization suggests that the functional derivative of $J(\psi)$ vanishes with all variations $\psi + \delta\theta \in L^2([0, 1]), \delta \in R$

$$\frac{d}{d\delta} J[\psi + \delta\theta]_{\delta=0} = 0. \tag{4}$$

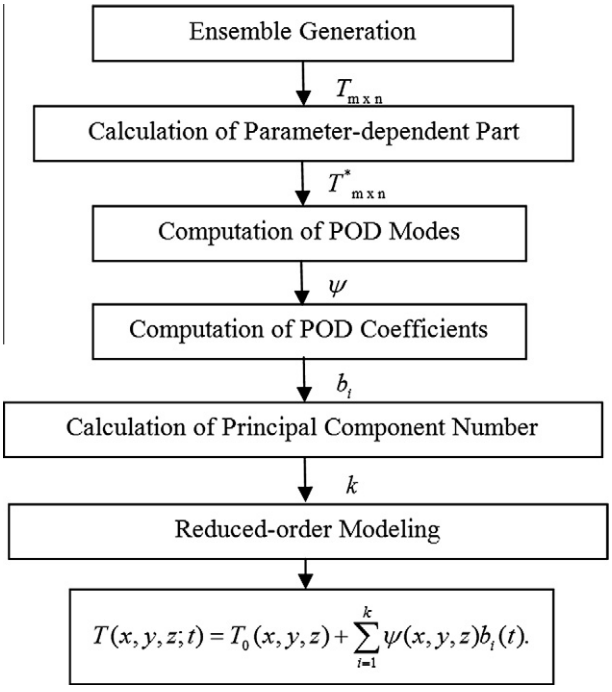

Fig. 1. POD-based dynamic reduced-order modeling algorithm with time as the parameter. The algorithm is applied on an ensemble of transient temperatures, T_{mxn}., acquired at a discrete set of time instants, $t_i (i = 1, 2, \ldots, n)$, called snapshot instants. First, the parameter-dependent part, T^*_{mxn}, is calculated. Then, by applying singular value decomposition, POD modes, ψ_i, are computed. While at snapshot instants, POD coefficients are calculated via matrix inversion; at intermediate times, POD coefficients are computed via interpolation, and outside of the snapshot time domain, they are computed via extrapolation. The principal components are determined via 99% eigenvalue-fraction criterion (Eq. (10)). Finally, the transient temperature field is constructed with a better temporal resolution.

Simplification [15,16] of Eq. (4) leads to the governing equation for POD modes:

$$R\psi = \lambda\psi, \tag{5}$$

where,

$$R = \frac{1}{m}(T^*)^{Tr}T^*. \tag{6}$$

The superscript '*Tr*' denotes the Transpose of the matrix.

After determining POD modes by solving Eq. (5), corresponding POD coefficients are determined via matrix inversion.

$$b^{n\times n} = inv(\psi^{m\times n})T^{*m\times n}. \tag{7}$$

Eq. (5) is an eigenvalue equation with eigenvalues , λ_i, with the following property:

$$\lambda_i > \lambda_j; i > j. \tag{8}$$

The relative energy content of a POD mode , E_i, is estimated by the fraction of corresponding eigenvalue, λ_i :

$$E_i = \frac{\lambda_i}{\sum_{i=1}^{n}\lambda_i}. \tag{9}$$

Using monotonically increasing eigenvalues, the number of principal components , k, can be calculated such that:

$$\left(\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}\lambda_i} > \text{C.E.P.}\right)\bigcap(\min(k)), \tag{10}$$

where, C.*E.P.* is defined as the captured energy percentage by the reduced-order framework. Eq. (10) states that $k \in [1, n]$ is the minimum value of a set of natural numbers for which the sum of the fraction of eigenvalues are greater than C.*E.P.* Eq. (10) implies the temperature field can be constructed by including only $k(<n)$ POD modes instead of n(ensemble size) modes. The numerical value of C.*E.P.* can be chosen based on the modeling accuracy requirement (M.A.R.): for a crude model, C.E.P. can be chosen as low as 80% (or even lower); while for a high-fidelity model, C.E.P. can be as high as 99.5% (or even higher).

While POD modes are time-independent, POD coefficients represent the time-dependent part of the temperature field. POD coefficients determined by Eq. (7) correspond to the ensemble time domain: the time instants at which the temperature ensemble, T_i, is generated. With the parameter of the problem as time, the generalized POD coefficients can be found by either interpolation or extrapolation in time. Although the Galerkin projection-based technique [15], alternative to the interpolation or extrapolation-based technique, exists for the determination of POD coefficients, the interpolation/extrapolation-based technique is particularly suitable for the analysis of experimental data collected from a stochastic process [17–20] e.g. the transient evolution of a turbulent temperature field. If the desired time instant is intermediate to time instants at which the ensemble was created, the interpolation-based numerical technique is used. Otherwise, the extrapolation-based numerical technique is used. As POD coefficients are determined for a desired time instant, Eq. (11) models the temperature field at that time instant:

$$T(x,y,z,t) = T_0(x,y,z) + \sum_{i=1}^{k}\psi(x,y,z)b(t). \tag{11}$$

POD/interpolation or POD/extrapolation-based numerical schemes have their inherent prediction errors ($E_{\text{Prediction}}$) which are quantified as the deviations between the experimental data ($T_{\text{Experiemnt}}$) and the corresponding POD-based predictions (T_{POD}), shown in Eq. (12):

$$E_{\text{Prediction}} = (T_{\text{Experiment}} - T_{\text{POD}}). \tag{12}$$

At a new time instant other than the time instants at which the temperature ensemble is compiled, the prediction error ($E_{\text{Prediction}}$) is computed a posteriori: starting from a temperature ensemble $T_i(x, y, z; t_i)$, POD/interpolation framework computes new temperature data $T_p(x, y, z; t_p)$ where $\min(t_i) < t_p < \max(t_i)$, and POD/extrapolation framework computes new temperature data $T_q(x, y, z; t_q)$ where $\max(t_i) < t_q <$ E.*H.* E.H is the acronym for the extrapolation horizon which is defined as the extent of time beyond the $\max(t_i)$ when the POD/extrapolation framework breaks down. The estimation of E.*H.* for a reliable temperature extrapolation depends on the corresponding prediction error, $E_{\text{Prediction}}$, against which the analytical error, $E_{\text{Analytical}}$, is calibrated. It is proposed a reliable extrapolation window should not go beyond a time instant for which $E_{\text{Prediction}}$ reaches a critical limit. The critical limit for the breakdown of the extrapolation scheme is assigned to be less than a fraction of the scale of the temperature difference which is a characteristic of the thermal system. Quantitatively, E.H. is a time instant till which $E_{\text{Prediction}}$ satisfies following condition:

$$E_{\text{Prediction}} \leqslant f\Delta T_{\text{Scale}}^{\text{Measurement}}, \tag{13}$$

Where, f is the scale factor, which can be any numerical value between 0 and 1.

After finding T_{POD}, the computation of the prediction error ($E_{\text{Prediction}}$) requires corresponding experimental data. Such a posteriori measurement requirement makes the framework sluggish and expensive, particularly for being useful as a dynamic control algorithm. As an alternative to the prediction error ($E_{\text{Prediction}}$), the analytical error ($E_{\text{Analytical}}$) is formulated. The analytical error ($E_{\text{analytical}}$) is defined as the difference between the exact solution of the governing equation for air temperatures (T_{Exact}) and the corresponding POD-based prediction (T_{POD}).

$$E_{\text{Analytical}} = (T_{\text{Exact}} - T_{\text{POD}}). \tag{14}$$

The analytical error ($E_{\text{Analytical}}$) can be determined by semi-analytical methods described in [14,21]. The detailed methodology is outlined in the Appendices A-B.

For the POD/interpolation scheme, $E_{\text{Analytical}}^{\text{POD/Interpolation}}$, is given by:

$$E_{\text{Analytical}}^{\text{POD/Interpolation}} = c_0\sum_{i=k+1}^{n}\lambda_i, \tag{15}$$

where c_0 is an empirical constant. Eq. (15) essentially means the sum of the discarded eigenvalues manifests as the error.

For the POD/extrapolation scheme, $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$, is given by:

$$E_{\text{Analytical}}^{\text{POD/Extrapolation}} = c_1(\sigma^{-1}(t_m)k^t + h^p) + \left[\frac{l\exp(\theta)}{1-\theta}(c_2 + c_3k)\sum_{n=d+1}^{m}\lambda_n\right]^{\frac{1}{2}}. \tag{16}$$

where, $\Delta t(=t_f - t_i)$ is the extrapolation interval, $\theta(=t/t_f)$ is non-dimensional time; and c_1, c_2, c_3 are empirical constants. $\sigma^{-1}(t_m) = \min(1, t)$. k^t: = Experimental time scale. h^p: = Experimental length scale. l: = Dimension of the ensemble set.

It is apparent from Eqs. (15) and (16) that complete determinations of $E_{\text{Analytical}}^{\text{POD/Interpolation}}$ and $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$ require optimal numerical values of for the arbitrary constants c_0 and (c_1, c_2, c_3). It is obvious that the numerical values of these constants depend on the specific initial data. Therefore, the numerical values of these constants are determined via a statistical optimization procedure. The central philosophy of this procedure is that the fractional difference between $E_{\text{Analytical}}$ and $E_{\text{Prediction}}$ is optimally minimized for the different values of optimization parameter(s): c_0 for the POD/interpolation framework, and (c_1, c_2, c_3) for the POD/extrapolation framework. The fractional difference between $E_{\text{Analytical}}$ and $E_{\text{Prediction}}$ is defined as the error functional (e):

$$e = \frac{abs(E_{\text{Prediction}} - E_{\text{Analytical}})}{abs(E_{\text{Prediction}})}. \quad (17)$$

For the POD/Extrapolation framework, the optimization problem is:

$$\min[e(c_0)], c_0 \in \mathrm{R}. \quad (18)$$

For the POD/Extrapolation framework, the optimization problem is:

$$\min[e(c_1, c_2, c_3)], (c_1, c_2, c_3) \in \mathrm{R}. \quad (19)$$

$E_{\text{Prediction}}$, $E_{\text{Analytical}}$, and e are multi-dimensional vectors. The minimization of e is conducted statistically: for a given c_0 or (c_1, c_2, c_3), e is calculated. Thereafter, average (μ) and standard deviation (σ) across the various dimensions of e are calculated:

$$\mu = \frac{\sum_i e_i}{\dim(e)}. \quad (20)$$

$$\sigma = \left(\frac{1}{\dim(e) - 1}\sum_i (e_i - \mu)^2\right)^{\frac{1}{2}}. \quad (21)$$

A low value of μ suggests that average values $E_{\text{Prediction}}$ and $E_{\text{Analytical}}$ are proximal to each other. On the other hand, a low value of σ suggests the difference between $E_{\text{Prediction}}$ and $E_{\text{Analytical}}$ does not deviate much from μ. A low μ together with a low σ suggests $E_{\text{Analytical}}$ tends to approximate $E_{\text{Prediction}}$ within a confidence interval determined by μ. Such an approximation will obviate the necessity of a posteriori experimental measurements for estimating the validity of the POD-based framework. By the definition in Eq. (14), T_{POD} can be directly added to $E_{\text{analytical}}$ to obtain a temperature value whose accuracy depends upon the quality of the optimization procedure. For difference values c_0 and (c_1, c_2, c_3), different μ and σ can be obtained. The relative importance of μ and σ in the optimization framework can be mathematically quantified by a weighting factor, ω. Finally, to choose optimal values of c_0 and (c_1, c_2, c_3), a unified decision-making index (I) can be modeled:

$$I = \omega\mu + (1 - \omega)\sigma. \quad (22)$$

For various choices of c_0 (for POD/interpolation) or (c_1, c_2, c_3) (for POD/extrapolation), the choice that makes I smallest gives the chosen parameter(s).

3. Experimental setup

As shown in Fig. 2, the experimental setup is a data center that employs a raised floor plenum supply and overhead plenum return air flow scheme. The servers and other IT equipment are mounted in cabinets, or racks, on a raised floor. An alternating "cold aisle" and "hot aisle" configuration is employed, where the inlet side of the servers faces a cold aisle, and the outlet side faces a hot aisle. The computer room air conditioning (CRAC) unit supplies pressurized cold air into the underfloor plenum. The cold air flows up through perforated tiles, and is entrained into the servers by server fans. The hot air from the server outlets is cooled by chilled water circulating in air-to-water and rear-door heat-exchangers mounted on the rear cabinet doors prior to discharge into the hot aisle. It then returns to the CRAC though an overhead plenum for further cooling to the supply temperature. Fig. 3 shows the plan view of the experimental setup which is populated with 16 standard size server cabinets or racks of height: 2134 mm, depth: 1067 mm, and width: 584 mm. The racks are arranged in an 8×2 architecture with alternating cold and hot aisles. The facility has three CRAC units. However, in the present case study, CRAC-1 is the only active unit which supplies cooling air at 4.6 kg/s at its 100% capacity. Additional pertinent specifications, including the hardware housed within the racks and their power dissipations are listed in Table 1. Racks are numbered as R-I, where I = 1 – 16.

3.1. Air temperature measurement

A grid-based [22] thermocouple network deployed in a 3-D telescopic mechanism measures air temperatures between racks, R-5 and R-6. Fig. 4 shows the thermocouple grid unit. In the cold aisle, the positional uncertainty of the thermocouple grid is 25.4 mm(1"). As shown in Fig. 4, the thermocouples are arranged in a square grid and the grid consists of 21 T-type copper-constantan thermocouples made from 28 gauge wire (0.9 mm diameter). Each thermocouple provides a response time around 20 ms, which is suitable for rack-level air temperature measurements (time scale of ~1 s). The temperature data obtained are processed by a National Instruments digital data acquisition system (http://www.ni.com/pdf/manuals/373344a.pdf) and subsequently transmitted to I/O terminals by a Cisco© WET200 Wireless-G Business Ethernet Bridge router. With 100 Mb/s digital data acquisition rate, a LabVIEW©-based measurement capability that can capture transient temperatures from the entire array at a frequency of 1 Hz is

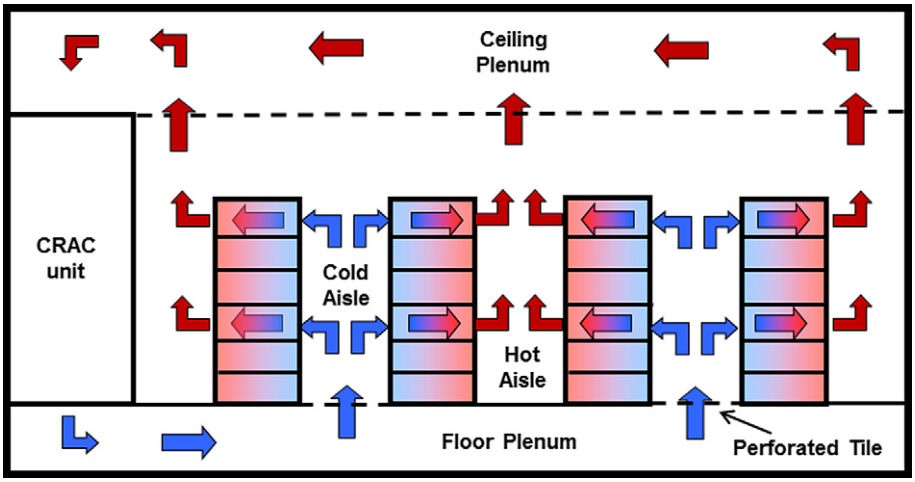

Fig. 2. Layout of cooling air distribution in a typical data center. Servers and other IT equipment are mounted in racks on a raised floor. The racks are arranged in alternating cold/hot aisles. While the server inlets face cold aisles, the exhaust sides face hot aisles. Cooling air is blown by the CRAC unit into the plenum and provided through perforated tiles to the server inlets. Heated exhaust air is returned to the CRAC unit via the overhead plenum.

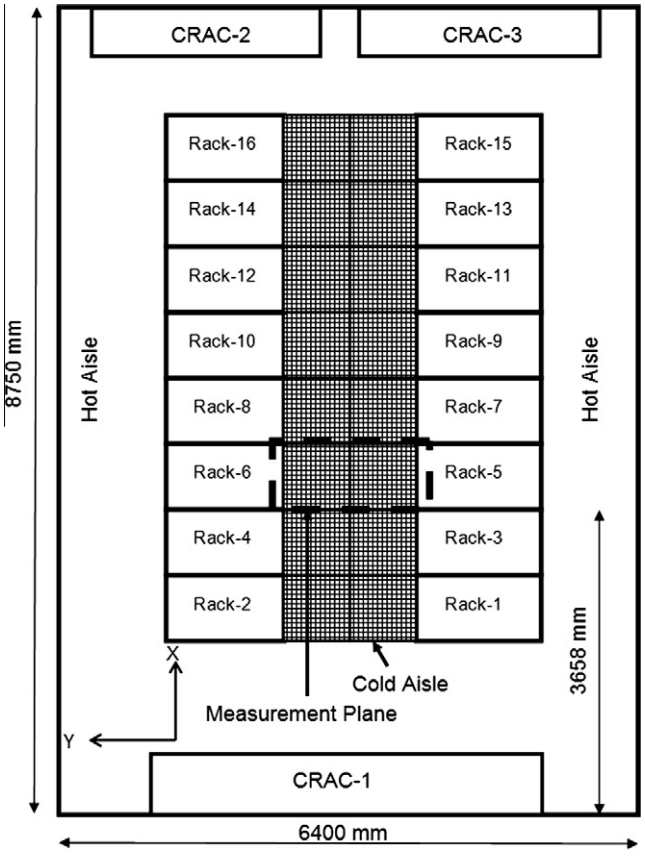

Fig. 3. Plan view of the experimental set up. The facility has 16 racks, labeled Rack-1-Rack-16, and three CRAC units, CRAC-1-CRAC-3. The region with grids indicates perforated tiled floor in the cold aisle. Transient temperatures were measured in the cold aisle between racks, Rack-5 and Rack-6, within the box with ashed line.

Table 1
Specifications of the experimental setup.

Component	Specification	Comment
R-1	5.2 kW	Network rack
R-2	5.2 kW	Storage rack
R-3	8.48 kW	IBM blade center
R-4	6.4 kW	IBM blade center
R-5	10.08 kW	IBM blade center
R-6	10.08 kW	IBM blade center
R-7	8.8 kW	IBM blade center
R-8	10.72 kW	IBM blade center
R-9	9.6 kW	IBM blade center
R-10	6.4 kW	IBM blade center
R-11	9.6 kW	IBM blade center
R-12	0	Empty
R-13	10.48 kW	IBM blade center
R-14	0	Empty
R-15	0	Empty
R-16	0	Empty
Perforated tiles	610 mm × 610 mm; 56% Porosity	Passive tile
Floor plenum	914 mm Height	Cooling air supply
Ceiling plenum	1524 mm Height	Hot air exhaust

developed. At the beginning of a measurement, typical latency period for the digital system is 13.2 ± 0.5 s. However, with the correction of the initial offset, the data acquired from the array are considered to be nearly instantaneous. Additionally, the thermocouple based data acquisition system has an uncertainty on the order of ± 0.5 °C, determined by Omega CL122 thermocouple calibrator (www.omega.com/Manuals/manualpdf/M2931.pdf), utilizing a NIST traceable calibrated thermometer.

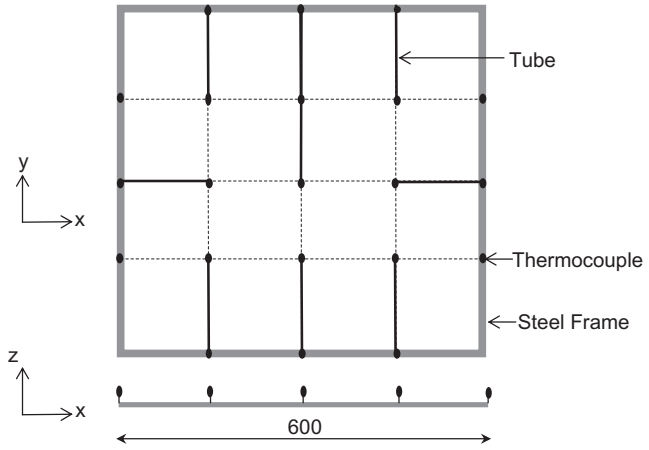

Fig. 4. A projection of the thermocouple-based temperature measurement grid which has 21 T-type copper-constantan thermocouples arranged in a square symmetrical geometry. *X*-axis is parallel to the rack width; *Y*-axis parallel to the tile; and *Z*-axis parallel to the direction of the rack height.

4. Results and discussion

This paper focuses on the transient temperature field following a step-change in the capacity of CRAC-1. At $t = -120$ s, the CRAC fan unit is shut down. After remaining inactive for 2 min, the CRAC unit is powered back at $t = 0$. Subsequently, the data acquisition system measures temperatures until $t = 300$ s. Fig. 5 shows transient air temperature variations after the CRAC unit is powered back at $t = 0$. These temperatures are measured at the center of the perforated floor tile in front of Rack-5 (Fig. 3) at different heights: h = (1960 mm, 1644 mm, 1288 mm, 932 mm, 576 mm, 220 mm.) from the tile surface. Cooling airflow from CRAC-1 reduces average air temperatures. However, air temperatures decrease disparately at different heights. As evident from Fig. 5, the temperature decrease near the top of the rack is rather gradual. This trend is attributed to hot air re-circulation near the top of the rack. As a result, a marginal temperature drop on the order ~1 °C is observed near the top of the rack. On the other hand near the perforated tile (h = 220 mm), temperature drops precipitously. With hardly any hot air re-circulation, the temperature field near the perforated tile surface is dominated by convective airflow from the perforated tile. Consequently, a sharp temperature drop of the order ~10 °C is observed near the perforated tile surface.

Following the transient data acquisition, a temperature ensemble is constructed by taking snapshots of the data at t = 10, 20, ..., 190, 200 s. An ensemble of size 252 × 20 is developed. The row size, 252 (= 21 × 12), characterizes the number of temperature sensors and the column size, 20, represents the number of transient observations used to form the temperature ensemble. The choice of the number of transient observations is arbitrary. However, a few snapshots yield a crude approximation, whereas a large number of snapshots will be computationally sluggish. Also, the choice of t = 200 s as our final snapshot time instant is chosen since large transient temperature variations are observed in this time window. After the ensemble is developed, Eq. (5) uses the singular value decomposition to compute POD modes. Then, Eq. (9) determines fractions of energy contents of different POD modes. The bar chart in Fig. 6 shows the energy contents of different POD modes. Fig. 6 suggests the energy content of the first POD mode is more than 50% of entire energy spectrum, and that of the second is about 10%.

As Eq. (10) outlines, the number of principal components, k, is plotted against different values of C.E.P.s in Fig. 7. For a crude

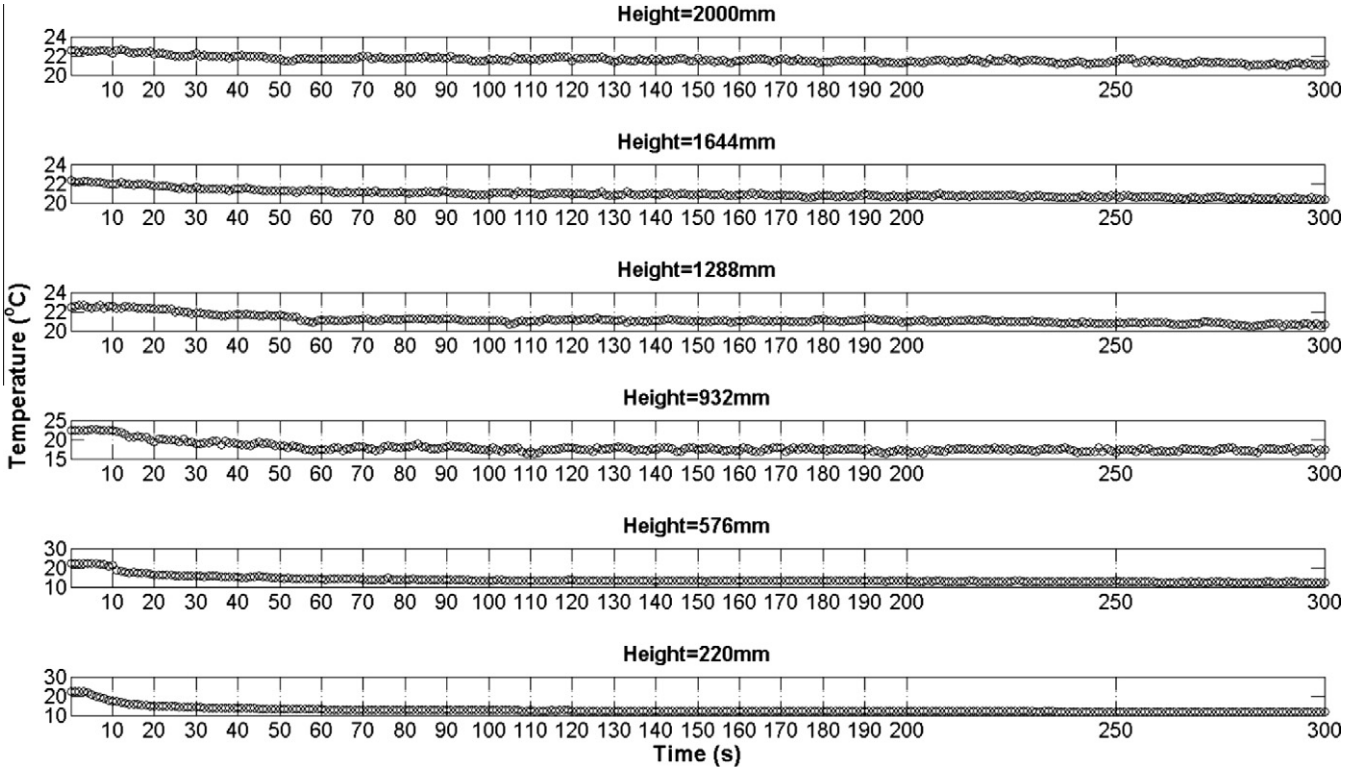

Fig. 5. Transient air temperature at different heights (1960 mm, 1644 mm, 1288 mm, 932 mm, 576 mm, 220 mm). Near the top, the transient temperature variation is approximately equal to 1.5 °C and that near the bottom is about 10 °C.

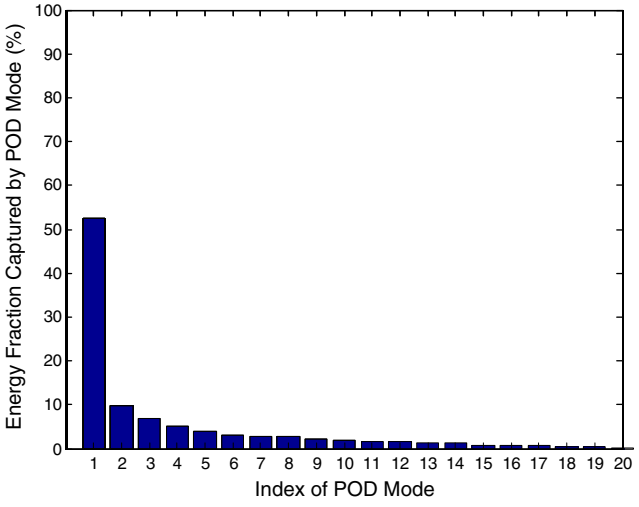

Fig. 6. Energy contents of different POD modes: Optimality of POD modes. The horizontal axis denotes the index of a POD mode, and the vertical axis denotes the energy fraction captured by a POD mode. First 10 POD modes capture more than 90% of the energy spectrum. The principal component number is equal to 17 for 99% C.E.P.

model where the accuracy requirement is low, C.E.P. is also low which in effect makes the number of principal components low. As it can be observed from Fig. 7, for C.*E.P.* = 75% :k = 5, and *C.E.P.* = 99.5%:k = 18. Between these two extremes, k varies non-linearly with C.E.P.: after C.*E.P.* = 95%, k increases rapidly. Since the focus of this paper is development of a high-fidelity surrogate model, a high C.*E.P.* = 99% is chosen for the results reported in this document. The corresponding number of principal components is determined as k = 17. Based on this choice of principal components, the POD/interpolation or the POD/extrapolation frameworks constructs the temperature data following the model stated in Eq. (11).

For the inlet of Rack-5, Fig. 8 shows experimental and POD predicted temperatures at an intermediate time instant of t = 92 s. The deviation between these two temperature data sets, i.e., the prediction error distribution is also shown. While the data acquisition system captures a 3-D temperature field, for convenience of data analysis, Fig. 8 presents a subset: a 2-D plane located at the rack inlet. The temperatures at the rack inlets are critically important because they are directly responsible for cooling heated servers. The POD/interpolation-based prediction for the rack-inlet temperature at t = 92 s (intermediate to snapshots at t = 90 s and t = 100 s) requires 4 s with an Intel© Core 2 Duo CPU at 2.54 GHz. It is, therefore, faster than an independent experiment. Also, this framework obviates the necessity of high-resolution data acquisition because it is capable of improving granularity of low-resolution experimental data. For example in this case, new temperature data can be constructed from an ensemble comprised of 20 temperature data at t = 10, 20, 30, ..., 200 s. Both experimental data (Fig. 8(a)) and POD-based predictions (Fig. 8(b)) suggest stratifications of local temperature fields. Hot layers of air near the top of the rack can be explained by air recirculation from the hot aisle. Moving down the rack inlet plane, the temperature gradually drops due to increasing convective cooling effect near the perforated tile. At the lowest location, a hot air pocket is observed which is explained by the Venturi effect of the constricted airflow through the gap between the rack and the floor. Furthermore, the deviation is obtained by subtracting POD-based predictions from corresponding experimental data. Fig. 8(C) shows the deviation which is of the order 1 °C. The distribution of the deviation, dependent upon the local stochastic dynamics, has not been investigated in this study.

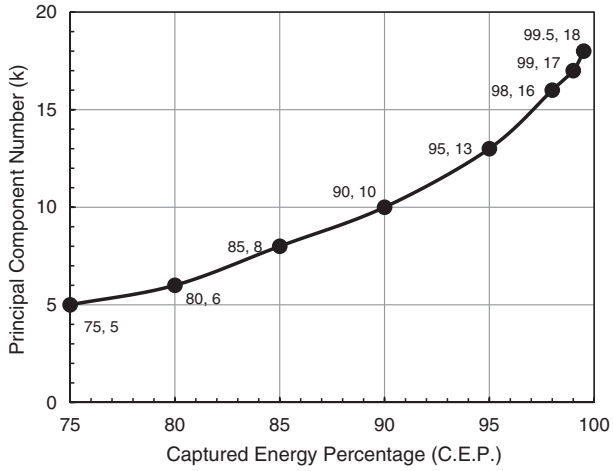

Fig. 7. The variation of captured energy percentage (C.E.P.) vs. the number of principal components (k).

As discussed in Section 2, an optimized analytical estimate of error bound for the POD/interpolation framework obviates the need for finding prediction error which requires a posteriori independent experiments. Alternatively, an optimal $E^{POD/Extrapolation}_{Analytical}$ can be added to T_{POD} to obtain reliable temperature data. Based on the lowest value of the unified decision-making index (I), the constant, c_0 (Eq. (15)) is empirically determined as 0.1. Table 2 shows the optimization procedure for determining c_0. Based on the optimal value of c_0, $E^{POD/Interpolation}_{Analytical}$ is calculated. On the other hand, Eq. (12) gives the prediction error, $E^{POD/Interpoaltion}_{Prediction}$, which is determined by subtracting experimental data to corresponding POD prediction (Eq. (11)). Fig. 9 compares the temporal variations of $E^{POD/Interpolation}_{Prediction}$ and $E^{POD/Interpolation}_{Analytical}$. It suggests the maximum deviation between $E^{POD/Interpolation}_{Prediction}$ and $E^{POD/Interpolation}_{Analytical}$ is equal to 0.2 which is less than 1% of the average temperature scale in the air-temperature field (~20 °C). Therefore, $E^{POD/Interpolation}_{Analytical}$ can substitute $E^{POD/Interpolation}_{Prediction}$ without compromising the solution accuracy.

The POD/extrapolation-based framework, which can also compute temperature fields beyond the ensemble time domain, i.e., time instants greater than the final snapshot time instant which is $t = 200$ s in the present case-study, is developed. It involves spline-based extrapolation of POD coefficients at snapshot time-instants. Fig. 10 shows experimental temperature field (Fig. 10 (a)), POD-based temperature prediction (Fig. 10 (b)), and the deviation (Fig. 10 (c)) between them at a time instant outside the ensemble domain: $t = 207$ s. A comparison between experimental data and POD-based predictions indicates average temperatures predicted from these two methods are close. Nevertheless, a close scrutiny reveals that local temperature distributions are moderately different, which is reflected in the deviation between experimental data and POD-based predictions. As Fig. 10(c), shows the deviation is within a scale of [−2.5 °C, 1.5 °C].

As discussed in Section 2, an optimized analytical estimate of error bound for the POD/interpolation framework obviates the need for finding prediction error which requires a posteriori independent experiments. An optimal $E^{POD/Extrapolation}_{Analytical}$ can be added to T_{POD} to obtain reliable temperature data. As discussed in reference to Eq. (13), the scale of the temperature difference chosen as $\Delta E^{Measurement}_{Scale} = (20-12)^0C = 8^0C$ (the difference between minimum initial temperature and temperature of supplied cooling air), which is indeed a characteristic of the thermal system involved in this case study. The scale factor, f, in Eq. (13) is arbitrary chosen to be 0.25. A larger f will increase E.H. at the cost of the extrapolation accuracy and vice versa. Based on these arbitrarily chosen parameters, the extrapolation horizon is calculated to be equal to 24 i.e., till $t = 224$ s. Once E.H. is derived, the determination of $E^{POD/Extrapolation}_{Analytical}$ requires identifying the case-specific constants (k, h^P, l, θ) and conducting the optimization procedure for identifying the arbitrary constants (c_1, c_2, c_3). The case-specific constants depend upon the experimental setup and conditions. The time-

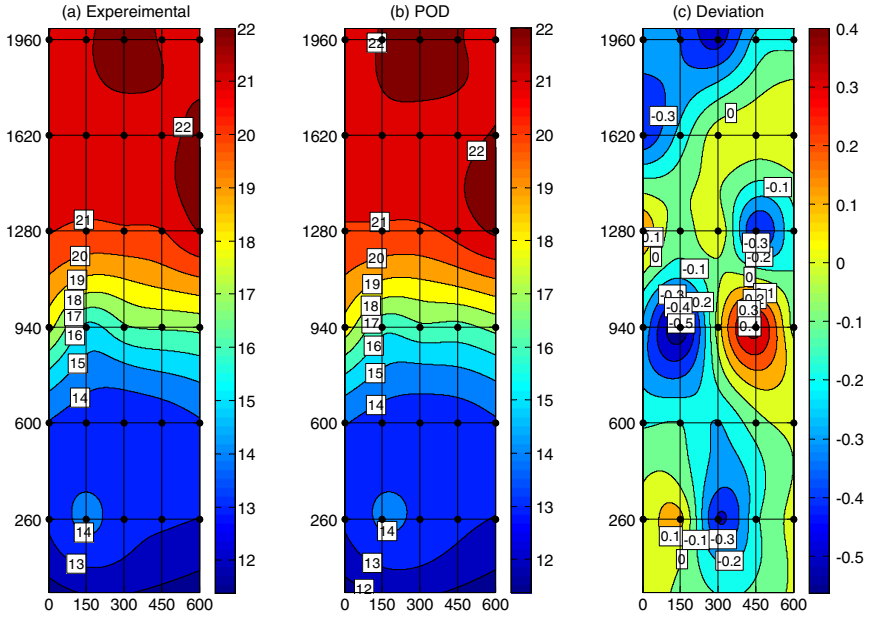

Fig. 8. The contour plots for temperature distributions at the inlet of Rack-5 at $t = 92$ s. The horizontal direction of 600 mm length indicates the width of the rack, and the vertical direction of 2000 mm length indicates the height of the rack, (a) shows experimentally-acquired temperature field and, (b) shows POD-predicted temperature field. The POD-based algorithm uses interpolation to compute the temperatures. The temperature scales are almost identical [14 °C, 22 °C]. Indeed, as shown in (c), the deviations between experimental data and POD-predicted data are within a scale of [−0.5 °C, 0.4 °C]. The black filled markers are the locations of temperature sensors. Remaining data points are produced by Delaunay triangulation.

Table 2
Numerical method for determination of c_0.

	$c_0 = 0.01$	$c_0 = 0.10$	$c_0 = 0.25$	$c_0 = 0.5$	$c_0 = 0.75$	$c_0 = 1$
Average	0.060	0.050	0.134	0.331	0.529	0.728
Standard deviation	0.055	0.031	0.054	0.057	0.057	0.057
Decision-making Index	0.0575	0.0405	0.094	0.194	0.293	0.3925

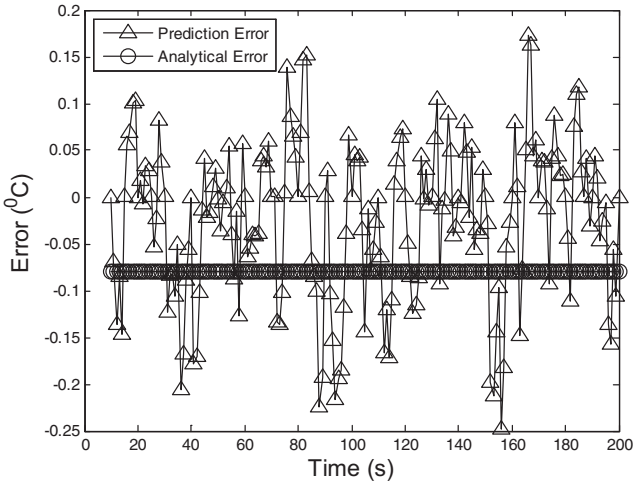

Fig. 9. Analytical error bound for POD-based interpolation. The solid line with the triangular marker shows the transient deviations in prediction error between the experimentally-acquired temperature data and the POD-predicted temperature data. The solid line with the circular marker shows the analytically-determined transient deviation or error between the exact solution data and the POD-predicted temperature data.

step for the POD/extrapolation framework is $k = 1$, since the extrapolation is carried out at a frequency of 1 Hz beyond $t = 200$ s. The normalized length scale (h^p) is defined as the ratio of the distance between two neighboring sensors (=150 mm in this case) and the characteristic length of the measurements system (=600 mm is the length of the square grid). Hence, it is calculated: $h^p = 0.25$. The number of snapshots included in the temperature ensemble is equal to 20: $l = 20$. The non-dimensional time, θ, defined as the time normalized against end of the transient measurement window (=300 s in this case). Hence, it is calculated $\theta \in [0.67, 0.74]$ for the derived E.H. The optimization procedure for determining arbitrary constants (c_1, c_2, c_3) is outlined in Table 3. The 'Average' column lists the average of e, and the 'Std. Dev.' column lists the standard deviation of e. The fractional difference, e, between $E^{\text{POD/Extrapolation}}_{\text{Analytical}}$ and $E^{\text{POD/Extrapolation}}_{\text{Prediction}}$ is calculated based on Eq. (17). The unified decision-making index (I) is calculated based on an optimization weightage, $\omega = 0.5$, which is arbitrarily assigned. As listed in Table 3, the minimum value of I is equal to 1 which corresponds to (c_1, c_2, c_3) = (- 15, 1.7, 2).

An optimally small value of I indicates $E^{\text{POD/Extrapoaltion}}_{\text{Analytical}}$ is closely matching with $E^{\text{POD/Extrapoaltion}}_{\text{Prediction}}$. Indeed, as shown in Fig. 11, $E^{\text{POD/Extrapoaltion}}_{\text{Analytical}}$ and $E^{\text{POD/Extrapoaltion}}_{\text{Prediction}}$ are closely matching at least to

Table 3
The optimization procedure for the determination of (c_1, c_2, c_3).For different combinations of (c_1, c_2, c_3), the unified decision-making indices (I) are calculated. The combination ($c_1 = -15$, $c_2 = 1.7$, $c_3 = 2$) is the best choice because it optimally minimizes I.

c_1	c_2	c_3	Average	Std. dev.	w	I
−10.0	1.8	2.0	17.5	13.8	0.5	15.6
−5.0	1.8	2.0	33.7	27.9	0.5	30.8
−15.0	1.8	2.0	1.5	0.7	0.5	1.1
−15.0	2.0	2.0	3.1	1.2	0.5	2.1
−15.0	1.5	1.5	4.0	4.9	0.5	4.5
−15.0	1.8	1.8	1.3	1.6	0.5	1.4
−15.0	1.8	2.0	1.7	0.8	0.5	1.2
−15.0	1.7	2.0	1.3	0.6	0.5	1.0
−16.0	1.7	2.0	2.3	3.6	0.5	2.9
−14.0	1.7	2.0	4.3	2.2	0.5	3.3

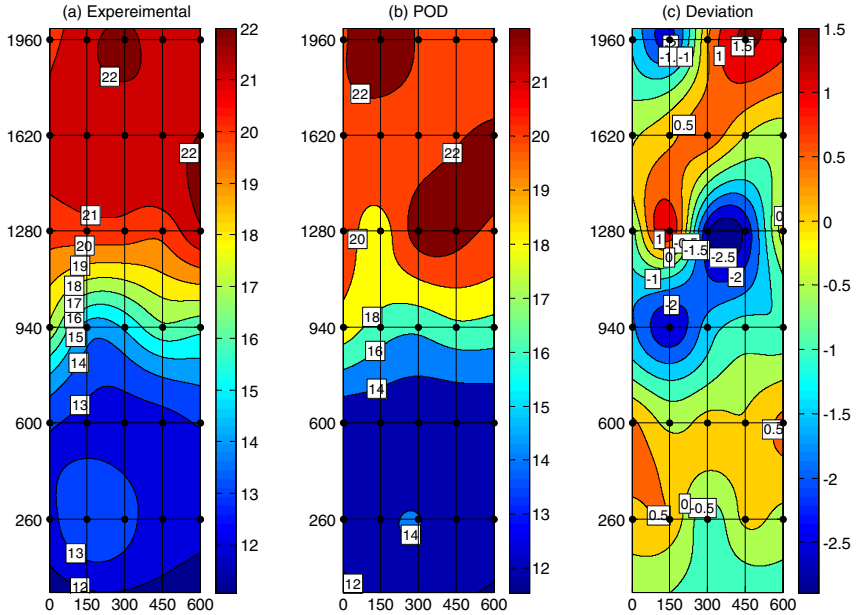

Fig. 10. The contour plots for temperature distributions at the inlet of Rack-5 at $t = 207$ s. The horizontal direction of 600 mm length indicates the width of the rack, and the vertical direction of 2000 mm length indicates the height of the rack, (a) shows experimentally-acquired temperature field and, (b) shows POD-predicted temperature field. The POD-based algorithm uses extrapolation to compute the temperatures. The temperature scales are almost identical [14 °C, 22 °C]. Indeed, as shown in (c), the deviations between experimental data and POD-predicted data are within a scale of [−2.5 °C, 1.5 °C]. The black filled markers are the locations of temperature sensors. Remaining data points are produced by Delaunay triangulation.

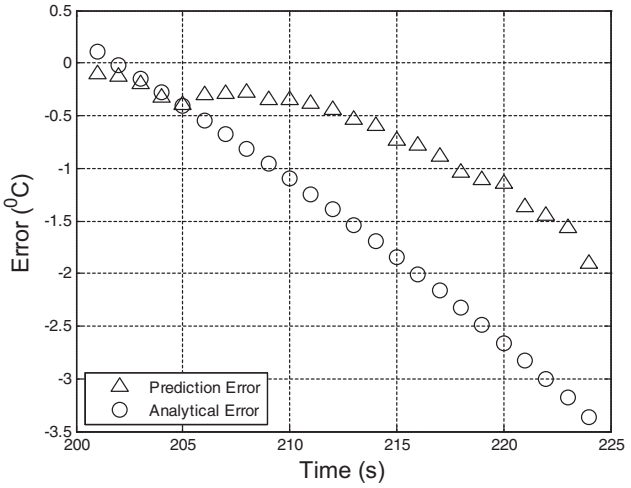

Fig. 11. Analytical Error Bound for POD-based extrapolation. The triangular markers show the transient deviations between the experimentally-acquired temperature data and the POD-predicted temperature data or prediction error. The circular markers show the analytically-determined transient deviation between the exact solution data and the POD-predicted temperature data or analytical error.

an order of 1 °C. It means $E_{\text{Analytical}}^{\text{POD/Extrapoaltion}}$ can be directly added to T_{POD} to obtain a reliable set of temperature data.

5. Conclusions

The POD-based model presented in this paper uses an ensemble of air temperature data at discrete time instants inside the data center facility, and predicts temperature fields both inside and outside of the discrete time domain. This capability presents the prospect of utilizing this methodology for future corrective actions within the facility, based on present measurements. The POD-based model is tested for its accuracy and efficiency in predicting real-time data. It is found that such predictions save computational time, without significantly compromising solution accuracy. For both interpolation and extrapolation-based predictions, the trends in their deviations from the corresponding experimental data are analyzed, and their respective error bounds are semi-analytically obtained. By adding corresponding error bounds to POD-based predictions, the developed framework can yield high-fidelity predictions with uncertainty on the order of original experiments. The presented dynamic modeling framework can reduce data acquisition costs by improving the granularity of the experimental data, and utilizing limited data to predict possible future outcomes in similar physical conditions.

In the context of data centers, a high-fidelity dynamic modeling framework opens up the possibility of designing control systems that can monitor and analyze existing data center conditions and react, if required, appropriately. Such control systems would potentially be useful in improving efficiency of thermal management of data centers [23]. Also, a high-fidelity transient extrapolation capability could be useful for early detection of thermal hazards, e.g., fire, in data centers to allow increased reaction time to take preventive measures.

Although the developed modeling framework is promising, it has a number of limitations. First, the experimental data is limited only to rack-level time and length scales. Smaller scales in a multi-scale data center, such as those associated with the server unit, remain unresolved. A higher resolution data set is needed to develop a more holistic framework. Second, the developed reduced-order framework is a single parameter model with time as the only parameter. A more realistic model must include multiple secondary parameters such as CRAC fan speeds, server heat loads etc. Lastly, error bounds determined are partially empirical. Nevertheless, arbitrary constants are closely related to different experimental parameters.

Acknowledgments

The authors acknowledge support for this work from IBM Corporation, with Dr. Hendrik Hamann as the Technical Monitor. Acknowledgements are also due to the United States Department of Energy as the source of primary funding. Additional support from the National Science Foundation award CRI 0958514 enabled the acquisition of some of the test equipment utilized.

Appendix A

A.1. Analytical Error for the POD/Interpolation Framework

For determining the analytical error of the POD/interpolation scheme, $E_{\text{Analytical}}^{\text{POD/Interpolation}}$, a linear algebra-based analysis, as documented in Section 2.3 of [14], is utilized. The important features of the analysis are outlined in the remaining part of Appendix A.

Let $T^1, T^2, \ldots, T^l$ are snapshots and let $\zeta := span\{T^1, T^2, ..., T^l\} \in T$ with $m := \dim(\zeta)$. Assume $\{\psi\}_{i=1}^{m}$ is an orthonormal basis of ζ:

$$T^j = \sum_{i=1}^{m} (T^j, \psi_i)\psi_i, \quad \text{for } j = 1, \ldots, l. \tag{23}$$

The fundamental principle of a reduced-order modeling is finding $d(< m)$ orthonormal basis vectors $\{\psi_i\}_{i=1}^{d} \in T$ such that the mean square error between the elements of the ensemble set and corresponding dth partial sum is minimized on average:

$$\min_{\{\psi\}_{i=1}^{d}} \frac{1}{l}\sum_{j=1}^{l} \|T^j - \sum_{i=1}^{d} (T^j, \psi_i)_v \psi_i\|_1^2 \quad \text{subject to } (\psi_i, \psi_j) = \delta_{ij} \quad \text{for } 1 \leqslant i \leqslant d, 1 \leqslant j \leqslant i. \tag{24}$$

POD error can be reformulated:

$$\min_{\{\psi\}_{i=1}^{d}} \frac{1}{l}\sum_{j=1}^{l} \|T^j - \sum_{i=1}^{d} (T^j, \psi_i)\psi_i\|_v^2 = \min_{\{\psi\}_{i=1}^{d}} \frac{1}{l}\sum_{j=1}^{l}\sum_{i=d+1}^{m} |(T^j, \psi_i)_1|^2 = \sum_{i=d+1}^{m} \lambda_i. \tag{25}$$

In addition, a constant, c_0, is multiplied to the sum of the eigenvalues corresponding to the discarded POD modes to fully specify $E_{\text{Analytical}}^{\text{POD/Interpolation}}$. The arbitrary constant, c_0, quantifies the interpolation error in determining POD coefficients.

Appendix B

B.1. Analytical Error Bound for POD-based Extrapolation

For determining the analytical error of the POD/extrapolation scheme, $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$, a weak formulation-based functional analysis, as documented in [14], is used. Instead of a weak formulation-based functional analysis for the Navier–Stokes equations as conducted in [14], the analytical error for the POD/extrapolation framework requires a functional analysis of the energy equation. The governing equation for the convective air temperature field, $T(x, y, z, t)$ inside a data center is:

$$\frac{\partial T}{\partial t} - (\alpha + E_H)\nabla^2 T + \vec{u}.\nabla T = \dot{q}. \tag{26}$$

For the sake of simplicity, the initial condition is chosen to be independent of spatial locations: $T(t = 0) = T_0$. The boundary conditions for air temperatures in a data center are often complicated:

following [14], the boundary temperatures are chosen to be equal to zero. The resultant error is taken care while determining the constant coefficients by the optimization procedure as outlined in Eqs. (17)–(22). Both the Navier–Stokes equations and Eq. (26) are conservation equations; therefore, both of them have similar forms except for Eq. (26) does not have the *pressure gradient* term like the Navier–Stokes equations. Nevertheless, same analytical methodology [14] is used considering the *pressure gradient* term does not feature in the weak formulation in Eq. (27).

$$\begin{aligned} &(T_t, v) + a(T, v) + b(u, T, v) = (q, v).\\ &a(T, v) := \alpha \int_\Omega \nabla T : \nabla v dx, b(U, T, v) = \int_\Omega (U.\nabla) T.v dx. \end{aligned} \tag{27}$$

The determination of the analytical error, $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$, in [14] is essentially a two-step procedure: first, the estimation of the deviation between the exact solution and the numerical solution [21,24], and second, the estimation of the deviation between the numerical solution and the reduced-order solution. The second part of the procedure is exhaustively derived in [14]. Finally, the errors determined from previous two steps are added to obtain the bound for the deviation between the exact solution and the reduced-order model solution, $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$.

The deviation between the exact solution and the POD-based prediction is:

$$\begin{aligned} E_{\text{Analytical}}^{\text{POD/Extrapolation}} \leqslant\ & c_f(\sigma^{-1}(t_m)k + h^p)\\ &+ \left[\frac{l\exp(\theta)}{1-\theta}(\lambda + C_5 k)\sum_{n=d+1}^{m}\lambda_n\right]^{\frac{1}{2}}, C_6 k \leqslant \theta < 1. \end{aligned} \tag{28}$$

Where, c_f, λ, C_5, C_6 are arbitrary constants. $\sigma^{-1}(t) = \min(1, t)$. k: = Time step. h^p: = Finite element size. l: = Number of snapshots. λ_n: = Eigenvalues corresponding to POD modes.

With k and h^p featuring in Eq. (28), it is evident that the discretization of the numerical scheme is an integral part for determining $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$. By definition, a numerical solution framework involves discretization which is essentially transforming continuous equations into its discrete counterparts. Similarly, experimental data can be modeled as a discrete sample set of the solution space of the governing equation. For an experimentally-derived discrete dataset, the time step, k, can be modeled as the time difference between two consecutive observations, and the finite element size, h^p, can be modeled as the normalized distance between two neighboring sensors. After the functional form of the analytical error, $E_{\text{Analytical}}^{\text{POD/Extrapolation}}$, is determined, its complete specification involves a multi-dimensional optimization.

References

[1] J.G. Koomey, Growth in Data Center Electricity Use 2005 to 2010, Analytics Press, 2011.
[2] G. Meijer, Cooling energy-hungry data centers, Science 328 (2010) 318–319.
[3] M.A. Bell, Use best practices to design data center facilities, Gartner Research 22 (2005).
[4] D. Cole, Data center infrastructure management, Data Center Knowledge (2012).
[5] R. Zhou, Z. Wang, A. McReynolds, C. E. Bash, T. W. Christian, and R. Shih, Optimization and Control of Cooling Microgrids for Data Centers, in: 13th IEEE ITherm Conference, San Diego, California, USA, 2012, pp. 338–343.
[6] S.V. Patankar, Airflow and cooling in a data center, J. Heat Transfer 132 (2010). 73001-1-73001-17.
[7] J. D. Rambo, Reduced-order modeling of multiscale turbulent convection: Application to data center thermal management, Ph.D. Dissertation, Georgia Institute of Technology, Atlanta, GA, 2007.
[8] J. Biazar, H. Ebrahimi, Existence and uniqueness of the solution of non-linear systems of Volterra integral equations of the second kind, J. Adv. Res. Appl. Math. 2 (4) (2010) 39–51.
[9] R. Genesio, A. Tesi, Harmonic balance methods for the analysis of chaotic dynamics in nonlinear systems, Automatica 28 (1992) 531–548.
[10] G. Berkooz, P. Holmes, J.L. Lumley, The proper orthogonal decomposition in the analysis of turbulent flows, Ann. Rev. Fluid Mech. 25 (1993) 539–575.
[11] E. Samadiani, Reduced order modeling based energy efficient and adaptable design, in: Y. Joshi, Pramod Kumar (Eds.), Energy Efficient Thermal Management of Data Centers, Springer, 2012.
[12] V. López, H.F. Hamann, Heat transfer modeling in data centers, Int. J. Heat Mass Transfer 54 (2011) 5306–5318.
[13] H. F. Hamann, V. López, A. Stepanchuk, Thermal zones for more efficient data center energy management, in: Proceedings of 12th IEEE Intersociety Conference on Thermo and Thermomechanical Phenomena in Electronic System, Yorktown, NY, June 2010.
[14] S. Ravindran, Error estimates for reduced order POD models of Navier-Stokes equations, in: ASME Int. Mech. Eng. Cong. Expos., Boston, Massachusettes, USA, 2008, pp. 652–657.
[15] P. Holmes, J.L. Lumley, G. Berkooz, Turbulence, Coherent Structures, Dynamical Systems and Symmetry, Cambridge Univ Pr, 1998.
[16] L. Sirovich, Turbulence and the dynamics of coherent structures Part I: Coherent structures, Quarter. Appl. Math. 45 (1987) 561–571.
[17] P. Druault, P. Guibert, F. Alizon, Use of proper orthogonal decomposition for time interpolation from PIV data, Exp. Fluids 39 (2005) 1009–1023.
[18] D. Alonso, A. Velazquez, J. Vega, A method to generate computationally efficient reduced order models, Comput. Methods Appl. Mech. Eng. 198 (2009) 2683–2691.
[19] M. Joyner, Comparison of two techniques for implementing the proper orthogonal decomposition method in damage detection problems, Math. Comput. Model. 40 (2004) 553–571.
[20] H. Banks, M.L. Joyner, B. Wincheski, W.P. Winfree, Real time computational algorithms for eddy-current-based damage detection, Inverse Prob. 18 (2002) 795–823.
[21] R. Temam, Navier–Stokes equations: theory and numerical analysis, vol. 2, American Mathematical Society, 2001.
[22] G. M. Nelson, Development of an Experimentally-Validated Compact Model of a Server Rack, Master's Thesis, Georgia Institute of Technology, Atlanta, GA, 2007.
[23] J. Liu, F. Zhao, X. Liu, W. He, Challenges towards elastic power management in Internet data centers, in: 29th IEEE International Conference on Distributed Computing Systems Workshops, Readmond, WA, USA, 2009, pp. 65–72.
[24] V. Girault, P.A. Raviart, Finite Element Methods for Navier–Stokes Equations: Theory and Algorithms, Springer, Berlin, 1986.