2018

Data Driven Prediction Model (DDPM) for Server Inlet Temperature Prediction in Raised-floor Data Centers

Raymond Lloyd
IBM Research Dublin

Marek Rebow
Technological University Dublin, marek.rebow@tudublin.ie

Follow this and additional works at: <https://arrow.tudublin.ie/engscheleart>

Part of the [Electrical and Electronics Commons](#)

Recommended Citation

Lloyd, R. & Rebow, M. (2018) Data Driven Prediction Model (DDPM) for Server Inlet Temperature Prediction in Raised-floor Data Centers, *2018 17th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)* DOI:10.1109/itherm.2018.8419650

This Conference Paper is brought to you for free and open access by the School of Electrical and Electronic Engineering at ARROW@TU Dublin. It has been accepted for inclusion in Conference papers by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie, gerard.connolly@tudublin.ie.

Data Driven Prediction Model (DDPM) for Server Inlet Temperature Prediction in Raised-floor Data Centers

Raymond Lloyd, Marek Rebow*

IBM, IBM Research Dublin, Ireland

*Dublin Institute of Technology, Dublin Energy Lab, Dublin, Ireland

Email: raylloyd@ie.ibm.com, marek.rebow@dit.ie

ABSTRACT

Data centers account for approx. 1.4% of the world's electricity consumption, of which up to 50% of that power is dedicated to keeping the actual equipment cool. This represents a huge opportunity to reduce data center energy consumption by tackling the cooling system operations with focus on thermal management.

This work presents a novel Data Driven Predictive Model (DDPM) for temperature prediction of server inlet temperatures that utilises high resolution empirical temperature measurements from 52 real-life data centers. A knowledge-base of temperature data and related physical features, created via clustering techniques was used to train a series of artificial neural networks (ANN). The ANNs are used to make predictions of server inlet temperatures based on inputs which describe the boundary conditions.

The temperature predictions are made for each server rack to estimate the vertical temperature distribution (s-curve) from the bottom to top of the rack spaced at one foot intervals. Each ANN predicts a temperature at a corresponding vertical height for the given inputs, producing the s-curve reconstructed from the combination of ANN outputs. Furthermore, one ANN predicts the s-curve cluster which is used to provide a prediction confidence.

The model only requires local boundary conditions such as rack power, perforated tile airflow rate and temperature, ceiling temperature and a rack adjacency identifier (RAI), in addition to average delta T for air conditioners and an identifier for data center layout type. Both RAI and data center identifier are assigned programmatically given a data centers layout information while the rack power, perforated tile airflow rate and delta T can be measured, metered or calculated.

Unlike other statistical approaches which are specific to a single data center room, the DDPM is trained with data from a wide range of data centers, therefore can be used to predict server inlet temperatures in many different types of data centers. Performance of the model on an Intel i7 powered laptop computer with 16GB memory is in the order of 660,000 temperature location predictions per second.

The prediction is accompanied by a confidence level in the prediction. DDPM results gave a prediction accuracy of 0.76C

RMSE with a 0.12 probability of one point on the s-curve crossing the upper or lower bounds of the confidence interval.

The model can perform in real-time, giving way to applications for real-time monitoring, input to optimize control of air conditioning units, and can complement sensor networks.

I. INTRODUCTION

The energy consumption in the ICT sector increased tremendously over the last years up 7% of electricity worldwide mainly due to the cloud computing and the rapid growth of the use of Internet services such as real-time video streaming, online gaming and mobile devices, but can reach 13% by 2030 due to the development of the Internet of Things remote digital sensors, devices and driverless cars [1]. The data centre sector in particular was estimated to account for 1.4% of the global electricity consumption for 2011 [2], with US data centres consumed about 70 billion kilowatt-hours of electricity in 2014, representing 1.8% of the countrys total energy consumption, according to results of a new study of data centre energy use by the Lawrence Berkeley National Laboratory [3]. In small countries such as Ireland, the expanding data centre sector will pose an enormous challenge in meeting countrys CO2 emission targets where total electricity demand is forecast to grow by 15% in the low demand scenario, or by 36% in the high scenario with the data centre capacity to reach 1400 MVA by 2026 [4]. Efficiency improvements in data centres have contributed significantly to restrain the growth rate of their energy consumption. In the US, data centre electricity consumption increased by 4% from 2010-2014, a significant change from the 24% increase estimated from 2005-2010 and is expected to continue increasing by 4% from 2014-2020 to reach 73 billion kWh in 2020 [3]. The data centre efficiency is measured by using the standardized Power Usage Effectiveness metric (PUE) which is defined as the ratio of the data centre total input power to the power used by the IT equipment. The higher the PUE value is the lower the efficiency of the facility is as more power is used by the HVAC systems and auxiliary equipment. Recent studies of 268 facilities in Europe show that their PUE average value is 1.8, but there are 30 sites that reported much better efficiencies between 1.4 and 1.1 [1]. Similar studies in the US reveal that the industry-wide weighted average PUE

remains at 1.89 due to the slower turnover rate of data centers HVAC infrastructure relative to the IT equipment, especially in smaller data centres with lower ability to install more efficient and advanced cooling technologies or to develop better airflow management solutions. Therefore, smaller data centers have PUE values greater than 2.0 while large hyperscale cloud data centers can achieve a record PUE value of 1.1 or even less [3]. In data centres the air delivered to the IT equipment has to meet the special temperature requirements in which the equipment should operate. Failure to comply with the requirements can void warranties and/or more importantly, cause a system failure resulting in losses of revenue, customers, integrity and potential penalties. A recent report by the Ponemon Institute estimated the average cost of unplanned down-time in data centres at \$740,357 up 38% from 2010 with average cost of downtime per minute at \$8,851 and as much as \$17,000 per minute [5]. The requirement to manage server inlet temperatures is vital to reducing unplanned outages. Control of a data centre room is not simply a task of matching cooling capacity with heat generated from IT equipment. Rather, cooling must be targeted to specific locations to ensure that only the inlets of servers consume the cold air while minimising mixing due to recirculation caused from the hot exhaust air from the servers. Real-time monitoring or predictions of inlet temperatures of servers can help in operational decisions within a specific facility to reduce its PUE and to mitigate risks of unplanned very costly down-times. Many modelling approaches exist today for estimating server inlet temperatures, each modelling approach has its own specific advantages and disadvantages. For example, Computational Fluid Dynamics (CFD) models are physics-based numerical methods which are quite accurately predict temperature and air flow fields in a data centre, but CFD is computationally expensive, costly to purchase and in general require trained personnel to develop, run and interpret the results. Moreover they are not suitable for real-time applications where quick results are necessary that may be used as input to settings of a Computer Room Air Conditioner (CRAC) control system. Other existing simulation techniques based on Potential Flow Theory (PFT)[6], Enhanced Potential Flow Model (EPFM) [7] or measurement-based physical modelling [8] [9] offer faster solution times but at the expense of accuracy and still need a full domain solution. Recently, a novel Data Centre Air Flow Model (DCAM) for the prediction of server inlet temperatures, which is real-time acceptable and accurate enough to be useful, was developed and implemented. This simplified physics-based model is grounded in the round free turbulent jet theory and adapted to the data centre environment in such a way that can make temperature predictions based on local parameter changes. The method was implemented as a programmatic algorithm, tested, validated and the results compared to high resolution temperature data collected on a production data centre which has undergone 25 different CRAC control setting changes and has returned an overall prediction error of 1.2°C [10]. On the other hand, statistical approaches offer fast computational speeds, but require training data which is not

always available and their accuracy degrades significantly as a data centre changes its layout so the need for re-training may outweigh the adoptability and practicality of these solutions. This paper presents a novel Data Driven Predictive Model (DDPM) for a statistical prediction of server inlet temperatures using a knowledge-base of created temperature profiles, called s-curves, with help of a newly developed hierarchical clustering algorithm. The knowledge-base is utilised to train series of artificial neural networks (ANN) to make real-time predictions of inlet temperatures and provide upper and lower probability bounds to the predictions. Since the knowledge-base is derived from multiple data sets collected from 52 various types of data centres containing over 11 million temperature measurements or 1,387,422 s-curves, so unlike other statistical models, this model can be re-used for other data centres. The data sets comprise of high resolution temperature fields collected using the IBM Measurement and Management Technology (MMT) mobile temperature cart [11].

II. ASSUMPTIONS AND LIMITATIONS

The research is based on data from data centres utilising a raised-floor cooling arrangement. Individual rack power numbers were not available in the data set, however room level power from PDU's were available. Rack level power is estimated by means of room power allocation based on temperature differential between inlet and outlets of the servers. The model predicts inlet temperatures in front of servers and is limited to server racks which are not placed at the end of aisles.

III. DATA DRIVEN PREDICTIVE MODEL - DDPM

We now describe the DDPM model where the goal is to limit the solution domain and boundary conditions to a local level and provide high speed performance prediction of inlet temperatures. The local level is defined as the area in front of the server, the perforated tile supplying the cool air, and the server itself.

A. Collecting Real World Empirical Data

Temperature data for training and validation was collected in the data centres using IBM Measurement and Management Technology (MMT) which was developed in IBM and is widely used in data centres around the world [12]. MMT uses a specially developed cart to collect temperatures from the data centre room. The cart consists of layers spaced vertically at 305mm intervals with nine thermocouples per layer and one humidity sensor. The cart has motion encoding wheels and hardware to capture the data relative to the location in the data centre. The MMT cart fits on a standard 600mm tile and is traversed tile by tile throughout the data centre to digitise a high resolution three dimensional snapshot of the thermal environment. Simultaneously, information is recorded from the CRAC units, power distribution units and perforated tiles.

B. Knowledge-base Creation

The next step is to extract a feature-set relevant to each s-curve which can be used to identify a specific s-curve and store as a knowledge-base consisting of the temperature profiles (s-curves) and the corresponding physical features that define the shape of the s-curve. The physical featured can be described as;

a) *Rack Adjacency Identifier RAI*: RAI is an ontology to describe a server rack and its surroundings relative to other racks, aisles and gaps between racks. Here a general form is used and assigned to each sensor location on each rack.

- R defined as Rack (either Server, Disk, Network or Tape)
- A defined as Aisle (Walkway)
- G defined as Gap

The format of the RAI is X-X-X where the centre X is the rack which is being identified. The letters listed above are assigned to the racks, aisles or gaps as appropriate. A more detailed RAI defines the rack type, compute, network, disk and tape and is illustrated in Fig. 1

Fig. 1. Results of the RAI Ontology Classification - determines the RAI for each rack in the data centre based on its geometric location

b) *Volumetric flow rate*: The volumetric flow rate data is measured for each perforated tile. Each s-curve position under a perforated tile is assigned this volumetric flow rate.

c) *Estimated rack power*: As individual rack power numbers were not available, rack power is estimated based on allocation of IT power P_{IT} calculated from room level power taken from PDU's Eqn. 1.

$$P_{IT} = \sum_1^p P_{pdu} - \sum_1^l P_{lights} - \sum_1^c P_{crac} + P_{misc} \quad (1)$$

$$P_{IT} = \sum_1^s P_{rack} \quad (2)$$

where s, p, l, c are number of servers, PDUs, lights and CRACs respectively.

The energy balance equation Eqn. 3 is the basis of the estimation,

$$Q = \rho \times c_p \times V \times \Delta T \quad (3)$$

however instead of a value of V_{rack} , the volumetric airflow rate across the rack, a scaling factor is calculated in the form

$$V_{scale} = V_{perf} \times \frac{n}{z} \times \Delta T_{rack} \times w \quad (4)$$

where V_{perf} is the perforated tile flow in front of the rack, z is the height (in layers) of data where the cold air from the perforated tile becomes equal to the ambient air temperature and n is the number of layers to top of the rack and w is the width of the rack in units of tiles. If the servers in the rack consumes all of the perforated tile air half way up the rack (air beyond the half way mark is recirculated air) the flow rate is scaled by a factor of 2 ($\frac{n}{z} = \frac{8}{4} = 2$). Similarly if the width of the rack is 1.5 tiles wide, a scale factor of 1.5 is applied. The scaling operation is performed for all racks. To predict the power of a specific rack, the power from the P_{IT} is allocated based on the scaling factor for the rack V_{scale}

$$P_{rack}^i = P_{IT} * \frac{V_{scale}^i}{\sum_{i=0}^s V_{scale}} \quad (5)$$

d) *Data centre type*: Since the layout of a data centre plays a significant roles in the air transport of hot air returning to the CRAC units, a data centre layout type classification is assigned to identify the type of layout arrangement of the data centre. The layout of equipment and placement of the CRAC units can have an impact on the routing of hot exhaust air returning to the CRAC units. In more modern data centres, the rows of server equipment are perpendicular to the CRACs which are placed in-line with the hot aisles, promoting a direct return to the CRAC units. Older data centres can have non-optimal designs and can be prone to hotspots as hot air becomes stagnant and recirculates as it is trapped in the server rows. The procedure classifies the data centre type based on the following criteria and illustrated in Fig. 2.

- 1: Rows of racks are perpendicular to CRACs
- 2: Rows of racks are parallel to CRACs no return path.
- 3: Rows of racks mixed between parallel and perpendicular to CRACs no return path.
- 4: Operating CRACs per m^3 of data centre floor space.

Fig. 2. Data Centre Site Type Classification

e) *Average temperature difference between CRAC return and discharge*: The average temperature difference between all operating CRAC unit return and discharge temperatures is calculated for each data centre.

C. Clustering

Next is to cluster together s-curves based on the similarity of their shape and for this task an adaptation of hierarchical agglomerative clustering is used. The clustering process generates sets of s-curves that have a controlled variance by setting a max allowable variance. The variance step $\vec{v}_i$ is incrementally increased to create the clusters until $\vec{v}_i = \vec{v}_{max}$ is reached.

The knowledge-base now consists of multiple sets of s-curves, with physical features corresponding to each set. Features include perforated tile flow rates, perforated tile discharge temperature, a rack ontology, data centre layout type classification, rack power and the average temperature difference between the CRAC return and discharge temperatures.

IV. MODEL TRAINING AND PREDICTION

Given the data has been clustered into groups based on curve shape similarity, the challenge now is to link the features to the corresponding cluster temperatures. To create and train this model, a series of artificial neural networks (ANN) are used. A supervised ANN from the neuralnet package in R software [13] is utilised. Neuralnet focuses on Multi Layers Perceptrons (MLP) on a directed graph of nodes with weighted directed edges. The nodes are arranged in layers. In this implementation up to four ANN layers are used, the input layer, one/two hidden layer and the output layer. The input layer comprises of a node for each covariate or observation feature. Each node connects only to subsequent nodes by a weighted edge. The structure of the graph and threshold parameters require tuning to the application and in this case were tested by trial and error to systemically arrive at the chosen configuration.

The output prediction is a function of the input covariates and graph structure and weights. It can be defined as [14]

$$o(\mathbf{x}) = f \left(w_0 + \sum_{j=1}^j w_j \cdot f \left(w_{0j} + \sum_{i=1}^n w_{ij} x_i \right) \right) \quad (6)$$

where $x_i \in \mathbf{x}$ are the co-variables, j is the j th hidden node, w_0 the intercept (bias) to the output node and w_{0j} the intercept to the j th hidden node. The activation function f is a sigmoid function $f(x) = \frac{1}{1+e^x}$.

To train the ANN a supervised learning approach with resilient backpropagation and backtracking is used. The network weights are initialised with random values from a normal distribution. The co-variables from the temperature observations are introduced together with the observation outputs. The observation outputs are compared to the predicted outputs and a error E (sum of squared errors) is calculated. The weights are adjusted based the learning rule, for resilient backpropagation this is [14]

$$w_k^{(t+1)} = w_k^{(t)} - \eta_k^{(t)} \cdot \text{sign} \left(\frac{\partial E^{(t)}}{\partial w_k^{(t)}} \right) \quad (7)$$

where t and k index the iteration and weights respectively and η_k the learning rate for weight. The learning rate is added

to the weight if the partial derivative of the error w.r.t. weight is negative and subtracted if it is positive, essentially seeking the local minimum for the partial derivative. In order to speed up the convergence η_k is increased if the corresponding partial derivative has kept its sign from the previous step and reduced if not. This can lead to the derivative swinging around the local minimum. Backtracking allows for the last adjustment to be undone and mitigates this problem [14].

a) *Training*: The training procedure for DDPM takes the input parameters consisting of perforated airflow rate, ceiling temperature, rack power, CRAC delta T and data centre type. The model comprises of 7 artificial neural networks as follows:

- Cluster Prediction - predicts the cluster (trains with cluster number as output)
- Temperature Prediction L2 = predicts the temperature at layer L2 (trained with L2 actual temperature as output)
- Temperature Prediction L3 = predicts the temperature at layer L3 (trained with L3 actual temperature as output)
- Temperature Prediction L4 = predicts the temperature at layer L4 (trained with L4 actual temperature as output)
- Temperature Prediction L5 = predicts the temperature at layer L5 (trained with L5 actual temperature as output)
- Temperature Prediction L6 = predicts the temperature at layer L6 (trained with L6 actual temperature as output)
- Temperature Prediction L7 = predicts the temperature at layer L7 (trained with L7 actual temperature as output)

Temperatures, rack power and flow rate values are normalised to values between 0 and 1 using the following general formula

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (8)$$

where x_n is the normalised value, x the value, x_{min} the minimum value of x and x_{max} the maximum value of x .

A general schematic representing the structure of the neural networks for cluster and temperature are shown in Fig. 3 and Fig. 4. The number of hidden layers and number of nodes require tuning and are discussed in Section V.

The cluster prediction ANN is trained with 75% (25% withheld for testing) of the data set by feeding in the input parameters for each s-curve and the associated cluster number as the output in training. The split between training and testing is based on the rule of thumb where the fraction of s-curves reserved for the validation set should be inversely proportional to the square root of the number of free adjustable parameters (input nodes) [15]. The temperature prediction ANNs are fed with the same training input data. Each temperature training ANN is trained with the output for a specific layer or vertical height, so one is required for each layer in the data. The combination of the individual temperature prediction ANNs enables reconstruction of the complete s-curve for that server inlet. While testing and tuning the ANN, it was found that creating an ANN for each layer increases the likelihood of convergence over outputting all layers at the one time in a single ANN model.

Fig. 3. Trained Artificial Neural Network - Cluster Predictor

Fig. 4. Artificial Neural Network schematic used to train the DDPM model.

b) Prediction: Prediction of inlet temperatures is made with a combination of all the ANN models. The trained models accept the input features on the input nodes of the model and calculate the outputs from the stored graph structure and edge weights. The prediction is twofold. Firstly a cluster is determined from the input features, then the temperature for each layer is determined and joined together to construct a complete s-curve. The complete output of the prediction provides for a full s-curve and a probabilistic band of upper and lower temperatures in which the actual temperature resides. The upper and lower bands are calculated from the predicted cluster. When the cluster is determined, the maximum and minimum temperatures of all cluster members are determined for each height to provide the confidence band. The probability of the actual temperature existing out of these upper and lower bands is calculated.

V. RESULTS

The DDPM model is implemented using a series artificial neural networks (ANN) which are trained from data contained in the knowledge-base. The knowledge-base is created from

real-life data from 52 data centres containing 11,099,376 temperature measurements or 1,387,422 s-curves. Extracting only s-curves in front of servers this reduces to 32,888 s-curves which further reduces to 18,138 when cleansed for missing data. Rack Adjacency Identification (RAI) is set to extract server racks which have other server racks on each side (R-R-R), resulting in 4,781 s-curve matches and 38,248 temperature measurements. After clustering the s-curves, we restrict the result to clusters with 20 or more members yielding 41 clusters from a total of 988.

When the ANNs are trained, prediction can be made by feeding the input features to the input nodes of the ANNs. There are 7 ANNs in total. One ANN is trained for cluster prediction from the knowledge-base while the others are for individual temperature predictions for each vertical temperature layer (in 305mm (1ft) layers). Temperature layer 1 is transposed layer, always equal to zero. Layer 8 is temperature over the server rack and is used as the ceiling temperature. So only layers 2-7 have an ANN trained.

The knowledge-base data is split into 75% training and 25% testing representing 1600 training and 660 testing samples randomised from the 2260 knowledge-base s-curves. The randomisation of the samples is checked to ensure all 41 clusters are represented in both the training and test data. Table I shows the cluster coverage between training and testing data, including the missing clusters from the result.

The structure of the ANN requires tuning of the number of node layers, node count per layer and the threshold for the partial derivatives of the error function. Different structures were tested with some performing better than others. The criteria for performance judgement is based on minimising the RMSE between the actual temperature measurements and either the predicted cluster means or predicted temperatures per layer depending on which ANN is being analysed. Additionally for the clustering prediction ANN, the accuracy of the cluster prediction to the actual cluster assigned is measured as the percentage of correctly matched clusters. Although many of the tested ANN structures produced similar results, this validation was carried out with ANN structures listed in Table II and shown in the figures following.

In training the ANNs, categorical data such as the data centre type and cluster number require a binary representation with a node for each possible value. Therefore the input nodes require a DCTYPE node for each possible value, similarly the cluster number output requires a node for all 41 cluster types.

A. DDPM ANN Prediction Results - Cluster

The trained model contains 41 possible cluster outcomes from the data. A probability for each outcome is calculated by the neural network with all scores equal to 1 when summed. The highest score is selected as the prediction of the cluster. The test data is feed into the DDPM model and the outcome of the prediction accuracy is shown in Table III.

The result table shows that the predicted cluster to actual cluster is 63% correct for the input features specified. The remaining 37% are assigned to a different cluster other than

Coverage	Cluster <i>k</i>	Clusters included in the prediction																							
	Test	1	2	4	5	6	7	8	9	11	12	17	19	20	23	26	27	30	33	35	36	37	38	40	41
	Training	15	83	31	65	5	6	11	12	10	12	5	17	5	29	29	48	10	14	4	14	17	5	19	
	Predicted	35	188	75	143	17	26	22	26	17	32	16	59	26	70	62	155	25	42	28	18	31	33	15	54
		14	136	27	111	5	11	11	1	8	22	22	51	20	39	36	56	9	12	18	4	13	10	5	19

Coverage	Cluster <i>k</i>	Cluster missing from the prediction																
	Test	3	10	13	14	15	16	18	21	22	24	25	28	29	31	32	34	39
	Training	12	6	20	6	18	8	6	10	17	9	10	11	9	8	20	5	9
	Predicted	12	16	38	21	23	25	16	16	16	33	28	20	32	12	40	17	15
		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Totals	Check	660	
	Test	Test	660
	Training	Training	1600
	Predicted	Predicted	660

[illegible]

the actual cluster. Of the 41 clusters, the prediction utilises only 24 clusters and does not assign any test samples to the 17 clusters that are missing from the result.

Interestingly, since the ANN calculates the probabilities of all clusters with the highest probability value being the one selected as the prediction result, it also inherently enhances the categorical cluster data by providing a proximity of clusters to each other through the probability values. Without this there is no quantitative value to judge similarity of clusters to one another. Therefore, although the results of cluster matching is 63%, the remaining clusters are assigned to clusters which are close to the correct cluster. Fig. 5 shows the probabilities of the prediction outcome for a single prediction. In this case the predicted cluster index is 26 while the actual cluster index is

Cluster #		
Predicted	Actual	Match
4	4	1
26	26	1
41	41	1
5	5	1
2	3	0
2	2	1
12	12	1
27	27	1
.	.	.
.	.	.
5	5	1
<i>Correctly matched</i>		417
<i>Test data size</i>		660
Accuracy		63%

TABLE III
CLUSTER MATCHING RESULTS OF DDPM MODEL. TABLE SHOWS PARTIAL
VIEW OF COMPLETE RESULT.

24. The sum of all probabilities equals 1.

B. DDPM ANN Prediction Results - Temperature Prediction

The trained ANN for the temperature prediction for each temperature layer are fed with the same sample test data as the clustering prediction ANN. The result of each individual temperature layer prediction for a given test sample are joined together to construct an s-curve which represents a complete prediction.

The prediction error is calculated as the RMSE between the actual temperature s-curve and the reconstructed s-curve from the individual ANN predictions. The overall RMSE is 0.76°C and on a layer by layer bases is 0.50, 0.57, 0.64, 0.82, 0.98, 0.95 for layers L2-L7.

Fig. 6 shows the spread of actual data, predicted cluster centroid data and the predicted cluster temperatures constructed as s-curves.

Comparing the prediction error of both the cluster centroid and individual layer temperatures as overlaid histograms in Fig. 7, the individual layer predictions show an overall small RMSE improvement of 0.07°C . Fig. 8 provides a scatter plot for all test data results for each of the layers L2-L7 for both temperature prediction and cluster centroid prediction against the actual temperature values.

The DDPM model utilises the prediction of the individual layer temperatures, reconstructed as a complete s-curve as the inlet temperature distribution prediction. The cluster prediction is used to provide upper and lower bounds to the prediction and are set at the maximum and minimum temperatures values per layer for all members of the predicted cluster in the knowledge-base. Fig. 9 shows the prediction results for some test samples. The prediction includes the temperature prediction, cluster centroid prediction, the upper and lower spread of the in-cluster members contained in the knowledge-base and the actual temperature. Inclusion of the upper and lower bounds in the prediction adds a level of confidence in the prediction accuracy. Taking the upper and lower bounds into account in the prediction, the error between the actual temperature crossing the bounded area in the graph has an RMSE of 0.25°C . The probability of crossing the bounded area at any layer is 0.12 over all s-curves.

The results of the ANN temperature and cluster prediction together with the accuracy of the bounded confidence range and are summarised in Table IV where

- 1) *Cluster accuracy RMSE* is the in-cluster member accuracy to the cluster centroid expresses as RMSE.
- 2) *ANN cluster prediction accuracy RMSE* is the actual temperature to the cluster centroid expressed as RMSE.
- 3) *ANN temperature prediction accuracy RMSE* is the actual temperature to the predicted temperature expressed as RMSE.
- 4) *Cluster bounds accuracy* is calculated based on how far outside the bounds the actual temperature is for each prediction point on each s-curve expressed as RMSE.
- 5) *Number of occurrences actual temperature outside the bounds* is the number of times a single temperature value lies out of the upper or lower bounds.

Fig. 6. Actual data compared to predicted cluster centroids and predicted temperatures.

- 6) *probability actual temperature out of bounds* is the probability of an actual temperature crossing the upper or lower bounds.

The ANN model was developed using R Software [16] with the neuralnet package [13]. The model execution performance is very fast. When tested on an i7 laptop computer with 16GB RAM a performance of 110,000 s-curve calculation per second were obtained. This includes calculations for all 7 ANNs.

Fig. 7. Distribution of prediction errors for both the cluster centroid prediction and the temperature prediction. Figures show that the temperature prediction yields a more accurate result

		Layer								Total
		L1	L2	L3	L4	L5	L6	L7	L8	
1	Clustering Accuracy RMSE	0.00	0.44	0.45	0.47	0.53	0.57	0.58	0.61	0.52
2	ANN Cluster Prediction Accuracy RMSE	0.00	0.51	0.60	0.85	0.96	1.17	1.10	-	0.83
3	ANN Temperature Prediction Accuracy RMSE	0.00	0.50	0.57	0.64	0.82	0.98	0.95	-	0.76
4	Cluster Bounds Accuracy RMSE	0.00	0.02	0.01	0.23	0.24	0.41	0.33	0.45	0.25
5	# Occurances actual T out of bounds	0	14	5	146	93	103	123	91	-
6	Probabilities actual T out of bounds	0	0.02	0.01	0.22	0.14	0.16	0.19	0.14	0.12

TABLE IV
SUMMARY OF FULL DDPM RESULTS

Fig. 8. Scatter plot of actual temperatures against predicted temperatures in red and predicted cluster centroids in green.

Fig. 9. Prediction results

VI. CONCLUSIONS

The data centre is a complex dynamic environment in terms of airflow and temperature distribution. Current physics-based modeling approaches require solving the full numeric calculations for the complete data centre domain which can be expensive in term of computation time and is usually done offline. Our approach is based on machine learning methodologies are used over empirical data. Initially a knowledge-base is created to decompose the data into sets (clusters) of s-curve data with similarities in the shape of the s-curves. Physical features which are used to identify the cluster sets are extracted from the data and together with the temperature s-curves form the knowledge-base which in turn is used to train the model. The model itself comprises of a series of artificial neural networks trained with cluster output and temperature output and using the same input criteria. The performance of the ANNs is $\approx 110,000$ s-curve predictions per second with an overall prediction to actual temperature accuracy of 0.76° RMSE. The temperature prediction is accompanied by upper and lower range obtained from the maximum and minimum values in the predicted cluster set with a just a 0.12 probability that one prediction point on the s-curve will cross the upper

or lower bounds.

ACKNOWLEDGMENT

We wish to acknowledge the support of various IBM data centre staff who have helped us over the years but especially Hendrik Hamann.

REFERENCES

- [1] M. A. P. Bertoldi and L. Castellazzi, "Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency," in *Energies*, vol. 10, 2017, p. 1470.
- [2] H. R. H. Z. S. X. C. Li and C. Hu., "Optimizing energy consumption for data centres," in *Renew. Sustain. Energy Rev. Vol. 58*, pp. 674691, 2016.
- [3] A. S. S. D. S. R. B. M. H. J. K. E. M. N. H. I. Azevedo and W. Lintner, "United states data center energy usage report," in *Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775*, 2016.
- [4] I. EirGrid Group, "All-island generation capacity statement 2017-2026," 2017.
- [5] P. Institute, "Research report: Cost of data center outages," January 2016.
- [6] M. Toulouse, G. Doljac, and C. Bash., "Exploration of a potential-flow-based compact model of air-flow transport in data centers," in *Proceedings of the 2009 International Mechanical Engineering Congress and Exposition*, November 2009.
- [7] J. W. VanGilder, X. Zhang, and C. M. Healey, "Data center airflow prediction with an enhanced potential flow model," in *Thermal Management; Data Centers and Energy Efficient Electronic Systems.*, vol. doi:10.1115/ipack2013-7307, 2013.
- [8] H. F. Hamann, V. Lopez, and A. Stepanchuk, "Thermal zones for more efficient data center energy management," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2010 12th IEEE Intersociety Conference on*, June 2010, pp. 1–6.
- [9] V. Lopez and H. F. Hamann, "Measurement-based modeling for data centers," in *Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2010 12th IEEE Intersociety Conference on*, June 2010, pp. 1–8.
- [10] R. L. J. H. M. Rebow and B. Norton, "A data centre air flow model for predicting computer server inlet temperatures," in *IEEE ITherm 2017. The Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems, May 30 - June 2, 2017, Orlando, Florida, USA*, May 2017.
- [11] M. I. H.F. Hamann, M. Schappert and A. C. T.V Kessel, "Methods and techniques for measuring and improving data center best practices," in *2008 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2008.
- [12] H. F. Hamann, M. Schappert, M. Iyengar, T. van Kessel, and A. Claassen, "Methods and techniques for measuring and improving data center best practices," in *ITHERM*, 2008.
- [13] S. F. F. G. M. S. S. M. Mueller, "R neuralnet package," <https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf>.
- [14] F. G. S. Fritsch, "Neuralnet: Training of neural networks," *The R Journal*, vol. 2/1, June 2010.
- [15] I. G., "A scaling law for the validation-set training-set size ratio," 1997.
- [16] "R project," <https://www.r-project.org/> (accessed Jan 2018).