Physics-informed Denoising Autoencoders for missing data imputation in commercial buildings: an ablation study

ANTONIO LIGUORI*, RWTH Aachen University, Germany
MATIAS QUINTANA, National University of Singapore, Singapore
CHUN FU, National University of Singapore, Singapore
CLAYTON MILLER, National University of Singapore, Singapore
JÉRÔME FRISCH, RWTH Aachen University, Germany
CHRISTOPH VAN TREECK, RWTH Aachen University, Germany

The aim of this paper is to propose the use of Physics-informed Denoising Autoencoders (PI-DAE) for missing data imputation in commercial buildings. In particular, the presented method enforces physics-inspired soft constrained to the loss function of a Denoising Autoencoder (DAE). In order to quantify the benefits of the physical component, an ablation study between different DAE configurations is conducted. First, three univariate DAEs are optimized separately on indoor air temperature, heating and cooling data. Then, two multivariate DAEs are derived from the previous configurations. Eventually, a building thermal balance equation is coupled to the last multivariate configuration to obtain PI-DAE. The results proved the robustness of PI-DAE over varying missing intervals. Additionally, the average reconstruction error on the indoor air temperature data decreased by up to 6.3%, compared to the same model configuration without physical component.

CCS Concepts: • **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

Additional Key Words and Phrases: Physics-informed neural networks, Denoising autoencoder, Deep learning, Missing data

ACM Reference Format:
Antonio Liguori, Matias Quintana, Chun Fu, Clayton Miller, Jérôme Frisch, and Christoph van Treeck. 2018. Physics-informed Denoising Autoencoders for missing data imputation in commercial buildings: an ablation study. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2018), 10 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Building retrofit refers to the process of renovating a building during its operation phase, particularly for energy efficiency and insulation [29]. As such, it is of major importance in order to reduce the environmental footprint of the built environment. In the last years, the increasing number of installed meters has enabled the use of high-resolution data for building energy simulation models (BEMs) [6]. These data are usually used for calibration purposes, in order to correctly simulate the behavior of a building [3]. When using BEMs for retrofit analysis,

*Corresponding author.

Authors' addresses: Antonio Liguori, liguori@e3d.rwth-aachen.de, RWTH Aachen University, Germany; Matias Quintana, matias.quintana.r@gmail.com, National University of Singapore, Singapore; Chun Fu, patrickfu0302@gmail.com, National University of Singapore, Singapore; Clayton Miller, clayton@nus.edu.sg, National University of Singapore, Singapore; Jérôme Frisch, frisch@e3d.rwth-aachen.de, RWTH Aachen University, Germany; Christoph van Treeck, treeck@e3d.rwth-aachen.de, RWTH Aachen University, Germany.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2018 Association for Computing Machinery.
2476-1249/2018/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

the objective should be to improve both the energy efficiency and occupant thermal comfort [24]. Therefore, calibrating these models against high-resolution heating, cooling and indoor air temperature data is an important issue [7, 17].

As observed by Baltazar and Claridge [4], retrofit analysis might be hindered by the quality of the collected data. Building monitoring datasets are indeed characterized by a large number of anomalies and missing data that might be caused by a variety of reasons [20]. A study performed on a database containing energy ~~use~~ recordings from over 600 buildings in US~~;~~ ~~found out~~ that around 60% of ~~the~~ missing data were below six consecutive hours [4]. Considering that daily totals might be useful for building savings determination, incorrect estimation of these faulted data points might lead to wrong decision making [7]. For instance, the presence of this short-term gaps might lead to the discarding of the whole day of observation, with subsequent loss of relevant information.

In order to avoid this scenario, imputation should be performed. Imputation is the process of replacing missing data with new estimated values [10]. In ~~the~~ literature, several statistical and machine learning-based imputation approaches have been proposed [12]. As observed by Emmanuel et al. [12], simple statistical imputation techniques, such as the mean or linear interpolation, are often used ~~for~~ their simplicity. However, these same techniques are also prone to produce inaccurate estimations. More advanced data-driven solutions, such as regression or machine learning-based imputation models, are usually necessary to achieve more stable and precise missing value replacements. However, these models suffer from the same issues as their predictive counterparts. In other words, they require an extensive amount of historical data for accurate predictions [12]. This could be a major problem, especially since the quality of a dataset that needs to be imputed is, by definition, already poor.

In case of building energy systems, there is a variety of physical laws that pure data-driven models ignore at the beginning of the learning process [35]. Encoding such knowledge from scratch results in an increasing amount of information ~~needed~~ by the algorithm. To solve this issue, the use of machine learning for solving partial differential equations (PDEs) has been recently proposed as a valid approach [19]. By encoding prior physical knowledge in the form of PDEs, these models can indeed achieve more accurate predictions by using a smaller amount of monitoring data [28]. Physics-informed Neural Networks (PINNs), recently proposed by Raissi et al. [28], are a family of neural networks that embed~~s~~ physical principles by applying soft constraints to the loss function. Here, the authors took advantage of the automatic differentiation principle of neural networks, in order to approximate the PDEs' residuals of generic physical problems. The results proved the effectiveness of the method when applied to different forward and inverse problems in fluid and quantum mechanics and reaction-diffusion systems.

So far, PINNs have been applied to different fields, such as biophysics, quantum chemistry, material science and geophysics [19]. However, to the best of the authors knowledge, the application to missing data imputation problems has not been yet investigated. When implementing an imputation model, it is particularly important that the reconstruction error is independent on the particular missing component. This is crucial to ensure the robustness of the imputation process [21]. In ~~the~~ literature, this goal has ~~been already~~ achieved by defining a specific data augmentation technique for Denoising Autoencoders (DAEs) [21]. The presented solution consisted in stacking together synthetic copies of a same training sample with pseudo-random masking noise. In the scope of this paper, it is aimed to improve the aforementioned approach by encoding prior physical knowledge. In particular, the prior physical knowledge refers to a simple resistance-capacitor (RC) model able to bind together the imputed heating, cooling and indoor air temperature data.

In summary, the contribution of this paper is to propose the use of Phyiscs-informed Denoising Autoencoders (PI-DAEs) for imputing low-quality building monitoring datasets. In order to quantify the benefits of using PI-DAE, an ablation study between different model configurations is performed. The used dataset was collected in an office building located in Berkeley, California and ~~it~~ is available open-source [23]. It is shown how Convolutional Denoising Autoencoders with data augmentation (CAE + Aug), as proposed in Liguori et al. [21], can be further improved by applying physics-inspired soft constrained to the loss function. By guiding the reconstructed heating,

cooling and indoor air temperature daily profiles towards physically meaningful values, the presented approach might be particularly valuable for improved retrofit analysis.

2 LITERATURE REVIEW

As discussed in the previous Section, pure data-driven models, such as standard neural networks, tend to rely heavily on the quality and size of the used data. This is generally due to the lack of initial knowledge regarding the underlying system. Physics-informed Machine Learning has been widely explored in different engineering areas in which the general system principles can be described using physical models, yet the volume of data that need to be proceeded exceeds the limit of conventional data analytic methods. Among others, Physics-informed Machine Learning has been widely explored in computational fluid dynamics (CFD) [31, 33], applied CFD in the context of climate modeling [15], geoscience modeling [18] as well as non-linear dynamic systems [27] and heat transfer [36].

The Physics-informed Data-Driven modeling has recently gained significant attention in the building energy modeling community. In one of the pioneering works on this topic, Drgona et al. [11] proposed a constrained Recurrent Neural Network (RNN) to model buildings' thermal dynamics. The encoded underlying graph structure is inspired by the physical building components. The results pointed out that the model can be generalized well to different buildings and could represent the buildings' thermal dynamics including the indoor climate conditions with high validity.

The initial results from Drgona et al. [11] demonstrated that applying physics-inspired soft constrained to the loss function of a neural network is beneficial for modeling the thermal dynamics of a building with limited data samples. This encouraged further research in the literature [5, 8, 9, 14, 25, 32, 34]. Gokhale et al. [14] implemented a Physics-informed Neural Network to model the temporal evolution of room temperature, power consumption and temperature of the building thermal mass. In particular, two different variants were proposed. The first configuration consisted of an encoder and dynamics modules, while the second was a simple Fully Connected network. To introduce physics ~~in~~ the network, a grey box approach with a two resistances and two capacitors (2R2C) model was used. The results showed that both ~~the~~ network architectures were able to accurately predict the room temperature. Additionally, the models proved to be data-efficient, hence requiring less training data. Di Natale et al. [8] proposed a Physically Consistent Neural Network (PCNN) for building zone thermal modelling. The model consisted of two parts, namely a black-box and a physics-inspired module. In particular, the thermal dynamics of the building was described by means of an ordinary differential equation (ODE) which represented a simple RC model. Even if the PCNN showed similar performance to the single neural network, the results proved that the temperature prediction could remain physically consistent over the input data. The same authors extended their work to multi-zone modeling, in order to consider the entire building [9]. The improved PCNN could outperform the used grey-box and black-box baselines. According to the authors, differently from PINNs, PCNNs can indeed provide guarantees of physical consistency. Nagarathinam et al. [25] implemented a PINN-based thermal model to predict air temperature, humidity and wall temperature of a building. Additionally, they validated their model for model predictive control (MPC). The results confirmed that the proposed approach could respect the underlying physics of the system over different conditions. Finally, the method could perform better than the Long Short-Term Memory Network (LSTM) used as baseline and consume 24% less energy with an increased user comfort of 2%. Xiao et al. [34] presented a Physically Consistent Deep Learning (PCDL) model for modeling the thermal dynamics of a building. Furthermore, they integrated the model into an MPC controller in order to evaluate the reduction of energy consumption and improvement of thermal comfort. By providing physical guarantees, the PCDL-based controller could reduce the energy consumption by a maximum of 8.9% and improve the thermal comfort by a maximum of 64%. Chen et al. [5] proposed a PINN for demand response control in grid-integrated buildings. In particular, the model consisted of a Fully Connected Network with a

2R2C model to provide prior physical knowledge. This method could outperform purely data-driven models for predicting the indoor air temperature and thermal load demand of different types of buildings. Finally, Wang and Dong [32] developed a Physics-informed Input Convex Neural Network for indoor environment data time-series prediction. The model was further integrated into an hierarchical data-driven predictive control (HDDPC) for space cooling and airside coil loads minimization. It was observed that the proposed network could provide physically consistent predictions of indoor air temperature and ~~CO2~~ data. Finally, the HDDPC could achieve a reduction of cooling and airside coil energy of about 35% and 70%, respectively.

In summary, in the building energy modeling community, the Physics-informed Data-Driven modeling has been mainly used for predicting the indoor environmental conditions and power consumption of a building. This has led to the first attempts to integrate these models into the MPC controllers of buildings. However, no study has been found to deal with missing data imputation problems. In the scope of this paper, it is proved how PINN-based imputation models can indeed provide more robust and accurate replacements for missing building monitoring data.

3 METHODOLOGY

3.1 Dataset description

The used dataset was collected in a multi-storey office building located in Berkeley, California. For detailed information about the building, the reader is referred to Luo et al. [23]. Out of three years of measurements, only the last one is considered. This is based on the limited data availability concerning the heat pump (HP) operation. Since the cooling and heating flow rates were not available, these are derived as follows:

$$\dot{Q}_{cool_i} = \dot{m}_{sa_i} \cdot \rho_a \cdot c_{p_a} \cdot (T_{sa_i} - T_{ma_i}) \,, \tag{1}$$

$$\dot{Q}_{heat} = f \cdot \dot{m}_{hw} \cdot \rho_w \cdot c_{p_w} \cdot (T_{shw} - T_{rhw}) = f \cdot \dot{Q}_{hw} \,, \tag{2}$$

where $\dot{Q}_{cool_i}$ is the cooling flow rate from the i-th roof terminal unit (RTU), $\dot{m}_{sa_i}$ is the i-th RTU filtered supply air flow rate, ρ_a and ρ_w are the densities of the air (1.204 $\frac{kg}{m^3}$) and water (1000 $\frac{kg}{m^3}$), c_{p_a} and c_{p_w} are the specific heat capacities of the air (1.006 $\frac{kJ}{Kg \cdot K}$) and water (4.200 $\frac{kJ}{Kg \cdot K}$), T_{sa_i} is the i-th RTU supply air temperature, T_{ma_i} is the i-th RTU mixed air temperature, f is an unknown correction factor, $\dot{m}_{shw}$ is the HP supply hot water flow rate, T_{shw} is the HP supply hot water temperature, T_{rhw} is the HP return hot water temperature and $\dot{Q}_{hw}$ is the reheat water heat flow rate.

Assuming perfect mixing of the air, the indoor air temperature of the thermal zone served by the i-th RTU is considered equal to the i-th RTU return air temperature T_{ra_i}. Additionally, for simplicity, data are aggregated at the building level. Here, the average indoor air temperature ($T_{ra_{avg}}$) and average outdoor air temperature ($T_{oa_{avg}}$) data are obtained by the mean operation, while the total cooling flow rate ($\dot{Q}_{cool_{tot}}$) is obtained by summing the contribution of each RTU. Finally, in line with the previous studies from which the ~~A~~utoencoders are adopted, the dataset is resampled to a 30 minutes-based frequency.

3.2 Building thermal balance

In order to implement a Physics-informed Neural Network, a PDE which guides the outputs towards physically meaningful values has to be determined. It is observed that the thermal balance of the introduced building can be described by means of the following ODE [13]:

$$\dot{Q}_{storage} = M \cdot c_{p_a} \cdot \frac{dT_{ra_{avg}}}{dt} = \dot{Q}_{env} + \dot{Q}_{ven} + \dot{Q}_{int} + \dot{Q}_{sol} - \dot{Q}_{cool_{tot}} + \dot{Q}_{heat} \,, \tag{3}$$

where M is the air volume mass. In the presented case study, it is difficult to obtain correct estimation for the ventilation ($\dot{Q}_{ven}$), internal ($\dot{Q}_{int}$) and solar ($\dot{Q}_{sol}$) heat flow rates. Therefore, in line with previous works in the literature [8, 14], these contributions are ignored. The heat flow rate from the environment ($\dot{Q}_{env}$) is defined as follows:

$$\dot{Q}_{env} = \frac{T_{oa_{avg}} - T_{ra_{avg}}}{R_{ra}} , \tag{4}$$

where R_{ra} is the thermal resistance of the external wall. By rearranging all the terms, Equation 5 is obtained:

$$\frac{dT_{ra_{avg}}}{dt} = a_1 \cdot (T_{oa_{avg}} - T_{ra_{avg}}) - b_1 \cdot \dot{Q}_{cool_{tot}} + c_1 \cdot \dot{Q}_{hw} , \tag{5}$$

where a_1, b_1 and c_1 are unknown physics-based parameters that should be greater than zero to be physically meaningful. By approximating Equation 5 with the forward difference method [30], a final formulation which is true for each timestep t is obtained:

$$(T_{ra_{avg_{t+1}}} - T_{ra_{avg_t}}) - (a \cdot (T_{oa_{avg_t}} - T_{ra_{avg_t}}) - b \cdot \dot{Q}_{cool_{tot_t}} + c \cdot \dot{Q}_{hw_t}) = 0 , \forall t . \tag{6}$$

3.3 From DAE to PI-DAE

The selected univariate DAE is a Convolutional Denoising Autoencoder implemented as part of Liguori et al. [20]. A detailed overview of the model architecture is presented in the Appendix of the successive paper [21]. It was already proved that the aforementioned Denoising Autoencoder could successfully generalize to alternative buildings [21, 22]. For that reason, the network architecture is preserved, while just some hyperparameters are optimized using Optuna [2]. As continuous missing scenarios are usually more complex than random ones [21], corrupted elements are selected based on the former. In particular, the corruption rate (CR) varies between few hours (20%) and around 20 hours (80%). The reader is referred to Liguori et al. [21] for additional information about the used continuous missing masking noise pseudo-algorithm. Eventually, three different univariate DAEs for missing indoor air temperature (Univariate_DAE_1), heating (Univariate__DAE_2) and cooling (Univariate__DAE_3) data imputation are optimized. Additionally, two different multivariate DAEs, where the missing intervals for each input time-series are assumed to occur at the same time, are implemented and optimized based on the same procedure. Namely, Multivariate_DAE_1 takes as input corrupted indoor air temperature, heating and cooling data. Multivariate_DAE_2 takes as input corrupted indoor air temperature, heating, cooling and full outdoor air temperature data. Here, since weather data can be usually obtained from nearby weather stations, the outdoor air temperature is not corrupted.

The proposed PI-DAE can be observed in Figure 1. It consists of two parts, namely a multivariate DAE with outdoor air temperature (Multivariate_DAE_2) and an approximated ordinary differential equations based on Equation 6 that tie the reconstructed outputs together (Approximated ODE). Here, the Approximated ODE depends on the unknown physics-based parameters that are optimized together with the weights (w) and biases (b) of the DAE component. In particular, in order to quantify the effects of the physical component on the Autoencoder, the hyperparameters of PI-DAE are the same as those of Multivariate_DAE_2.

The total loss function of the model ($Loss$) is therefore given by a regression loss ($Loss_{Mutlivariate\_DAE\_2}$) and a physics-based loss ($Loss_{ApproximatedODE}$), as follows [16]:

$$Loss = Loss_{Mutlivariate\_DAE\_2} + Loss_{ApproximatedODE} . \tag{7}$$

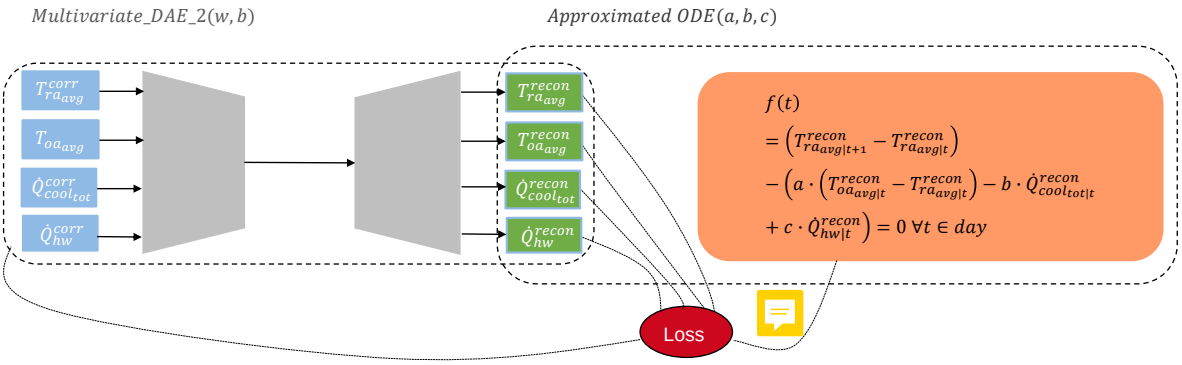

Fig. 1. Representation of the proposed PI-DAE with approximated building energy balance ODE. Figure partially reproduced based on Jagtap et al. [16].

3.4 Building operation periods

Embedding physical knowledge in a deep learning model, foresees the input and output features being tied by some specific mathematical expression. In the scope of this paper, it is assumed that the reconstructed variables follow the behavior represented by Equation 6. However, following the same procedure, the proposed Multivariate DAE might be coupled with any expression enforcing physical laws on the outputs. In the presented study, Equation 6 depends on strong hypothesis, e.g. absence of solar gains, which might lead to biased and inaccurate results. In order to respect these hypothesis, the considered variables should be highly correlated with each other. For instance, a low correlation between the cooling flow rate and the indoor air temperature could indicate that other factors, such as occupant behavior, should be taken into account.

In order to quantify the correlation attributes, the Pearson Correlation Coefficient (PCC) is used, since it is widely adopted in the literature [26]. Considering that higher PCCs are observed on days with higher heating and cooling flow rate variability, data are filtered based on the interquartile range (IQR) of these two variables. Namely, only days over a certain threshold are selected. This is based on the literature, where the IQR is a common measure of data variability [1]. Figure 2 shows different scatterplots that represent the correlations among the analyzed variables (vertical axis), with different IQR thresholds for the heating and cooling flow rates (horizontal axis). The black dots represent the points for which the positive or negative correlation is maximum. The red dots represent the points for which the thresholds is selected. It is observed that for most of the scatterplots, the red dots match the black dots. This indicates that the selected days have a higher correlation compared to the full dataset. Based on this consideration, models' evaluation is performed on two different building operation periods. Namely, Case 1 is the full dataset ~~which includes~~ 363 days of observations. Case 2 is the reduced dataset ~~which comprises~~ 19 days of observation with the highest correlation among the observed variables.

4 RESULTS AND DISCUSSION

In order to evaluate the performance of PI-DAE and to determine the impact of the physics-informed loss function, an ablation study is performed among different DAE configurations. The daily corruption rate is varied between few hours (20%) and around 20 hours (80%). Additionally, to analyze the influence of the number of training data on the models' performance, the training rate is varied from 10 % (36 days for Case 1 and 1 day for Case 2) to 50% (181 days for Case 1 and 9 days for Case 2). Eventually, following the data augmentation strategy to avoid DAEs' overfitting [21], training data are augmented 10 times for Case 1 and 80 times for Case 2.

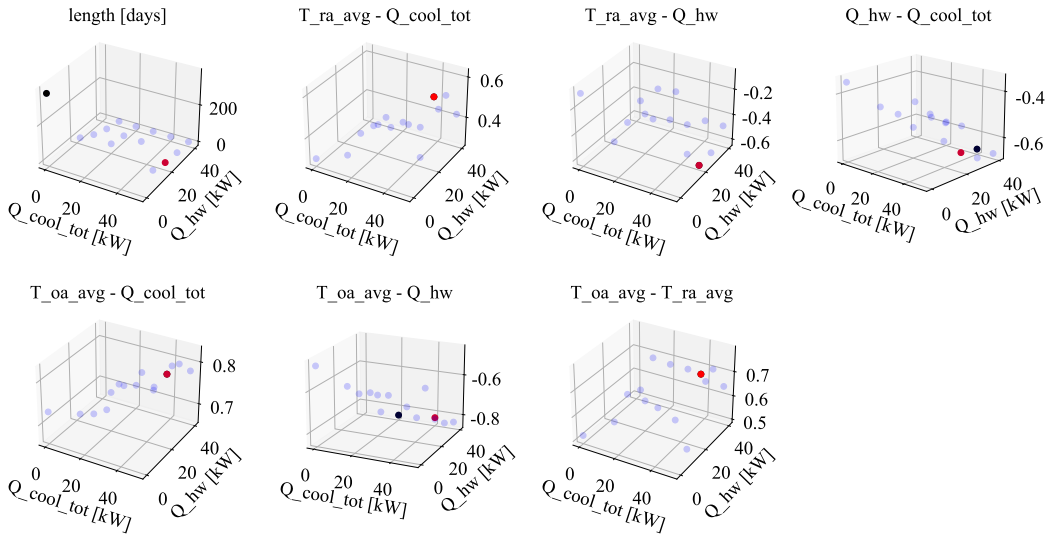

Fig. 2. Scatterplot for different thresholds of cooling and heating flow rates (horizontal axis). The black dots represent the points for which the positive or negative correlation (vertical axis) is maximum. The red dots represent the points for which the thresholds is selected.

The learning curves for the analyzed models can be observed in Figure 3, where the root mean squared error (RMSE) has been averaged over different corruption rates. For a detailed explanation about the used RMSE, the reader is referred to Liguori et al. [21]. It is noted that the average RMSE on the indoor air temperature data decreases if a physics-informed loss function is added to Multivariate_DAE_2. In particular, PI-DAE can achieve a maximum of 6.3% and 5.4% lower average RMSE than the same model configuration without physics-informed loss, respectively for Case 2 and Case 1. In case of cooling and heating flow rates, this impact is much lower. For Case 2, PI-DAE can achieve a maximum of 2.6% and 1.5% lower average RMSE, respectively for cooling and heating data. This reduction is negligible for Case 1. Finally, it is noted how the univariate model can generally perform better than PI-DAE, except for the cooling flow rate. For the latter, PI-DAE outperforms the univariate model with a 68.8% and 45.8% lower average RMSE, respectively for Case 2 and Case 1. This behavior can be explained by the higher correlation coefficients (positve and negative) between the cooling flow rate and the rest of the analyzed variables, as observed in Figure 2.

In order to analyze how the performance of the models change over different corruption rates, the standard deviation of the reconstruction errors is presented in Figure 4. Here, Multivariate_DAE_1 is not represented due to the high RMSEs shown in Figure 3. It is observed that the RMSE of the univariate Autoencoder has a much ~~larger~~ standard deviation compared to PI-DAE. This indicates that the reconstruction error of PI-DAE is more stable along different corruption rates. Finally, the RMSE of PI-DAE has a standard deviation which is comparable to the same model configuration without physics-informed loss, i.e. Multivariate_DAE_2.

Eventually, the optimized physics-based coefficient for both Case 1 and Case 2 are presented in Figure 5. Prior to model's training, all the coefficients have been initialized to one. It is noted that the optimized coefficients of Case 1 are much smaller than the ones of Case 2. This indicates that there is a small correlation between the temporal evolution of the indoor air temperature and the rest of the analyzed variables (see Equation 6). On the other hand, the impact of the outdoor air temperature is more significant than the other variables, especially for Case 2. This consideration is based on the coefficient a, which is always higher than the starting value.

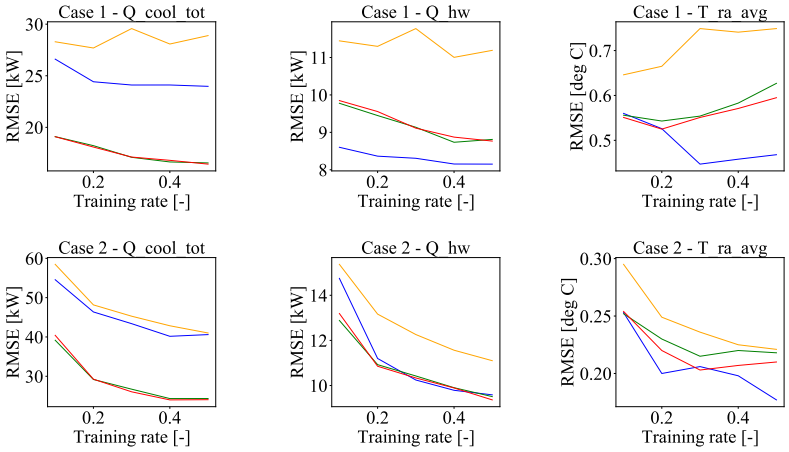

Fig. 3. Learning curves of the models (Univariate_DAE: blue curve, Multivariate_DAE_1: yellow curve, Multivariate_DAE_2: green curve, PI-DAE: red curve) for varied training rates and cases.

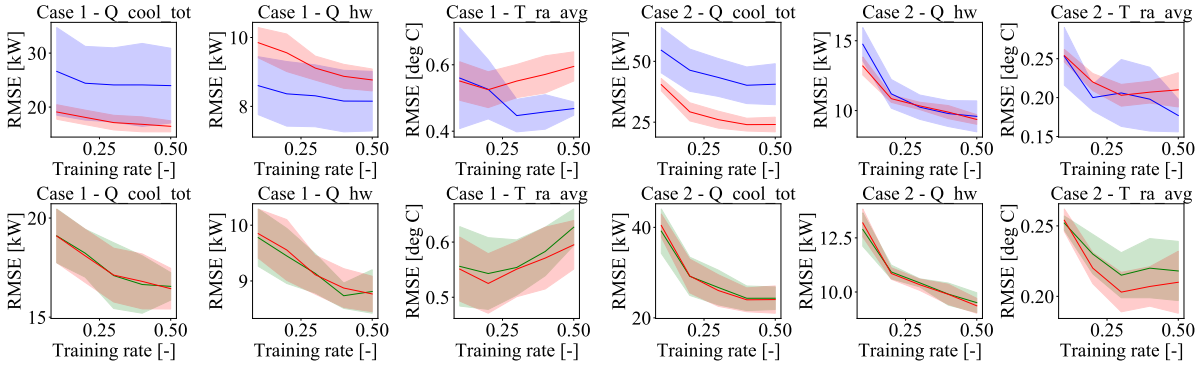

Fig. 4. Top: learning curves of the models PI-DAE (red curve) and Univariate_DAE (blue curve) with standard deviation for varied training rates and cases. Bottom: learning curves of the models PI-DAE (red curve) and Multivariate_DAE_2 (green curve) with standard deviation for varied training rates and cases.

5 CONCLUSION, LIMITATIONS AND FUTURE WORK

This paper presented an ablation study to quantify the effects of embedding physical knowledge in a Denoising Autoencoder for missing data imputation. The proposed PI-DAE combines the potential of DAEs with an approximated building thermal balance ODE which enforces physical laws on the reconstructed features. By guiding the imputed indoor air temperature, heating and cooling data within physically meaningful boundaries, the presented model could support ~~the~~ building retrofit analysis. For that purpose, it is observed that the unknown physics-based coefficients can be optimized together with the parameters of the black-box component.

While the varied training rates did not have a significant impact on the different performance of the models, the analysis confirmed the importance of selecting appropriate building operation periods before applying PI-DAE. Namely, the temporal profiles of the used features should be highly correlated with each other in order to validate

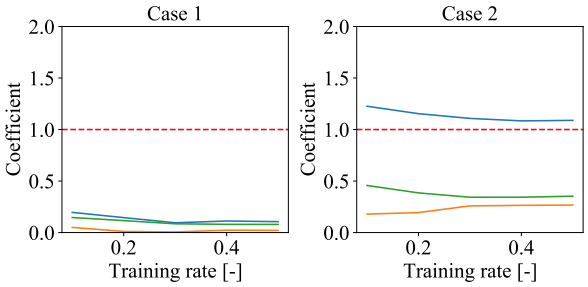

Fig. 5. Optimized physics-based coefficients (a: light blue curve, b: yellow curve, c: green curve) for varied training rates and cases. The starting value of one is represented by the dashed horizontal curve.

the hypothesis of the physical component. Furthermore, it was observed that the use of a physics-informed loss function could decrease the RMSE on the indoor air temperature data by up to 6.3%. On the other hand, the reduction was only up to 2.6% and 1.5% for the heating and cooling flow rates. In this regard, future work should focus on improving the presented method by adding explicit constrains on the reconstructed variables. As PI-DAE might include any mathematical expression enforcing physical laws on the outputs, the potential application of this model goes far beyond imputation in commercial buildings.

6 ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – TR 892/8-1. Simulations were performed with computing resources granted by RWTH Aachen University, Germany, under project rwth0622.

REFERENCES

[1] Shabbir Ahmad, Zhengyan Lin, Saddam Akber Abbasi, Muhammad Riaz, et al. 2012. On efficient monitoring of process dispersion using interquartile range. *Open journal of applied sciences* 2, 04 (2012), 39–43.

[2] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[3] Adriana Angelotti, Maricla Martire, Livio Mazzarella, Martina Pasini, Ilaria Ballarini, Vincenzo Corrado, Giovanna De Luca, Paolo Baggio, Alessandro Prada, Francesco Bosco, et al. 2018. Building energy simulation for Nearly Zero Energy retrofit design: the model calibration. In *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*. IEEE, 1–6.

[4] Juan Carlos Baltazar and David E Claridge. 2002. Restoration of short periods of missing energy use and weather data using cubic spline and Fourier series approaches: qualitative analysis. (2002).

[5] Yongbao Chen, Qiguo Yang, Zhe Chen, Chengchu Yan, Shu Zeng, and Mingkun Dai. 2023. Physics-informed neural networks for building thermal modeling and demand response control. *Building and Environment* 234 (2023), 110149.

[6] Adrian Chong, Yaonan Gu, and Hongyuan Jia. 2021. Calibrating building energy simulation models: A review of the basics to guide future work. *Energy and Buildings* 253 (2021), 111533.

[7] David E Claridge and Hui Chen. 2006. Missing data estimation for 1–6 h gaps in energy use and weather data using different statistical methods. *International journal of energy research* 30, 13 (2006), 1075–1091.

[8] Loris Di Natale, Bratislav Svetozarevic, Philipp Heer, and Colin N Jones. 2022. Physically consistent neural networks for building thermal modeling: theory and analysis. *Applied Energy* 325 (2022), 119806.

[9] Loris Di Natale, Bratislav Svetozarevic, Philipp Heer, and Colin Neil Jones. 2023. Towards scalable physically consistent neural networks: An application to data-driven multi-zone thermal building models. *Applied Energy* 340 (2023), 121071.

[10] A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. 2006. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology* 59, 10 (2006), 1087–1091.

[11] Ján Drgona, Aaron R Tuor, Vikas Chandan, and Draguna L Vrabie. 2020. *Physics-constrained Deep Recurrent Neural Models of Building Thermal Dynamics.* Technical Report. Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

[12] Tlamelo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. 2021. A survey on missing data in machine learning. *Journal of Big Data* 8, 1 (2021), 1–37.

[13] Simone Ferrari and Valentina Zanotto. 2015. *Building energy performance assessment in Southern Europe.* Springer, Chapter 1.1.

[14] Gargya Gokhale, Bert Claessens, and Chris Develder. 2022. Physics informed neural networks for control oriented thermal modeling of buildings. *Applied Energy* 314 (2022), 118852.

[15] Michael F Howland and John O Dabiri. 2019. Wind farm modeling with interpretable physics-informed machine learning. *Energies* 12, 14 (2019), 2716.

[16] Ameya D Jagtap, Kenji Kawaguchi, and George Em Karniadakis. 2020. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *J. Comput. Phys.* 404 (2020), 109136.

[17] Ying Ji and Peng Xu. 2015. A bottom-up and procedural calibration method for building energy simulation models based on hourly electricity submetering data. *Energy* 93 (2015), 2337–2350.

[18] Sadegh Karimpouli and Pejman Tahmasebi. 2020. Physics informed machine learning: Seismic wave equation. *Geoscience Frontiers* 11, 6 (2020), 1993–2001.

[19] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. 2021. Physics-informed machine learning. *Nature Reviews Physics* 3, 6 (2021), 422–440.

[20] Antonio Liguori, Romana Markovic, Thi Thu Ha Dam, Jérôme Frisch, Christoph van Treeck, and Francesco Causone. 2021. Indoor environment data time-series reconstruction using autoencoder neural networks. *Building and Environment* 191 (2021), 107623.

[21] Antonio Liguori, Romana Markovic, Martina Ferrando, Jérôme Frisch, Francesco Causone, and Christoph van Treeck. 2023. Augmenting energy time-series for data-efficient imputation of missing values. *Applied Energy* 334 (2023), 120701.

[22] Antonio Liguori, Romana Markovic, Jérôme Frisch, Andreas Wagner, Francesco Causone, Christoph van Treeck, et al. 2021. A gap-filling method for room temperature data based on autoencoder neural networks. In *BUILDING SIMULATION CONFERENCE PROCEEDINGS*, Vol. 17. 2427–2434.

[23] Na Luo, Zhe Wang, David Blum, Christopher Weyandt, Norman Bourassa, Mary Ann Piette, and Tianzhen Hong. 2022. A three-year dataset supporting research on building energy management and occupancy analytics. *Scientific Data* 9, 1 (2022), 156.

[24] Laurent Magnier and Fariborz Haghighat. 2010. Multiobjective optimization of building design using TRNSYS simulations, genetic algorithm, and Artificial Neural Network. *Building and Environment* 45, 3 (2010), 739–746.

[25] Srinarayana Nagarathinam, Yashovardhan S Chati, Malini Pooni Venkat, and Arunchandar Vasan. 2022. PACMAN: physics-aware control MANager for HVAC. In *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 11–20.

[26] Aqsa Saeed Qureshi, Asifullah Khan, Aneela Zameer, and Anila Usman. 2017. Wind power prediction using deep neural network based meta regression and transfer learning. *Applied Soft Computing* 58 (2017), 742–755.

[27] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. 2017. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561* (2017).

[28] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics* 378 (2019), 686–707.

[29] Emmanouil Thrampoulidis, Gabriela Hug, and Kristina Orehounig. 2023. Approximating optimal building retrofit solutions for large-scale retrofit analysis. *Applied Energy* 333 (2023), 120566.

[30] Christoph Alban van Treeck, ~~Ernst Rank, Gerd Hauser, and Jan Hensen~~, 2010. *Introduction to building performance modeling and simulation.* Ph.D. Dissertation. Habilitationsschrift, Technische Universität München, 2010.

[31] Jian-Xun Wang, Jin-Long Wu, and Heng Xiao. 2017. Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data. *Physical Review Fluids* 2, 3 (2017), 034603.

[32] Xuezheng Wang and Bing Dong. 2023. Physics-informed Hierarchical Data-driven Predictive Control for Building HVAC Systems to Achieve Energy and Health Nexus. *Energy and Buildings* (2023), 113088.

[33] Jin-Long Wu, Heng Xiao, and Eric Paterson. 2018. Physics-informed machine learning approach for augmenting turbulence models: A comprehensive framework. *Physical Review Fluids* 3, 7 (2018), 074602.

[34] Tianqi Xiao and Fengqi You. 2023. Building thermal modeling and model predictive control with physically consistent deep learning for decarbonization and energy optimization. *Applied Energy* 342 (2023), 121165.

[35] Ye Yao, Kun Yang, Mengwei Huang, and Liangzhu Wang. 2013. A state-space model for dynamic response of indoor air temperature and humidity. *Building and environment* 64 (2013), 26–37.

[36] Navid Zobeiry and Keith D Humfeld. 2021. A physics-informed machine learning approach for solving heat transfer equation in advanced manufacturing and engineering applications. *Engineering Applications of Artificial Intelligence* 101 (2021), 104232.