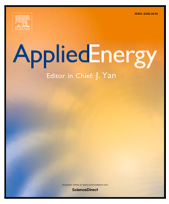
Applied Energy 378 (2025) 124763

Contents lists available at ScienceDirect

Applied Energy

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

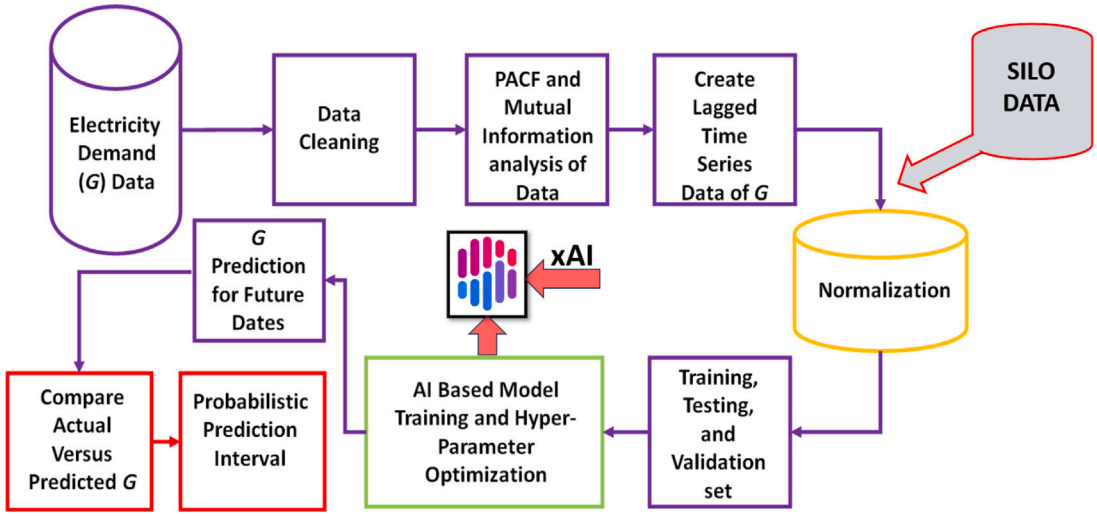


Explainable deeply-fused nets electricity demand prediction model: Factoring climate predictors for accuracy and deeper insights with probabilistic confidence interval and point-based forecasts

Sujan Ghimire [a], Mohanad S. AL-Musaylh [d], Thong Nguyen-Huy [b,c], Ravinesh C. Deo [a,*], Rajendra Acharya [a,g], David Casillas-Pérez [e], Zaher Mundher Yaseen [h], Sancho Salcedo-Sanz [a,f]

[a] Artificial Intelligence Applications Laboratory: School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD, 4300, Australia
[b] Centre for Applied Climate Sciences, University of Southern Queensland, Toowoomba, 4350, QLD, Australia
[c] Faculty of Information Technology, Thanh Do University, Kim Chung, Hoai Duc, Ha Noi, 100000, Vietnam
[d] Department of Information Technologies Management, Management Technical College, Southern Technical University, Basra 61001, Iraq
[e] Department of Signal Processing and Communications, Universidad Rey Juan Carlos, Fuenlabrada, 28942, Madrid, Spain
[f] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, 28805, Madrid, Spain
[g] International Research Organization for Advanced Science and Technology, (IROAST), Kumamoto University, Kumamoto 860-8555, Japan
[h] Civil and Environmental Engineering Department, King Fahd University of Petroleum & Minerals, Dhahran, 31261, Saudi Arabia

GRAPHICAL ABSTRACT



ARTICLE INFO

Dataset link: https://www.energex.com.au

MSC:
0000

ABSTRACT

Electricity consumption has stochastic variabilities driven by the energy market volatility. The capability to predict electricity demand that captures stochastic variances and uncertainties is significantly important in the planning, operation and regulation of national electricity markets. This study has proposed an explainable

* Corresponding author.
E-mail addresses: sujan.ghimire@usq.edu.au (S. Ghimire), mohanad.al-musaylh@stu.edu.iq (M.S. AL-Musaylh), thong.nguyen-huy@unisq.edu.au (T. Nguyen-Huy), ravinesh.deo@usq.edu.au (R.C. Deo), rajendra.acharya@usq.edu.au (R. Acharya), david.casillas@urjc.es (D. Casillas-Pérez), z.yaseen@kfupm.edu.sa (Z.M. Yaseen), sancho.salcedo@uah.es (S. Salcedo-Sanz).

https://doi.org/10.1016/j.apenergy.2024.124763
Received 5 August 2024; Received in revised form 7 October 2024; Accepted 16 October 2024
Available online 9 November 2024

1111

Keywords:
Machine learning models
Cities sustainability and development
Deeply fused nets model
Electricity consumption

deeply-fused nets electricity demand prediction model that factors in the climate-based predictors for enhanced accuracy and energy market insight analysis, generating point-based and confidence interval predictions of daily electricity demand. The proposed hybrid approach is built using Deeply Fused Nets (FNET) that comprises of Convolutional Neural Network (CNN) and Bidirectional Long-Short Term Memory (BILSTM) Network with residual connection. The study then contributes to a new deep fusion model that integrates intermediate representations of the base networks (fused output being the input of the remaining part of each base network) to perform these combinations deeply over several intermediate representations to enhance the demand predictions. The results are evaluated with statistical metrics and graphical representations of predicted and observed electricity demand, benchmarked with standalone models *i.e.*, BILSTM, LSTMCNN, deep neural network, multi-layer perceptron, multivariate adaptive regression spline, kernel ridge regression and Gaussian process of regression. The end part of the proposed FNET model applies residual bootstrapping where final residuals are computed from predicted and observed demand to generate the 95% prediction intervals, analysed using probabilistic metrics to quantify the uncertainty associated with FNETS objective model. To enhance the FNET model's transparency, the SHapley Additive explanation (SHAP) method has been applied to elucidate the relationships between electricity demand and climate-based predictor variables. The suggested model analysis reveals that the preceding hour's electricity demand and evapotranspiration were the most influential factors that positively impacting current electricity demand. These findings underscore the FNET model's capacity to yield accurate and insightful predictions, advocating its utility in predicting electricity demand and analysis of energy markets for decision-making.

1. Introduction

The United Nations advocates for global strategic measures to implement the Sustainable Development Goals (SDGs) for the 2030 Agenda [1]. Out of the 17 Sustainable Development Goals (SDGs), Goal 7 aims to optimize and improve the energy production and utilization system globally [1]. Prediction models for electricity demand (G) are a critical component of modern energy systems. For the construction of an efficient and sustainable energy platform, a reliable prediction model that incorporates the most relevant climate and social factors is essential. In addition to short-term (hourly, daily) and long-term (monthly, seasonal, annual) prediction horizons, G models must include techniques for evaluating uncertainties in electricity use patterns. However, constructing reliable prediction models poses significant challenges. These include, but not limited to the variability of climate conditions, the difficulty to predict social behaviours, and the integration or availability of diverse data sources. Additionally, due to the continuous changes in electricity demand, the models must be robust enough to handle non-linear interactions and adapt to rapidly changing energy consumption trends.

A number of earlier studies recommended soft computing algorithms based on machine learning (ML) algorithms for G predictions. These include statistical methods [2,3], Kernel models [2], regression analysis [4], hybrid ML-based nature inspired algorithms [5], neural network models [3], improved hybrid ML models using data pre-processing approaches [6], Trees models [6], Extreme Learning Machine (ELM) [7], Multiple Linear Regression (MLR) [3], and several others [8,9], Gaussian Process of Regression (GPR) [9] and Maximum Overlap Discrete Wavelet Transform (MODWT)-OS-ELM [7]. Due to the diverse capabilities of such ML models, whose accuracy varies based on data and region, the development of G prediction methods and understanding of their predicted uncertainties is an ongoing research area.

Recently, deep learning (DL) models have become increasingly popular for predicting G. A research study conducted by [10] shows that LSTM networks, convolutional neural networks (CNNs) and multilayer perceptrons (MLPs) perform best for G prediction. The LSTM models are widely used because of their ability to handle long-term dependencies [11], whereas in CNNs, convolution operations are used to extract features, which increases the accuracy of time series prediction by capturing high-level feature representations from multiple time series [12]. Locally connected algorithms have global sharing properties, which reduce training parameters and time, increasing time-series prediction accuracy [13]. In the study of [14], the authors evaluated the prediction ability of the Factored Conditional Restricted Boltzmann Machine (FCRBM) in comparison with Mutual Information-based ANNs (MI-ANNs), Bi-level, LSTM, and ANN-based accurate and fast convergence (AFC-ANNs). In terms of training time, FCRBM was demonstrated to be faster and more accurate than alternative models. In addition, the authors in [15] have connected a CNN and an ANN model to predict French energy demand. For such prediction tasks, Bidirectional LSTM (BILSTM), a variant of LSTM, has been demonstrated to be much faster and more accurate than traditional LSTM models [16]. Bidirectional memory in the BILSTM model is especially useful for exploring both previous and upcoming features; see, for instance, the works of [17,18]. To predict the G data, [19] applied Bi-LSTM model with attention mechanisms and compared with an SVR and a conventional Bi-LSTM. Overall, the study showed that the proposed Bi-LSTM model with an attention mechanism could be a viable and effective predictive model.

Because of the volatility and instability of electrical loads, standalone ML and DL models occasionally cannot precisely extract complex feature correlations in nonlinear and non-stationary G data. Many researchers have thus proposed hybrid models combining DL/ML models to show promising results through CNN/LSTM methods for electricity load prediction compared with non-hybrid models [20,21]. In [22], a hybrid model combining Multilayer Perceptron (MLP), Adaptive Network-based Fuzzy Inference System (ANFIS), and Seasonal Autoregressive Integrated Moving Average (SARIMA) was also proposed where its accuracy was demonstrated by reduced Mean Absolute Percentage Error and faster convergence rate. Other hybrid models applied to time-series prediction problems include LSTMs with Extreme Gradient Boosting (XGBOOST) [23] and the fractional ARIMA with enhanced Cuckoo search [24], outperforming their standalone counterpart models for electricity load prediction problems. In particular, ML, DL and hybrid models have largely been applied for point-based G prediction for distinct, deterministic, and definite outcomes. It is, however, impossible to completely eliminate prediction errors when G is non-stationary and chaotic. Thus, point prediction results can be hard to use in sound decision-making in critical power system infrastructure if these models are used. For electricity demand, the quantification of model uncertainty estimation is crucial in terms of a point-based and probability interval prediction outcome in order to better understand the model's fidelity and variability.

In the light of the aforementioned, the prediction intervals of an electricity demand model are constructed by assuming specific distribution functions. Due to the chaotic nature of G itself, it is typically impossible to determine the exact distribution and therefore, general assumptions must be made. A deviation from such assumptions can also affect the decisions made using the generated prediction intervals as well as on the estimated model parameter values, which could result in an over- or underestimation of the underlying risk of using

such predicted G values in real-time. Fortunately, re-sampling techniques, like Bootstrap (BS), enable the creation of prediction intervals without taking any sorts of distributional assumptions. As opposed to conventional methods such as Lower Upper Bound Estimation (LUBE), Monte Carlo Simulation (MCS), and Quantile Regression (QR), the BS technique uses original samples as a population of resampling.

In [25], a deeply-fused nets (FNET) method which embraces deep fusion or a combination of the intermediate representations of a base network with various other intermediate representations, was proposed. Importantly, this approach simultaneously learned the representations of the base networks, to mimic the highly successful methods such as GoogLeNet or deeply-supervised nets [26] and variants like Highway [27] and ResNet [28]. The FNET method was successfully applied on the CIFAR-10 and CIFAR-100 image-based datasets, to show 93.77–93.98% for the CIFAR-10 and 72.29–72.64% for the CIFAR-100 datasets with the Deep summation (fusion before ReLU) and the Deep summation (fusion after ReLU), respectively. Compared with several state-of-the-art algorithms baseline methods, the FNET model demonstrated significantly better performance. However, the application of the original FNET model has so far been restricted to only image and text-based datasets so its application to the time-series datasets, and especially in G prediction problems could provide new avenues to capitalize on the merits of the deeply-fused nets method.

In this study, we extend and significantly improve the scope and practical applicability of FNET [25] for time-series data fusion. We also adopt residual bootstrapping (BS) to generate prediction intervals of electricity demand in such a way that the proposed BS-based FNET model does not require prior information about the data distribution or the model parameters while it seamlessly adopts explainable artificial intelligence based on Shapley Additive Explanations (SHAP) to demonstrate an interpretable FNETS model for G predictions. The contributions and the novelty of this research are as follows:

1. To develop for the first time a new approach for electricity demand prediction by proposing Deeply Fused Network (FNET) model that seamlessly fuses the CNN and BILSTM algorithms for the point-based and confidence interval predictions.
2. To improve the efficiency of the proposed FNET model considering a fused net system with three fusions: a deep base network (1D-CNN), a set of CNN filters (ranging from 32–128) and a 4-layer BILSTM network with BILSTM unit (ranging from 16–128). As a regression model, we then adopt a single BILSTM layer at the end of the network before the fully connected or dense layer and apply the Scaled Exponential Linear Units (SeLU) activation function for 1D-CNN layer and Rectified Linear Unit (ReLU) for the Dense layer.
3. To evaluate the performance of the proposed FNET model using deterministic and probabilistic metrics against standalone and hybrid models (*i.e.*, BILSTM, LSTMCNN, Deep Neural Network (DNN), Multi-layer Perceptron (MLP), Multivariate Adaptive Regression Spline (MARS), Kernel Ridge Regression (KRR), and Gaussian Process of Regression (GPR)).
4. To improve the practicality of the proposed FNET model by generating the interval predictions of electricity demand that can inform model predicted uncertainties in electricity demand, enhancing the validity of using FNET in real-life scenarios for electricity forecasting.
5. To interpret model behaviour and better understand the underlying factors influencing electricity demand. Here, we adopted the Shapley Additive Explanations (SHAP) to reveal the intricate relationship between key variables and their contribution to the model's predictions.

Our study contributes to the development of a robust model that incorporates climate predictors for enhanced insights into energy markets to predict electricity demand daily with confidence intervals and point-based prediction. Consequently, this study contributes to a new deep fusion model that integrates intermediate representations from base networks (the fused output is used as input by the rest of the base networks) and enhances demand prediction by combining several intermediate representations deeply.

To enhance the contribution of this study, we adopted residual bootstrapping to quantify the uncertainties in the proposed FNET model to advance its practical implementation as previous studies [25] used FNET for only point-based predictions. In particular, the network comprising substantially fused CNN and BI-LSTM model utilizes the primary concept, to perform the fusion over intermediate representations of the base networks rather than just over the final representations. Such fusions are repeatedly performed at intermediate layers with the fused output serving as the input of remaining portion of each base network, and finally the base network is used to generate prediction intervals on the residuals obtained by the difference between observed G and predicted G generated from the proposed FNET model.

To train and evaluate the proposed FNET model, we have utilized historical (*i.e.*, time-lagged) G data as well as respective local climate variables for Annerley, Heathwood, Laidley and Zillmere substations located in Queensland, Australia, along with detailed statistical and probabilistic analysis of model performance (see later, in Table 3).

2. Overview of theoretical frameworks

This section describes the components of the proposed FNET model, the related theory in detail and benchmark models used to compare against the objective models.

2.1. The proposed Deeply Fused Networks (FNET) model

In this study, we have developed the Deeply Fused Networks (FNET) model, which is constructed from a series of base networks whose output representations are fused together [25]. In contrast to the shallow fusion model, as per Fig. 2(a), the deep fusion approach applies the feature fusion to both the final representation and the intermediate feature embedding, as schematized in Fig. 2(b). Typically, there are N blocks ($N \geq 1$) in the proposed FNET model and each block has L ($L \geq 1$) base networks. Each block fuses the feature representation from several base networks together, and the merged feature embedding is then handled as the input to the succeeding block. Furthermore, the L base networks are composed of various convolutional kernel scales while the number of convolutional layers can frequently vary [29]. In the study of [25], the FNET approach has achieved superior performance over Residual Network [28] and Highway Network [27]. In principle, the deep fusion approach can offer numerous advantages over the traditional shallow fusion method [30], which are as follows:

- The information flow during deep fusion can be enhanced regardless of whether it comes from the input to the intermediate layers or from the intermediate to the output layers.
- It is relatively easier to train an FNET model consisting of a very deep base network (*i.e.*, the first model, CNN) and a shallow model (e.g., ANN, SVR, etc.) or other deep models (second model, LSTM, BILSTM, GRU, etc.) compared to a deep base network alone. Despite having several base networks, the FNET model does not add more parameters or computational complexity but it facilitates training process quite seamlessly.
- The two models (deep-shallow or deep-deep) are likely to provide benefit from each other's own merits and are therefore trained simultaneously in order to learn more representative feature embedding.
- Because of its unique structure, the FNET model can extract multi-scale feature representations.

As a result of the advantages mentioned above, we expect the point-based and the interval prediction of daily electricity demand data to be performed quire satisfactorily by the proposed FNET model.

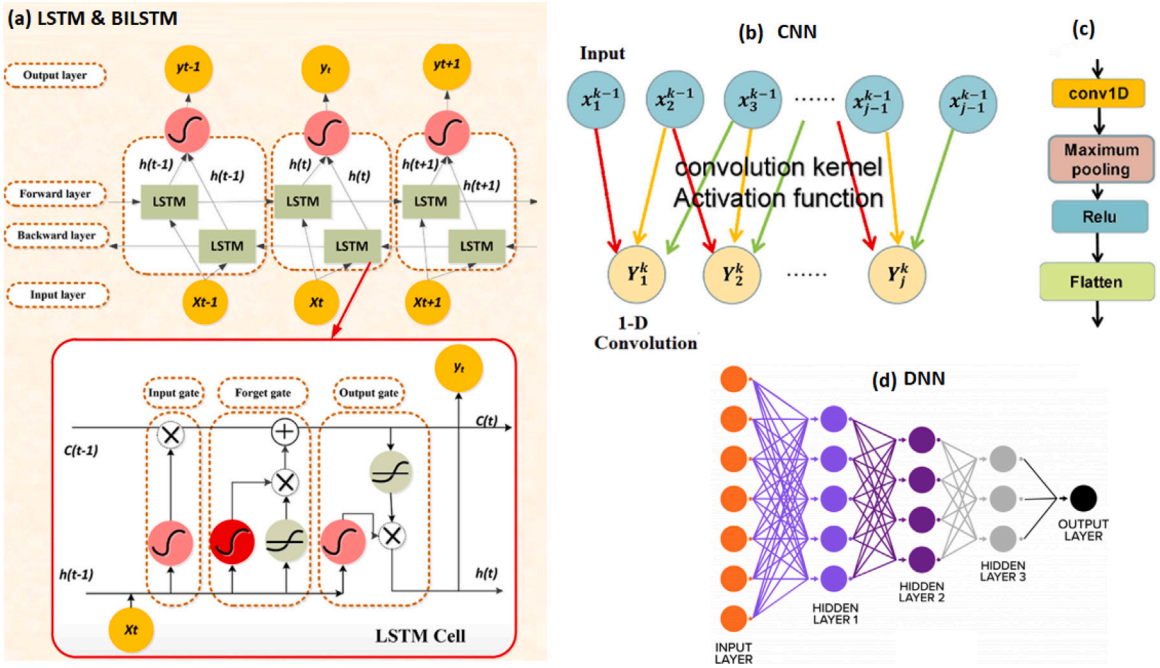


Fig. 1. A schematic view of the LSTM, BILSTM, 1D-CNN and DNN models employed to construct the proposed deep hybrid Fused Network (FNET) model for point-based and interval prediction of daily electricity demand.

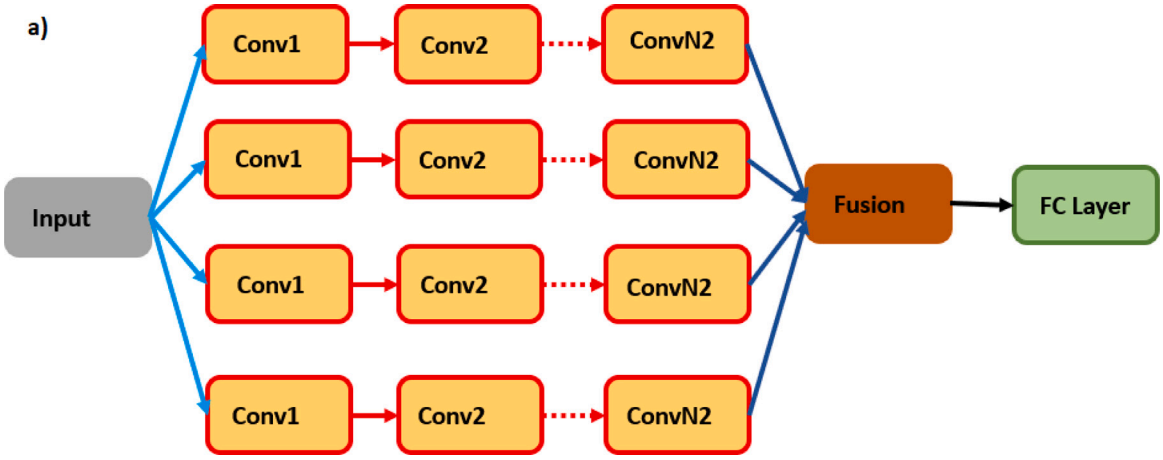


Fig. 2(a). The presentation of shallow fusion principle, where different scores from various CNN categories fused once prior the release to the regression analysis. Note that Conv1 is a CNN model, and FC is a fully Connected Layer.

2.2. Benchmark (deep and shallow learning) models

2.2.1. Long-short term memory network

The LSTM Network model, as a benchmark for the proposed FNET model, is a variation of Recurrent Network (RNN) [31,32]. In comparison to RNN, LSTM can manage long-term dependencies as well as gradient vanishing difficulties. The cell state and the gate structure are the basic concepts of LSTM, in which cell states are used to communicate information and solve the issues of short-term memory. The LSTM has three gate structures: input, forgetting, and output gates, each with its function [33]. The forget gate determines whether information should be discarded or maintained [34]. The input gate is utilized to update the state of the cell. The output gate is used to calculate the value of the next hidden state, which contains the previously entered data. Fig. 1 (a) depicts the structure and the equations below (Eqs. (1)–(6)) show the conventional equations for LSTM [35].

$$f_t = \sigma\left(w_f * \left[h_t - 1, x_t\right] + b_f\right) \quad (1)$$

$$i_t = \sigma\left(\mathrm{w_i}\left[\mathrm{h_t} - 1, \mathrm{x_t}\right] + \mathrm{b_i}\right) \quad (2)$$

$$c_t = \tanh\left(w_c \cdot \left[h_t - 1, x_t\right] + b_c\right) \quad (3)$$

$$c_t = f_t * C_{t-1} + i_t * C_t \quad (4)$$

$$O_t = \sigma\left(W_0 \cdot \left[h_{t-1,x_l}\right] + b_0\right) \quad (5)$$

$$h_t = O_t * \tanh\left(c_t\right) \quad (6)$$

where f_t is the forget gate, σ is the sigmoid function, W_f is the weight, h_{t-1} is the output of previous block, X_t is the input vector, and b_f shows the bias. The symbol $*$ signifies elementwise multiplication, C_t is the Cell state, h_t is the hidden state, O_t is the output gate and $\tanh(.)$ denotes the hyperbolic tangent activation function.

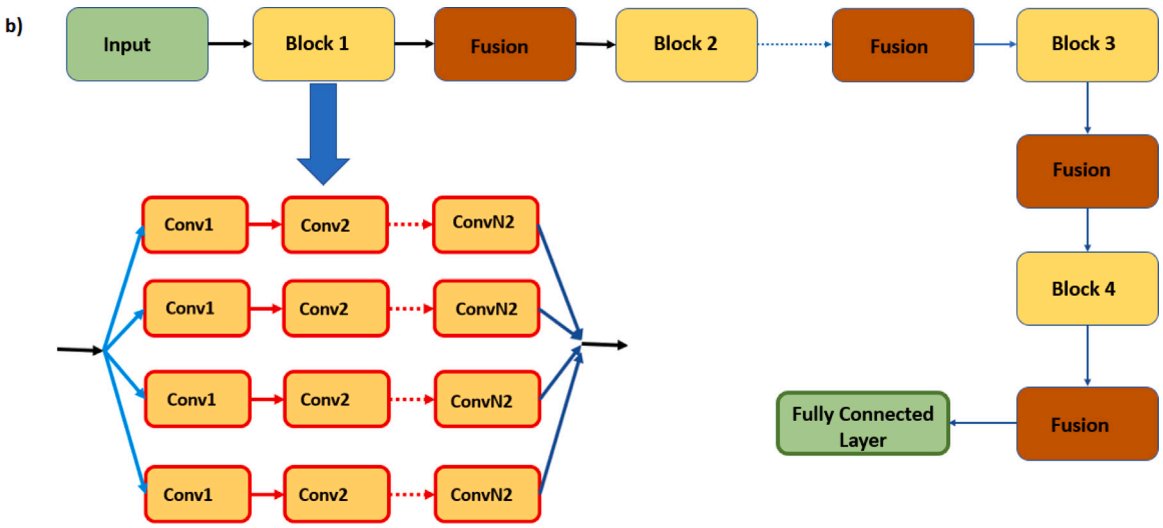


Fig. 2(b). Deep fusion principle where features extracted from different CNN branches are fused in a block before entering the intermediate results into the next block for further feature learning.

2.2.2. Bi-directional LSTM

By examining input vectors in one direction only, the typical LSTM model may lose crucial feature information in the training process, preventing the sequence information from being completely evaluated [36]. To overcome this issue, BILSTM is built with a bidirectional structure to gather both the forward and backward directions of time-series data representations, as illustrated in Fig. 1(a). As a result, the BILSTM produces a final output vector Y_t expressed by:

$$Y_t = \sigma\left(W_{fhy}h_{f(t)} + W_{bhy}h_{b(t)} + b_y\right) \tag{7}$$

This structure enables the internal state to store information in h_{ft} from the past time-series values in the forward direction and in h_{bt} from the future sequence values in the backward direction. The W_{fhy} and W_{bhy} symbolize forward and backward weighting scores from the internal unit to the output, respectively. σ is set to sigmoid or linear functions as the output layer activation function and b_y signifies the bias vector of the output layer.

2.2.3. Convolution neural network

Convolutional neural networks (CNN) are a type of feedforward neural network that uses convolutional computing. To extract feature information, CNN models employ convolution layers and pooling layers [37]. The core of a CNN model is the convolution layer that reduces the network's complexity and parameter numbers by connecting a neuron to only a subset of its neighbours [31]. Further, the pooling layer minimizes the number of parameters by lowering the dimensionality of the features. Adding a pooling layer not only speeds up the computation but also prevents over-fitting [8,38]. This study has employed a one-dimensional convolutional structure for sequential data (Fig. 1(b)), where the convolutional kernel is set to 3, depicted by red, orange, and green colour, and the CNN network structure is shown in Fig. 1(c).

The convolution outputs for each layer after convolution computation are treated non-linearly using the activation function known as Scaled Exponential Linear Unit (SeLU). This study has utilized SeLU because it avoids the self-dying problem associated with Rectified Linear Unit Activation function (ReLU) [39] and has a self-normalizing property [40] that makes the neuron activation automatically converge toward an average of 0 and a variance of 1. Due to this property of the SeLU activation function, many layers of CNN can be trained more robustly without gradient vanishing issues. The output $Y^{(r)}$ of the rth convolution layer can be defined as:

$$Y^{(r)} = f\left(\sum_{m=1}^{M} X_m^{(r-1)} \otimes W^{(r)} + B^{(r)}\right)(r = 1, 2, \ldots, l) \tag{8}$$

where the convolution operation is a dot product $\otimes$ between M feature maps $X^{(r-1)}$ and a set of filters $W^{(r)}$, which is the convolution kernels of the rth convolution layer. It is noted that when $r = 1$, $X^{(r-1)}$ is the reorganization of the input layer data; otherwise, $X^{(r-1)}$ is the output of the r^{1th} pooling layer. $B^{(r)}$ denotes the bias term. The activation function $f(x)$ is the SeLU function defined by following Eq. (9).

$$\mathrm{SeLU}(x) = \lambda\begin{cases} x & \text{if } x > 0, \\ \alpha e^x - \alpha & \text{if } x \le 0 \end{cases} \tag{9}$$

where x signifies the input to the activation function, $\lambda \approx 1.0507$ and $\alpha \approx 1.6733$ [41]. The pooling layer uses the output of the convolution layer as the input (Fig. 1(b)). The output $X^{(r)}$ of the rth pooling layer can be depicted by Eq. (10).

$$X^{(r)} = W\prime^{(r)} \oplus Y^{(r)} + B^{(r)}(r = 1, 2, \ldots, l) \tag{10}$$

where $\oplus$ denotes the pooling operation $\otimes$ is the dot product of feature maps $Y^{(r)}$ and pooling window $W^{(r)}$ with the bias $B^{(r)}$. The neurons in all the feature maps of the lth pooling layer are rasterized and displayed to one feature map by the full connection layer. After the transformation, the output $X^{(l+1)}$ of this layer is used to generate the final output of the CNN model expressed as:

$$Y^{(l+1)} = f\left(X^{(l+1)} * W^{(l+1)} + B^{(l+1)}\right) \tag{11}$$

where $W^{(l+1)}$ and $B^{(l+1)}$ represent the weight and bias of the output layer, respectively.

2.2.4. Multilayer Perceptron and Deep Neural Network

Multi-Layer Perceptrons (MLP) are a versatile and general-purpose type of Artificial Neural Network (ANN) [42], composed of an input layer, one or more hidden layers, and an output layer. An MLP network is comprised of simple neurons called perceptrons [43]. A perceptron integrates linear relationships based on input weights and even non-linear transfer functions (e.g., sigmoid or hyperbolic tangent) to form one output from multiple inputs. Whereas Deep Neural Network (DNN) is considered an ANN with many hidden layers between the input and output layers and has a stronger modelling and prediction capability. In feed-forward DNN, information flows forward from the input layer via the hidden layers (multiple) to the output layer [44]. As seen in Fig. 1(d), DNN has several layers stacked together for processing and learning from data. The output Y of MLP and DNN models can be mathematically formulated by the transfer functions F of input variables X, weights W, and bias values B with n neurons in an input layer and m neurons in the hidden layer as:

$$Y = F\left(\sum_{j=1}^{m} W_{kj} \cdot F\left(\sum_{i=1}^{n} W_{ji}X_i + B_j\right) + B_k\right) \tag{12}$$

where k, j, and i refer to the output, hidden, and input layers, respectively.

2.2.5. Multivariate Adaptive Regression Splines

The Multivariate Adaptive Regression Splines (MARS) technique is a non-linear and non-parametric regression model. This model uses piece-wise linear splines to evaluate the relationships between the dependent and independent variables. MARS mimics the model using basic functions (BFs). BFs are described as pairs based on a knot to establish an inflection region [45]. Mathematical derivation of the model can be found in [46,47]. The elegance of the MARS model is that no assumptions are required to build a link between the input and output variables. Therefore, the MARS model has been applied in many studies, such as financial management, prediction, and time series analysis, including solar radiation and wind power [4,48].

2.2.6. Kernel Ridge Regression

Kernel Ridge Regression (KRR), a regularized least squares-based method, is an extension of the conventional Ridge Regression model, which is extensively used for regression and classification of highly non-linear prediction tasks [49] . As a nonlinear procedure, KRR comprises a set of kernel tricks and RR to reduce over-fitting in nonlinear-large-multiple regression issues [50,51]. While the KRR model performance for regression problems is similar to the Support Vector Regression (SVR) [52] model, the key difference between the two models can be found in the loss function. More specifically, KRR implements a square error loss function, while SVR uses an epsilon-insensitive loss function. Furthermore, KRR fits faster than SVR for a small number of datasets. Complete mathematical derivation of the KRR model can be found in [53].

2.2.7. Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric modelling tool that does not dictate the type of relationship between the input and output [54]. Numerous applications of GPR have demonstrated its ability to make accurate probabilistic predictions in complex nonlinear situations [55]. Due to the complex relationships between G and other weather variables, a GPR is chosen as the benchmark model to predict the daily electricity demand. A detailed model description and the mathematical formulation are provided in [56,57].

2.3. Generating Bootstrap-based prediction intervals

While the aforementioned DL and ML models produce reliable G predictions, these model's predictions are subject to uncertainty. To address this issue, the Bootstrap Residual (BSR) method proposed by [58,59] is employed to quantify the uncertainties by generating Prediction Interval (PI) at the 95% confidence level. To implement the bootstrap prediction of G, we first generate a bootstrap sample of the residuals.

Using the residuals $\{\hat{\epsilon}_t : t = 1, 2, \ldots, n\}$, we define the empirical distribution $\hat{F}_\epsilon(\cdot)$ [60] by:

$$\hat{F}_\epsilon(x) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{I}_{(-\infty, \hat{\epsilon}_t)}(x) \tag{13}$$

From the empirical distribution $\hat{F}_\epsilon$, we draw an independent and identically distributed (i.i.d) sequence $\{\hat{\epsilon}_t^* : t = 1, 2, \ldots\}$, which is used as a bootstrap sample for constructing a bootstrap prediction interval. One-step ahead bootstrap prediction is carried out by:

$$\hat{X}_{n+1}^* \equiv \hat{X}_n^*(1) = \hat{\theta}^\top \mathbb{Y}_n + \hat{\epsilon}_1^* \tag{14}$$

The $100(1-\alpha)\%$ bootstrap prediction interval is computed as:

$$\left[\hat{X}_n^*(1)_{\alpha/2}, \hat{X}_n^*(1)_{1-\alpha/2}\right] = \left[\hat{\theta}^\top \mathbb{Y}_n + q_{\alpha/2}^*, \hat{\theta}^\top \mathbb{Y}_n + q_{1-\alpha/2}^*\right] \tag{15}$$

at $\alpha/2$ and $(1-\alpha/2)$ bootstrap quantiles, denoted as $q_{\alpha/2}^*$ and $q_{1-\alpha/2}^*$, respectively, of the bootstrap sample $\{\hat{\epsilon}_t^* : t = 1, 2, \ldots, N\}$, where N indicates the number of bootstrap replications.

2.4. Model interpretation

To foster transparency and build trust in the model's decision-making process, the SHAP (SHapley Additive exPlanations) method was employed. Rooted in game theory, SHAP offers a robust framework for understanding feature contributions to model output [61]. By decomposing the model's predictions into contributions from individual features, SHAP empowers stakeholders to comprehend the underlying logic and rationale behind the model's decisions. This interpretability is crucial for sectors such as electricity network operation, demand response aggregation, and electricity trading, where understanding user behaviour is paramount for effective strategy development and risk management.

3. Materials and method

3.1. Research methodology

A systematic methodology was implemented in this study to validate the efficacy of the proposed FNET for the prediction of daily electricity demand (G). Fig. 3 depicts the overall framework of the methodology. It consists mostly of eight major phases, as outlined below:

Phase 1: The data preparation step: the electricity demand (G, MW) data from 01/07/2011 to 30/06/2021 of the four sub-stations in South-east Queensland, Australia are collected from Energex website ().

Phase 2: Feature set scenario development: The collected G data were preprocessed to create the input features for prediction models. The partial Auto-correlation Function (PACF) and Mutual Information Test (MIF) are done to identify the suitable lags.

Phase 3: Integration of climate variables: The climate variables from the Scientific Information for Land Owners (SILO) database are extracted and integrated with lagged G data.

Phase 4: Final pre-processing of data: Further pre-processing is done by normalizing and splitting the data into training, validation and testing sets.

Phase 5: Predictive Model Development: The proposed model (*i.e.*, FNET) and benchmark Models (LSTMCNN, DNN, BILSTM, MLP, KRR, GPR and MARS) are developed and trained on training dataset to predict the daily G. Additionally, the proposed models have been optimized for tuning their hyperparameters (validation data and utilizing the HyperOpt Algorithm).

Phase 6: Model Evaluation: The prediction accuracy related to the eight predictive model configurations has been evaluated using several deterministic metrics.

Phase 7: Residual Bootstrap: The final residual was computed from the predicted and actual values of G, and bootstrapping was done to generate the Prediction Intervals PI at 95% confidence level.

Phase 8: Uncertainty Quantification: the generated PI are analysed using the probabilistic metrics to quantify the uncertainty associated with FNET and benchmark models.

The following sections further describe the phases of the Model development framework.

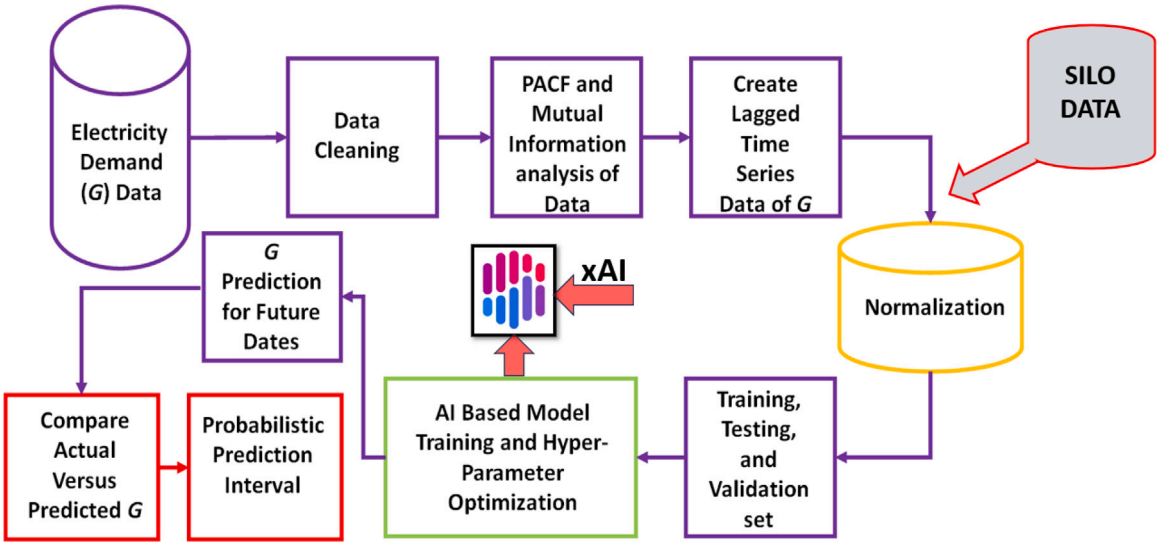


Fig. 3. Schematic diagram of model development.

Table 1
Descriptive statistics of daily electricity demand G (MW) at four substations of Southeast Queensland.

Statistical parameters	Annerley	Heathwood	Laidley	Zillmere
Median (MW)	345.39	659.85	193.58	694.43
Mean (MW)	371.62	649.86	195.94	709.37
Standard Deviation (MW)	90.55	113.49	41.22	106.11
Variance	8198.57	12 879.10	1699.27	11 259.13
Maximum (MW)	888.50	1005.26	535.64	1136.41
Minimum (MW)	−104.07	99.52	0.00	0.00
Range	888.50	1005.26	535.64	1136.41
Interquartile Range	70.19	127.22	36.41	143.99
Skewness	2.15	−0.30	−0.73	0.35
Kurtosis	8.57	3.38	11.05	4.87

3.1.1. Data preparation and feature scenario development

Since data-driven models (e.g., DL) heavily rely on past prognosis, the electricity data for four substations (Fig. 4; (a) Annerley, (b) Heathwood, (c) Laidley, and (d) Zillmere) in Southeast Queensland, Australia are collected from Energex website (). The dataset includes 280,560 measurements at a 30-minute sampling rate from 01/07/2011 to 30/06/2021 (120 months or 3653 days). The dataset has been downsampled from 30-min interval to daily interval using Eq. (16), where G_D is a function that employs a set of electricity demand data as input to down samples to a specific period with a downsampling rate n (*i.e.*, for the daily transformation of 30-min data, $n = 48$).

$$G_{Di} = \sum_{j=i*n}^{(i*n)+n} G_{Di} \tag{16}$$

It is important to note that for the interpretation of the proposed model performance in terms of the electricity demand that is typically measured in MWh, the respective timescale should be used and appropriate conversions to the time-based usage should be applied. Table 2 exhibits some descriptive statistics of daily G for the selected substations.

Fig. 5(a) provides further information on the distribution of annual G for the four substations. Box plots provide a visual representation of summary statistics (minimum, maximum, median, first quartile, and third quartile) for sample data and outliers are indicated with a circle ('o') outside the whisker. Fig. 5(a) shows that there are no significant differences in the G distribution between years for the Laidley substation, whereas for the Annerley substation, the year 2011 has the highest range of G variation compared to the period from 2012 to 2021. for the other two substations (Heathwood and Zillmere), there is an overlap between each G distribution box, indicating the parameters studied are not significantly different at 5% significance level.

Fig. 5(b) shows a box plot of the G for the entire substation monthly. According to this box plot, the highest and lowest changes of G occur in Autumn (March, April and May) and winter (June, July and August), respectively. In the summer season (December, January and February), G distribution is longer than in other seasons. The medians (generally close to the average) of Autumn, Spring, and Winter are all at the same level.

As a starting point for the nonlinear modelling of G time series, the underlying dynamics of the data were firstly examined. In principle, the PACF can be used to determine the temporal correlation structure and the lag dimensions of electricity demand dataset used to construct the proposed FNET model [62]. Fig. 6(a) shows the PACF for G data from 2011 to 2021 for four substations. In all G time series, the highest PACF was acquired at lag 1, which means the antecedent 1-day G data were highly correlated to the current day's G values. This also shows the classical Auto Regressive rapid decay patterns [63]. Furthermore, the Mutual Information Function (MIF) is applied to study the chaotic dynamics of the daily G time-series. The choice of the delay time (τ) or the lag is critical to the capturing the processes of correlation integral calculation and neighbouring trajectory separation within a minimum embedding space [64]. The first minimum Fig. 6(b) in the MIF plot generates the state vector that comprises components with minimal mutual information [65]. The delay times chosen for Annerley, Heathwood, Laidley and Zillmere are 4, 6, 5 and 6 days, respectively. The most effective inputs for G prediction can be mathematically expressed as Eqs. (17)–(20) for Annerley, Heathwood, Laidley and Zillmere, respectively.

$$G_{Annerley} = f\left(G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}\right) \tag{17}$$

$$G_{Heathwood} = f\left(G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, G_{t-5}, G_{t-6}\right) \tag{18}$$

$$G_{Laidley} = f\left(G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, G_{t-5}\right) \tag{19}$$

$$G_{Zillmere} = f\left(G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, G_{t-5}, G_{t-6}\right) \tag{20}$$

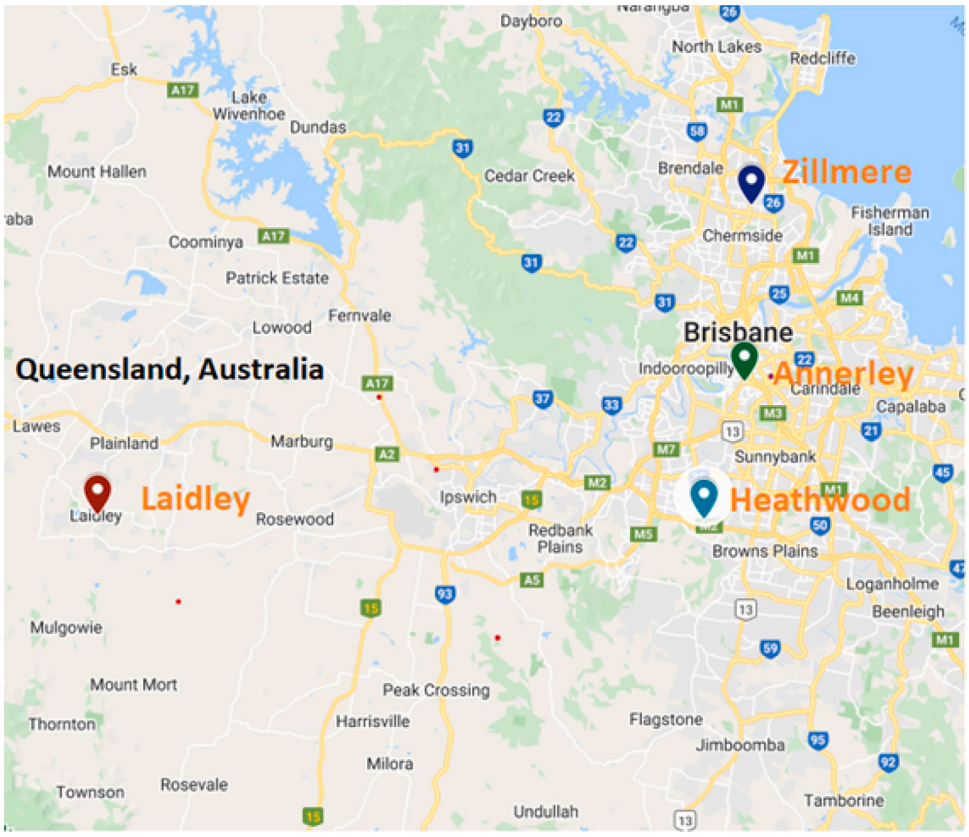


Fig. 4. Map of the location of the substations in Queensland, Australia where the deep hybrid Fused Network (FNET) model was implemented.

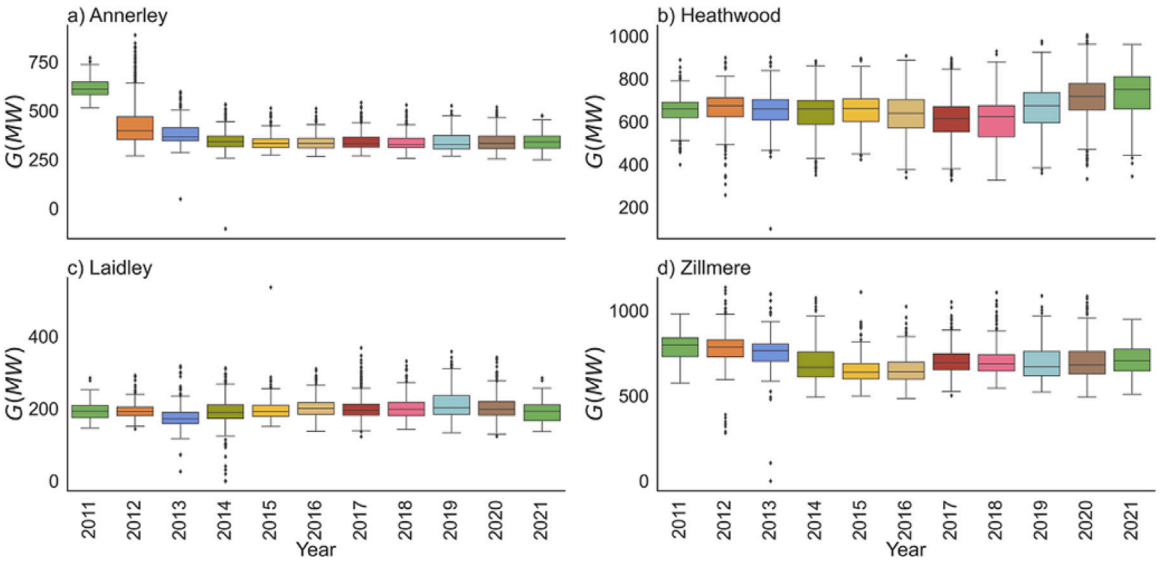


Fig. 5(a). Box plots of annual statistics for G at four substations. Note: The box represents the interquartile range, and whiskers extend to the 5th and 95th percentile.

3.1.2. Local climate variables and pre-processing of data

Apart from using the antecedent G series to build the proposed FNET model, this study has used ten different local climate variables from Scientific Information for Landowners (SILO) repository(). A SILO database system provides researchers with 'ready-to-use' climate data for their predictive models. A comprehensive description of the SILO database and spatial interpolation of Australian climate data can be found in [66]. The Queensland Department of Environment and Science hosts and organizes the SILO datasets.

Table 1 lists the variables from SILO database and Fig. 7 shows the heatmap of SILO predictors and the target variable (G) for all substations. Note that the acronyms used in Fig. 7 are defined in Table 1. According to Fig. 7, the Mean Sea Level Pressure ($MSLP$) is the most highly correlated predictive variable with G for all substations. For Laidley and Zillmere, the Maximum Temperature ($Tmax$) and Vapour Pressure Deficit (VPd) also have a high correlation with G.

As a further step required in data preprocessing, the min–max data normalization method (Eq. (21)) was applied to the lagged D data as well as the SILO variables since DL network performance is are

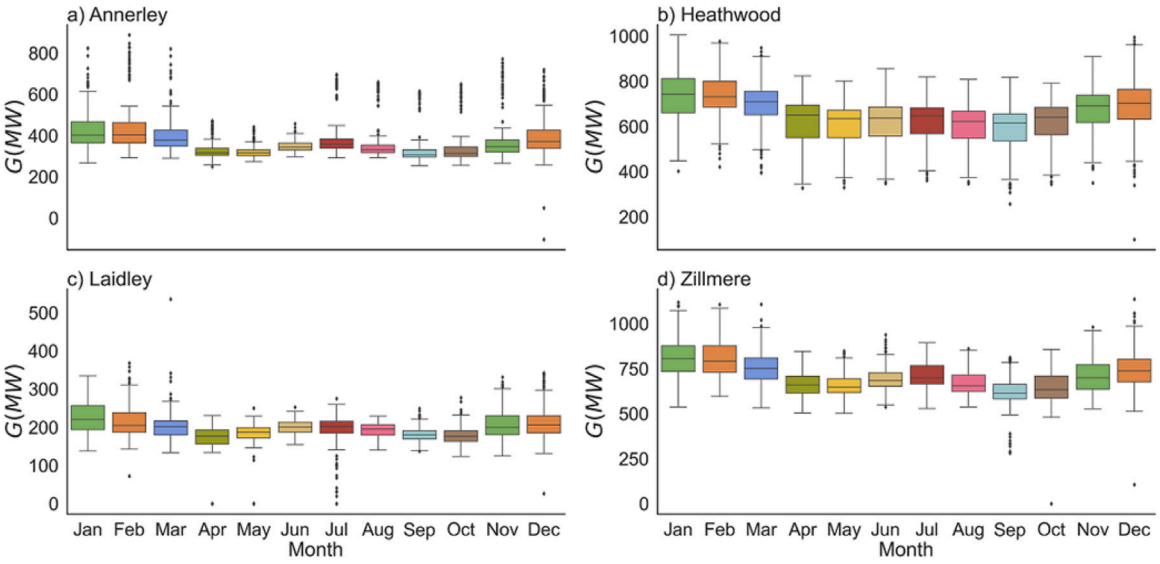


Fig. 5(b). Box plots of monthly statistics for G at four substations. Note: The box represents the interquartile range, and whiskers extend to the 5th and 95th percentile.

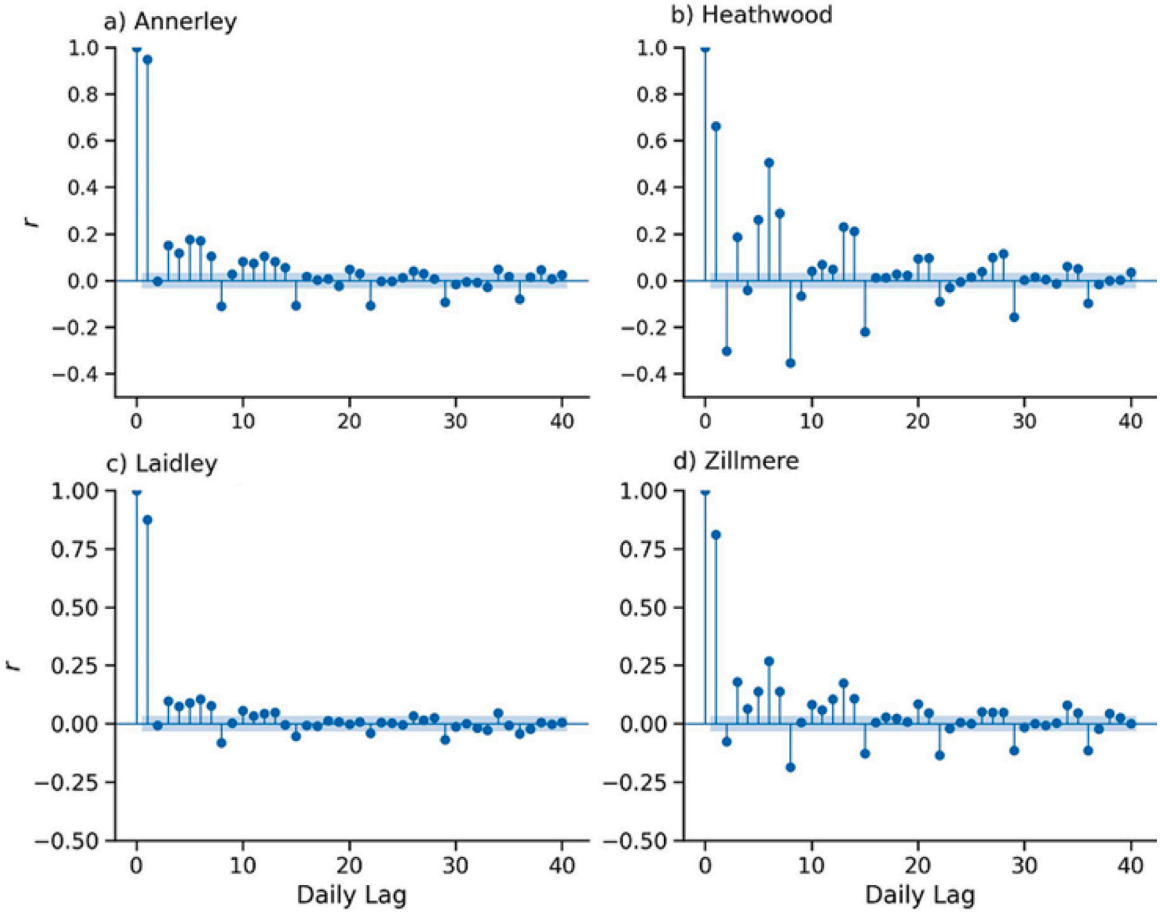


Fig. 6(a). PACF plots used for the selection of model degrees of FNET and Benchmark models of (a) Annerley, (b) Heathwood, (c) Laidley, and (d) Zillmere for daily G data from 01/07/2011 to 30/06/2021.

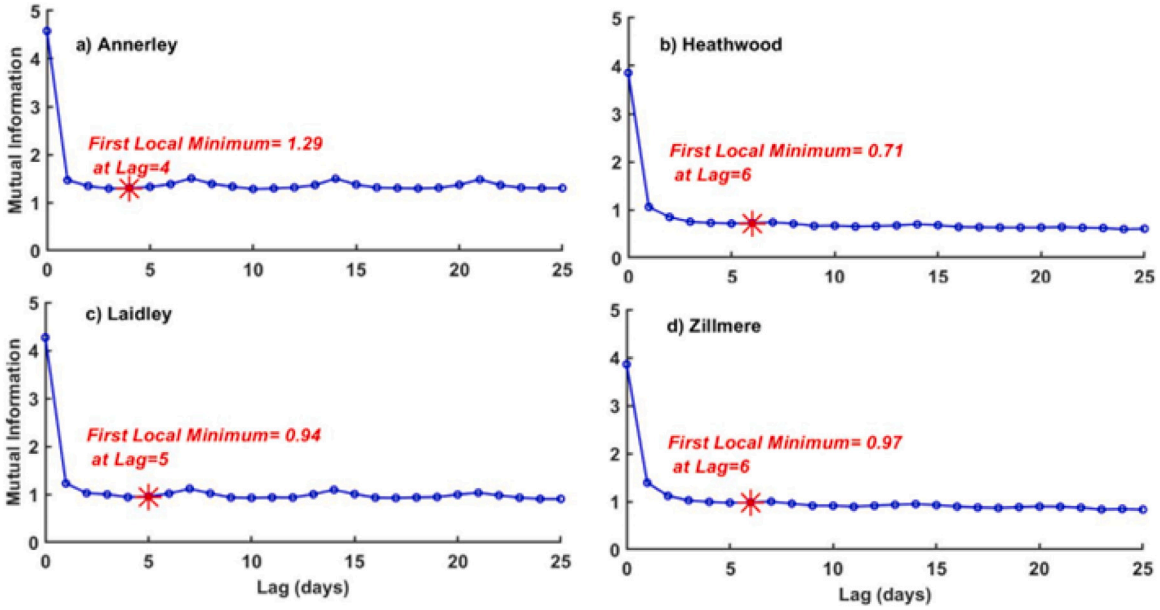


Fig. 6(b). Mutual information (MI) functions and their relative change for G time series from 01/07/2011 to 30/06/2021.

Table 2
Description of the pool of local climate predictor variables from Scientific Information for Landowners (SILO) database used for the point prediction and the interval prediction of $G(MW)$ at four substations in southeast Queensland, Australia.

Local climate predictor variables from SILO	Acronym
Maximum temperature (°C)	Tmax
Minimum temperature (°C)	Tmin
Vapour pressure (hPa)	VP
Vapour pressure deficit (hPa)	VPd
Evaporation - synthetic estimate (mm)	Esyn
Solar radiation - total incoming downward shortwave radiation on a horizontal surface (MJ/m^2)	GSR
Relative humidity at the time of maximum temperature (%)	Rhmax
Relative humidity at the time of minimum temperature (%)	Rhmin
Evapotranspiration - Morton's areal actual evapotranspiration (mm)	Etm
Mean sea level pressure (hPa)	MSLP

sensitive to the diversity of input data which requires normalization. After normalizing the data, the input and output matrices were created as per Eq. (22) and Eq. (23)(e.g. for Annerley)

$$X_{norm} = \frac{(X - X_{\min})}{(X_{\max} - X_{\min})} \tag{21}$$

where X = input/target, X_{min} = minimum point, X_{max} = maximum point and X_{norm} = anticipated normalized value.

In order to develop the proposed FNET model (and comparative benchmark models) using historical G and local climate variables, a model input matrix with the predictor variables was created as follows:

$$\begin{aligned} Input = (&G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, T_{max_{(t-1)}}, \\ &T_{min_{(t-1)}}, V_{p(t-1)}, VP_{d(t-1)}, \\ &E_{syn(t-1)}, GSR_{(t-1)}, Rh_{max(t-1)}, \\ &Rh_{min(t-1)}, Etm_{(t-1)}, MSLP_{(t-1)}) \end{aligned} \tag{22}$$

$$Target = (G_t) \tag{23}$$

where G_t is the current electricity demand, G_{t-1}, G_{t-2}, G_{t-3}, G_{t-4}, $T_{max_{(t-1)}}$, $T_{min_{(t-1)}}$, $V_{p(t-1)}$, $VP_{d(t-1)}$, $E_{syn(t-1)}$, $GSR_{(t-1)}$, $Rh_{max(t-1)}$, $Rh_{min(t-1)}$, $Etm_{(t-1)}$ and $MSLP_{(t-1)}$ are the lagged values of electricity demand, Maximum Temperature, Minimum Temperature, Vapour pressure, Vapour Pressure Deficit, Solar Radiation, Relative Humidity at Maximum Temperature, Relative Humidity at Maximum Temperature, Morton's Areal Actual Evapo-transpiration and Mean Sea level Pressure, respectively.

Data are divided into training, validation and testing sets with 90% of data set from 01/07/2011 to 30/06/2020 (3288 data points) dedicated to model training and validation, while remaining 10% (365 data points) from 01/07/2020 to 30/06/2021 is used for testing purposes. The training set is used to train the model learn hidden features or patterns in the data, while the test set is used to test the model after the training is complete. During model training, we use a validation set to validate our model performance, separate from the training set. We use this validation process to tune the model's hyperparameters and configurations accordingly. It acts as a critique that indicates whether or not the training is progressing properly. In this study, 20% of data from the training set are used for validation, *i.e.*, 658 data points. Thus, the input matrix for Annerley substation is [2630 × 14], [658 × 14] and [365×14] for training, validation and testing, respectively. Similarly, for Heathwood and Zillmere, [2630×16], [658×16] and [365×16] for training, validation and testing, respectively. However, [2630 × 15], [658 × 15] and [365 × 15] of data are used for training, validation and testing, respectively, for the Laidley substation.

3.1.3. Predictive model development and evaluation

The proposed FNET as well as the benchmark models were designed on the Microsoft Windows 10 platform with an Intel® $core^{TM}$ i9 Generation 10 processor operating at 3.8 GHz with 32 GB memory. Models were designed in Python programming language [67] and MATLAB R2020b was used for statistical analysis. Tensor Flow [68], Keras [69], and Scikit-Learn [70] are some of the key and important libraries available in Python for DL. As mentioned earlier, this study uses the hybrid Deep Fusion Network (FNET) to predict the Daily G at four substations in Southeast Queensland, Australia.

Fig. 8 shows the FNET model that takes the fused nets with three fusions composed of a deep base network (1D-CNN) with CNN filters ranging from 32 to 128 and 4-layer BILSTM network with BILSTM unit ranging from 16 to 128. Since we are using FNET for regression purposes, a single BILSTM layer at the end of the network is used before the fully connected or dense layer. The $SeLU$ was used as the activation function for the 1D-CNN layer and the $ReLU$ is used for the Dense layer in the proposed FNET model.

The architecture of the proposed FNET (and benchmark models) are presented in Table 3. It is noteworthy that this paper has selected the

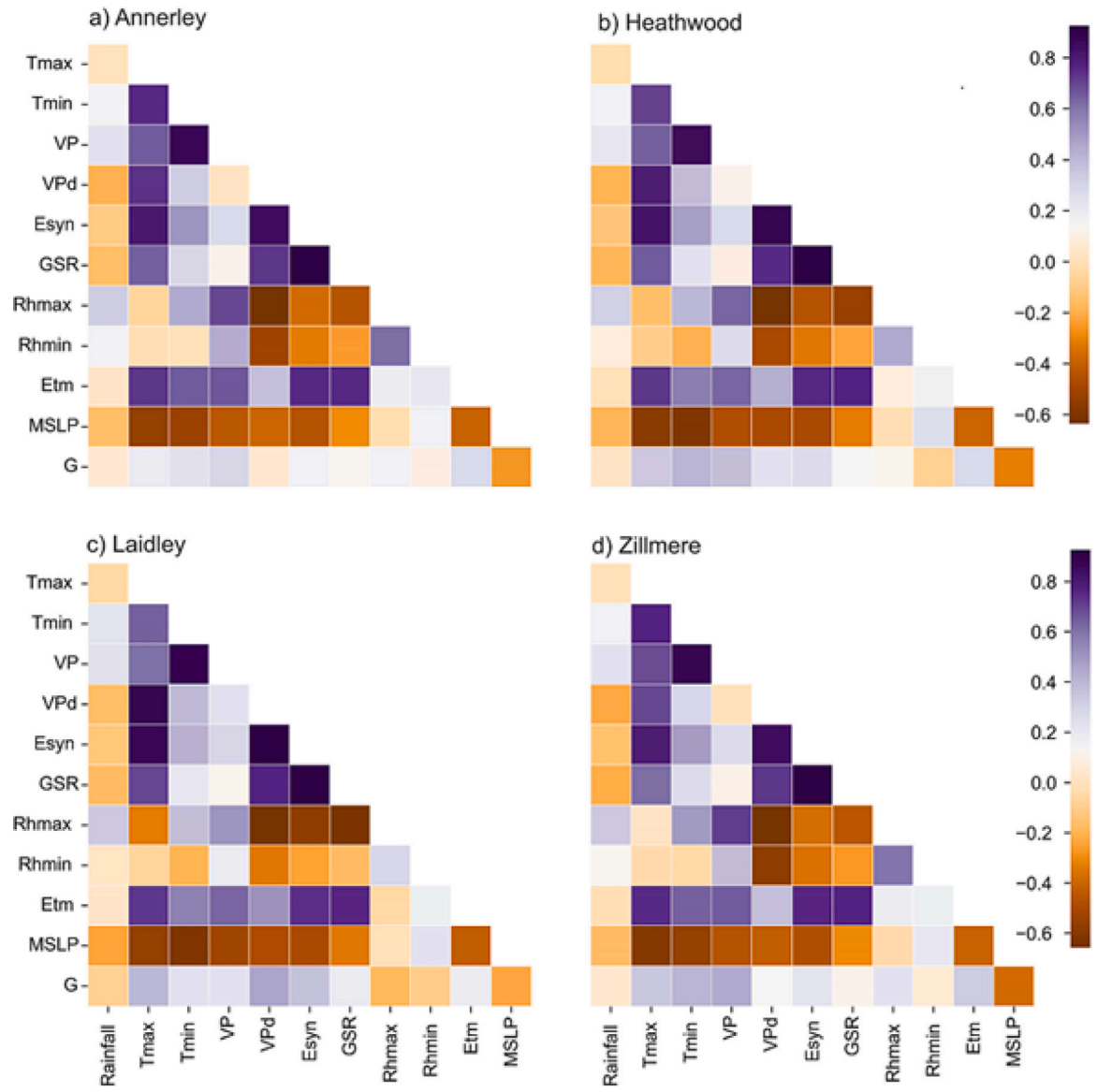


Fig. 7. Heatmap showing the correlation coefficient of SILO predictors and the target variable (G).

Adam algorithm as the model optimizer for the FNET, LSTMCNN, DNN and BILSTM models using Mean Square Error as a loss function. The choice of the Adam algorithm provides the advantage of maintaining momentum and gradient acceleration by considering both estimations of the first moment (mean gradient) and the second moment (variance of the gradient) [71]. This advantage allows enables the model to be trained faster and to predict the G data more accurately. Eq. (24) shows the back-propagation parameter adjustment using Adam and Eq. (25) shows the error function.

$$W_t = W_{t-1} - \alpha \frac{m_t}{\sqrt{v_t} + \varepsilon} \tag{24}$$

where w = weights of learning model, α = learning rate, and m_t and v_t = moving average.

$$L_{MSE} = \frac{1}{T} \sum_{t=1}^{T} (y_t - y_t^E)^2 \tag{25}$$

where y_t and y_{tE} = measured and predicted G at time t, respectively, and T = total prediction time period.

This study has also adopted the Python Hyperopt library [72] to deduce optimal hyperparameters, shown in Table 3 for BILSTM, LSTMCNN, DNN, MARS, MLP, KRR and GPR benchmark models. In this way, users can select their models or optimize their parameters simultaneously in Python programming environment. In fact, Hyperopt operates as a black box system in which the users can provide an evaluation function and parameter space to attain the best values based on the inputs [72]. When selecting an optimization algorithm through the Hyperopt, the distribution over the choice ('Adagrad', 'Adam', 'SGD', and 'RMSprop') is used. This study has used the following regularization parameters for the proposed FNET and all DL benchmark models (*i.e.*, BILSTM, LSTMCNN, and DNN).

- **Early Stopping** (es): This is used to overcome over-fitting, terminates the training once the performance stops improving on a validation data after an arbitrary number of epochs (patience). During training, the best model weights can be saved and updated with an es regularizer. After a certain number of iterations, the training is terminated, and the last best parameters are used [73]. One metric to monitor is MSE, which should be minimized. During training, the model will count the loss at each epoch. In subsequent epochs, if the MSE value does not change or the minimum value is already calculated, the training will be terminated. When training the model, the es patience was assumed to be 20.
- **ReduceLROnPlateau**: This stands for 'reduce', 'learning', 'rate', 'on', and 'plateau' - indicating the learning rate must be reduced

Table 3
Architecture of the Deep Hybrid Fused Network (FNET) model vs. LSTMCNN, DNN, BILSTM, MLP, KRR, GPR and MARS models developed for daily electricity demand G (MW) prediction at four sub-stations of Southeast Queensland. Note: - SeLU = Scaled Exponential Linear Unit; Adam = Adaptive Moment Estimation, ReLU = Rectified Linear Units; rbf=Radial Basis Function, logistic= Logistic Sigmoid Function, tanh= Hyperbolic Tangent Activation Function.

Predictive models	Model Hyperparameters	Hyperparameter Selection	Annerley	Heathwood	Laidley	Zillmere
Fused Net (FNET)	Filter1(CNN)	**32**				
	Filter 2 (CNN)	**32**				
	Filter 3 (CNN)	**128**				
	BILSTM cell 1 (BILSTM)	**16**				
	BILSTM cell 2 (BILSTM)	**64**				
	Epochs	**[1000]**				
	Activation function (CNN Layer)	**['SeLU']**				
	Activation function (Dense Layer)	**['ReLU']**				
	Solver	**['adam']**				
	Batch Size	**[5]**				
Long Short Term Memory Network Integrated with Convolutional Neural Network (LSTMCNN)	LSTM cell 1	**[50, 60,100,200]**	**100**	**60**	**100**	**100**
	LSTM cell 2	**[40,50,60,70,130]**	**40**	**70**	**70**	**60**
	CNN Filter 1	**[50, 60,100,200]**	**60**	**50**	**50**	**50**
	CNN Filter 2	**[40,50,60,70,130]**	**40**	**40**	**70**	**50**
	Activation function	**['relu']**				
	Epochs	**[1000]**				
	Batch Size	**[5,10,15,20,25,30]**	**10**	**5**	**10**	**5**
Bi-Directional LSTM (BILSTM)	BILSTM cell 1	**[50, 60,100,200]**	**60**	**50**	**50**	**60**
	BILSTM cell 2	**[40,50,60,70,130]**	**40**	**40**	**60**	**50**
	BILSTM cell 3	**[20,10,30,5]**	**30**	**20**	**20**	**10**
	Activation function	**['relu']**				
	Epochs	**[1000]**				
	Batch Size	**[5,10,15,20,25,30]**	**5**	**10**	**5**	**10**
Deep Neural Network (DNN)	Hiddenneuron 1	**[60,100,200,250,300,500]**	**200**	**250**	**100**	**250**
	Hiddenneuron 2	**[20,30,40,50,60,70]**	**70**	**50**	**70**	**30**
	Hiddenneuron 3	**[10,20,30,40,50]**	**10**	**20**	**10**	**30**
	Batch Size	**[5,10,15,20,25,30]**	**10**	**5**	**10**	**10**
	Solver	**['adam']**				
	Epochs	**[1000]**				
Multi-Layer Perceptron (MLP)	Hidden neuron	**[50,60,70,80,90,100]**	**90**	**60**	**70**	**90**
	Activation function	**['relu','logistic','tanh']**	**relu**	**tanh**	**logistic**	**relu**
	Learning rate	**[0.001,0.002,0.005,0.006]**	**0.002**	**0.001**	**0.005**	**0.001**
	Solver	**['adam']**				
Kernel Ridge Regression (KRR)	Kernel	**['rbf']**				
	alpha	**uniform (0,1)**	**0.0018**	**0.0021**	**0.0013**	**0.0012**
Gaussian Process Regression (GPR)	The kernel specifying the covariance function of the Gaussian Process.	**[DotProduct, WhiteKernel, DotProduct+WhiteKernel,**	**DotProduct + WhiteKernel**	**DotProduct+ WhiteKernel**	**DotProduct+WhiteKernel**	**DotProduct+WhiteKernel**
Multivariate Adaptive Regression Spline (MARS)	Maximum term generated by forward pass	**[10,20,30]**	**10**	**10**	**10**	**10**
	Maximum degree of terms generated by forward pass	**[5,10,15,20]**	**10**	**15**	**10**	**10**

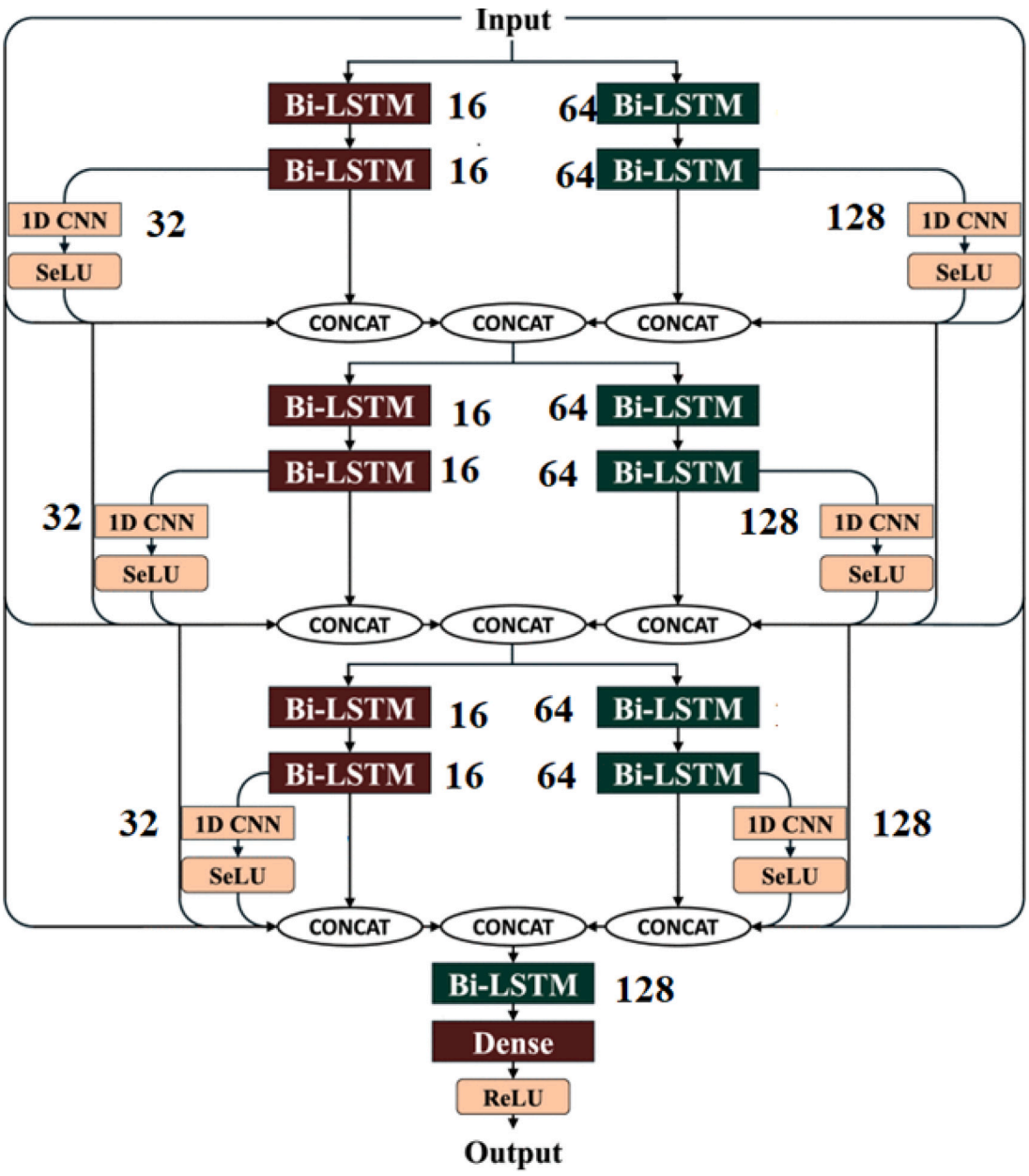


Fig. 8. The structure of the proposed FNET model using 1D-CNN and BILSTM layers used in the G prediction problem.

upon reaching a certain point. In this model, the regularizer is used to overcome under-fitting. Whenever the validation loss does not change, we dynamically update the learning rate [74]. After ten epochs without improvement during training, the learning rate will be reduced by 0.2; the lower bound is 0.00001.

- **Dropout**: Dropout rate (DOR) is an effective regularization tool for dealing with over-fitting. This prevents networks from becoming overly reliant on individual neurons. During the training phase, neurons are multiplied by a random variable following the Bernoulli distribution with a probability of p and the dropout rate is consistent with $(1-p)$. As part of this study, the DOR was set at 0.1 after every layer of the BILSTM, LSTMCNN, and DNN models.

To comprehensively evaluate the FNET model for G predictions, several deterministic metrics are used (see Tables 4(a)–4(c)).

- Class A metrics [Table 4(a)] are indicators of dispersion (or "error") of individually predicted G (0 for a perfect model). According to the study of [78], the relative errors represent model capability as being excellent ($0 \leq RRMSE$ or $RMAE \leq 10\%$), good ($10\% \leq RRMSE$ or $RMAE \leq 20\%$), fair ($20\% \leq RRMSE$ or $RMAE \leq 30\%$) and poor ($RRMSE$ or$RMAE \geq 30\%$).
- Class B metrics [Table 4(b)] are the normalized metrics whose maximum value is 1 for a perfect model [79–81].
- Class C metric [Table 4(b)] uses the KSI and $OVER$ to indicate the similarities in the distribution of predicted G (a lower value would indicate a better distribution similarity with observed value). In fact, KSI measures the distance between Cumulative Distribution Function of two datasets whereas $OVER$ measures the distance between them in parts where a critical value distance exceeds. The study also uses the Combined Performance Index (CPI), as per [82], to integrate $RMSE$, KSI and $OVER$ into a unified model performance indicator.

Table 4(a)
Class A - Deterministic performance measure.
Note that G^m and G^p = observed and predicted G, $\langle G^m \rangle$ and $\langle G^p \rangle$ = observed and predicted mean G, p = model prediction, x = observation, pr for perfect prediction (persistence), and r for the reference prediction, VAR = variance, SD = standard deviation, n = number of predictions [75].

Deterministic Performance Measure (Class A)	Definition	
Correlation Coefficient	$r = \frac{\sum_{i=1}^{n}(G^m-\langle G^m\rangle)(G^p-\langle G^p\rangle)}{\sqrt{\sum_{i=1}^{n}(G^m-\langle G^m\rangle)^2}\sqrt{\sum_{i=1}^{n}(G^p-\langle G^p\rangle)^2}}$	(26)
Root Mean Square Error (MW)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(G^p - G^m)^2}$	(27)
Mean Absolute Error (MW)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert G^p - G^m\rvert$	(28)
Relative Root Mean Square	$RRMSPE = \frac{RMSE}{G^m} \times 100\%$	(29)
Relative Mean Absolute Percentage Error (%)	$RMAPE = \frac{MAE}{G^m} \times 100\%$	(30)
Uncertainty at 95%	$U_{95} = 1.96(SD^2 - RMSE^2)^{0.5}$	(31)
t-statistic	$TS = \sqrt{\frac{(n-1) \times MBE^2}{RMSE^2 - MBE^2}}$	(32)
Mean Bias Error (MW)	$MBE = (100/\langle G^m \rangle)\sum_{i=1}^{i=N}\left(G^p{}_i - G^m{}_i\right)$	(33)
Standard deviation of the Relative Error	$STDRE = \left(\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{G^p-G^m}{G^m}\right)^2\right)^{1/2}$	(34)
Explained Variance Score	$E_{var} = 1 - \frac{\mathrm{Var}(G^m-G^p)}{\mathrm{Var}(G^m)}$	(35)
Absolute Percentage Bias (%)	$APB = \frac{\sum_{i=1}^{n}(G^m-G^p)*100}{\sum_{i=1}^{n}G^m}$	(36)
Skill Score	$SS = 1 - \frac{RMSE(p,x)}{RMSE(pr,x)}$	(37)

Table 4(b)
Class B - Deterministic performance measure.
Note that G^m and G^p = observed and predicted G, $\langle G^m \rangle$ and $\langle G^p \rangle$ = observed and predicted mean G, n = number of predictions, CV = Coefficient of Variation.

Deterministic Performance Measure (Class B)	Definition	
Willmot's Index	$E_{WI} = 1 - \frac{\sum_{i=n}^{n}(G^m-G^p)^2}{\sum_{i=n}^{n}(\lvert G^p-\langle G^m\rangle\rvert+\lvert G^m-\langle G^m\rangle\rvert)^2}$	(38)
Nash–Sutcliffe Equation	$E_{NS} = 1 - \frac{\sum_{i=1}^{n}(G^m-G^p)^2}{\sum_{i=1}^{n}(G^m-\langle G^m\rangle)^2}$	(39)
Legates and McCabe's Index	$E_{LM} = 1 - \frac{\sum_{i=1}^{n}\lvert G^m-G^p\rvert}{\sum_{i=1}^{n}\lvert G^m-\langle G^m\rangle\rvert}$	(40)
Theil's Inequality Coefficient	$TIC = \frac{\sqrt{\frac{1}{n}\times\sum_{i=1}^{n}(G^p-G^m)^2}}{\left(\sqrt{\frac{1}{n}\times\sum_{i=1}^{n}(G^m)^2}+\sqrt{\frac{1}{n}\times\sum_{i=1}^{n}(G^p)^2}\right)}$	(41)
Kling-Gupta Efficiency	$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\langle G^p\rangle}{\langle G^m\rangle} - 1\right)^2 + \left(\frac{CV_p}{CV_m}\right)^2}$	(42)

Table 4(c)
Class C - Deterministic performance measure.
Note that D_n = absolute difference between calculated and measured CDF. X_{min} and X_{max} = minimum and maximum D_n, A_c = critical area, D_c = statistical characteristic of the reference distribution or critical value, N = number of points and $\Phi(N)$ is a pure function of N [76,77].

Deterministic Performance Measure (Class C)	Definition	
KSI	$KSI = \frac{100}{A_c}\int_{X_{min}}^{X_{max}} D_n dx$	(43)
Critical Limit Overestimation Index	$OVER = \frac{100}{A_c}\int_{X_0}^{X_1}\max(D_n - D_c, 0)dx$	(44)
	where $A_c = D_c(X_{max}X_{min})$	(45)
	where $D_c = \Phi(N)/N^{1/2}$	(46)
Combined Performance index	$CPI = \frac{KSI+OVER+2RMSE}{4}$	(47)

This study has also used Global Performance Indicator (GPI) as a metric to rank the models [83] as well as Promoting Percentages (λ), Directional Symmetry (DS), Diebold–Mariano (DM) [84] and Harvey–Leybourne–Newbold (HLN) test statistics to compare the performance

Table 4(d)
Probabilistic performance measure (Class D).
Note: N denotes the number of test samples, y_i is the ith observation, $L(G_i)$ and $U(G_i)$ represent lower bound and upper bound of the ith. G Prediction Interval respectively, G^m is the observed value of G, R is the Range. [86]. In $CRPS$ metrics, $I(\cdot)$ is the Heaviside function, it takes the value of 1 when $t > y$ and equals 0 otherwise.

Deterministic Performance Measure (Class D)	Definition	
Prediction Interval Coverage Probability	$\text{PICP} = \frac{1}{N}\sum_{i=1}^{N} c_i$	(54)
	$where c_i = \begin{cases} 1 & \text{if } y_i \in \left(U(G_i), L(G_i)\right) \\ 0 & \text{otherwise} \end{cases}$	(55)
Mean Prediction Interval Width	$MPIW = \frac{1}{N}\sum_{i=1}^{N}\left(U(G_i) - L(G_i)\right)$	(56)
F Value	$F = \frac{PICP \times 2 \times \frac{1}{MPIW}}{PICP + \frac{1}{MPIW}}$	(57)
Average Relative Interval Length	$ARIL = \frac{1}{N}\sum_{i=1}^{N}\frac{\left(U(G_i) - L(G_i)\right)}{G^m{}_i}$	(58)
Winkler Score	$WS = \begin{cases} \Delta_i & L\left(G_i\right) \le y_i \le U\left(G_i\right) \\ \Delta_i + 2\left(L\left(G_i\right) - y_i\right)/\alpha & y_i < L\left(G_i\right) \\ \Delta_i + 2\left(y_i - U\left(G_i\right)\right)/\alpha & y_t > U\left(G_i\right) \end{cases}$	(59)
	where $\Delta_i = U(G_i) - L(G_i)$	(60)
Normalized Mean Prediction Interval Width	$PINAW = \frac{1}{N \cdot R}\left(\sum_{i=1}^{N}\left(U\left(G_i\right) - L\left(G_i\right)\right)\right)$	(61)
Continuous Rank Probability Score (MW)	$CRPS = \frac{1}{N}\sum_{i=1}^{N} crps\left(F_i, y_i\right)$	(62)
	where $crps(F, y) = \int_{-\infty}^{\infty}\left(F(t) - I(t - y)\right)^2 dy$	(63)

of the proposed FNET model with the benchmark models. The GPI is computed using six performance metrics as follows:

$$\text{GPI}_i = \sum_{j=1}^{6} \alpha_j (g_j - y_{ij}) \tag{48}$$

where α_j = median of scaled values g_j of the statistical indicators j for model i in which $j = -1$ is for r and $j = 1$ for $RMSE$, MAE, $MAPE$, $RRMSE$ and MBE ($j = 1, 2, 3, 4, 5$). A large GPI implies good performance.

Finally, the Directional Symmetry (DS), Promoting Percentage of Absolute Percentage Bias (λ_{APB}), Kling-Gupta Efficiency (λ_{KGE}) and Root Mean Square Error (λ_{RMSE}) [85] are also employed to evaluate the efficacy of the proposed FNET model:

$$DS = \frac{1}{n}\sum_{t=2}^{n} d_t \times 100\% \tag{49}$$

where,

$$d_t = \begin{cases} 1 & \text{if } (G_t^m - G_{t-1}^m)(G_t^p - G_{t-1}^m) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{50}$$

$$\lambda_{APB} = \left|\frac{(APB_1 - APB_2)}{APB_1}\right| \tag{51}$$

$$\lambda_{\text{KGE}} = \left|\frac{(KGE_1 - KGE_2)}{KGE_1}\right| \tag{52}$$

$$\lambda_{RMSE} = \left|\frac{(RMSE_1 - RMSE_2)}{RMSE_1}\right| \tag{53}$$

where APB_1, $RMSE_1$ and KGE_1 = objective model performance metric and APB_2, $RMSE_2$ and KGE_2 = benchmark model performance metric.

3.1.4. Quantifying uncertainty in electricity demand with residual bootstrapping method

In this study, the proposed FNET model is developed in such a way that the predicted uncertainties in G can be explored in detail to ascertain the suitability of the method for decision-making in the electricity industry. To pursue this, the bootstrap technique is adopted to study the residuals from the point-based G predictions to establish the bootstrap-driven Prediction Intervals (PI). To generate the PIs, the 95% confidence level or $100(1-\alpha)\%$ where $\alpha = 0.05$ was selected using $N = 1000$ bootstrap samples to assess the distribution of uncertainties generated by the FNET model.

By re-sampling the predictive model outcomes and analysing the errors encountered in several rounds of model emulations used to generate the predicted electricity demand, bootstrapping allows for the capture of the predicted uncertainties in electricity demand. After a trial and error process, we considered 1000 bootstrap samples that showed no significant difference in predictions beyond this value. Based on the results of the residual bootstrapping method, we computed several probabilistic metrics, as per Table 4(d) in respect to the prediction intervals and the model's uncertainties.

As a crucial measure of the model variability, we have analysed the Prediction Interval Coverage probability ($PICP$) whereby the probability of true G value limited by the upper and the lower boundaries of the predicted G values can be studies. Typically, the $PICP$ values range from 0 to 1 where a magnitude exceeding the confidence level (*i.e.*, 0.95) is preferred for a robust predictive model. However, increasing the range of the prediction interval can elevate the PICP while providing lesser information about the model's stability in respect to low error predictions. As a result, we also employed the Mean Prediction Interval Width ($MPIW$) as a supplemental metric to indicate the capability of the model to enclose genuine values inside the prediction boundaries. In general, the model with a lower $MPIW$ is expected to have a reduced uncertainty across the models with similar $PICP$ value [87].

Our study has also employed another comprehensive index, $F-value$ that combined both $PICP$ and $MPIW$ to evaluate the performance of the model based on PIs. Notably, a larger value of F would indicate a better performance of the prediction interval. Additionally, the PI Normalized Average Width ($PINAW$), Average Relative Interval Width ($ARIL$), Winkler Score (WS) [88] and the Continuous Rank Probability Score ($CRPS$) were also used to explore various other uncertainty measures. It is important to note that in probabilistic prediction model evaluations, the $CRPS$ is one of the most commonly used error measure as similar to the MAE in deterministic predictions, this metric can also generalize the MAE as a probabilistic prediction evaluation measure of the proposed FNET model [89].

4. Results and discussion

4.1. Results based on deterministic model evaluation metrics

This section describes the findings by a comparative analysis of the proposed FNET and the seven benchmark models, *i.e.*, BILSTM, LSTMCNN, DNN, MLP, KRR, GPR, and MARS. To conduct an accurate evaluation and avoid subjective conclusions, a comprehensive comparison is performed using a range of deterministic metrics, as per Table 4(a), Table 4(b), and Table 4(c). In addition, graphic tools consisting of bar charts, scatter plots, box plots, Empirical Cumulative Distribution Functions (ECDF) as well as cumulative frequencies and Taylor diagrams are used to support the analysis. In general, the proposed FNET model shows a persistently superior performance in respect to the daily G prediction problem to supersede the benchmark models for all four substations indicated through metrics.

Despite the varied performance of the benchmark models depending on the metric of choice, our results showed that the BILSTM and LSTMCNN model were generally the second-best models, after the proposed FNET model. However, the deep learning (DNN and the GPR) model yielded moderately accurate performance whereas the shallow models (MARS, KRR, and MLP) produced the least accurate predicted G values. It is noteworthy that further explanations of these model performance variations (e.g., the causes of model bias and the underlying physics) is not the primary interest of this study which is focused on the performance analysis of only the proposed FNET model. However, clear separation of this statistical performance no doubt highlights the importance of using a wide range of model metrics to evaluate the performance of different models, albeit at the same tested site. Next, we present a detailed evaluation of the model performance discussed in the following paragraphs.

Table 5 compares the models in terms of r, $RMSE$, and MAE as the most popular first order metrics. Here, the correlation coefficient measures the closeness between the observed and the predicted points through a scatter plot to generate a least-square regression line as shown in Fig. 9. The Root Mean Squared Error is the standard deviation of the distribution of prediction errors or residuals, while the Mean Absolute Error is measured as the average of the absolute prediction errors. The $RMSE$ penalizes the large prediction errors compared to MAE prediction errors. The values of these statistical performance metrics indicate a better predictive performance of the proposed FNET compared to the alternative models. There is often a direct relationship among these scores, for example, if $r = 1$, then $RMSE = 0$ when all points lie on the regression line; hence, there are no errors. For instance, the proposed FNET model for the Annerley substation produced higher scores of r ($\approx$ 0.974) and lower scores of $RMSE$ ($\approx$ 15.136 MW) and MAE ($\approx$ 11.641 MW) followed by LSTMCNN and BILSTM models ($r \approx 0.967$ and 0.965; $RMSE \approx 16.233$ and 16.481 MW; $MAE \approx 12.482$ and 12.484 MW, respectively).

It is important to note that the other models such as the MLP, MARS and KRR, registered the worst performance with $r \approx 0.950$, 0.936 and 0.926; $RMSE \approx 19.579$, 21.984 and 23.480 MW; $MAE \approx 14.889$, 17.219 and 17.593 MW, respectively. Similar results were also found for Laidley and Zillmere sub-stations. For the Heathwood site, despite the proposed FNET model still being the best model, the performance order of the benchmark models appeared to vary depending on the choice of the metric. For instance, the BILSTM model was the second-top model based on the $RMSE$ ($\approx$ 69.167 MW) and the MAE ($\approx$ 51.999 MW) but the worst based on the r value ($\approx$ 0.930). By contrast, the GPR model produced a high r ($\approx$ 0.939), just after the proposed FNET model but also the highest $RMSE$ ($\approx$ 76.208 MW) and MAE ($\approx$ 61.831 MW). This variation of the model performance may be partly explained by the differences in the distributions of G dataset in Heathwood, for example, having a higher standard deviation with extreme values compared with the other sub-stations (Figs. 5(a) and 5(b)) where the extreme values may have more influences on the scores using square roots.

Table 5
The testing performance of the Deep Hybrid Fused Network (FNET) model *vs.* benchmark models as measured by Correlation Coefficient (r), Root Mean Square Error ($RMSE, MW$) and Mean Absolute Error (MAE, MW).

Sub-Station	Predictive Model	Model Performance Metrics		
		r	*RMSE*	*MAE*
Annerley	FNET	**0.974**	**15.136**	**11.641**
	BILSTM	0.965	16.481	12.484
	LSTMCNN	0.967	16.233	12.482
	DNN	0.964	16.854	12.900
	MLP	0.950	19.579	14.889
	KRR	0.926	23.480	17.593
	GPR	0.957	18.533	14.110
	MARS	0.936	21.984	17.219
Heathwood	FNET	**0.947**	**66.045**	**49.392**
	BILSTM	0.930	69.167	51.999
	LSTMCNN	0.932	71.417	55.177
	DNN	0.932	71.264	55.155
	MLP	0.931	73.323	57.078
	KRR	0.935	73.016	56.945
	GPR	0.939	76.208	61.831
	MARS	0.939	73.084	57.773
Laidley	FNET	**0.963**	**13.038**	**10.266**
	BILSTM	0.957	14.449	11.488
	LSTMCNN	0.961	14.370	11.425
	DNN	0.953	14.730	11.655
	MLP	0.947	15.325	12.004
	KRR	0.937	16.842	13.376
	GPR	0.949	15.084	11.726
	MARS	0.955	15.114	11.768
Zillmere	FNET	**0.953**	**39.808**	**31.393**
	BILSTM	0.948	41.254	32.506
	LSTMCNN	0.950	41.488	32.176
	DNN	0.944	42.657	33.573
	MLP	0.923	49.463	39.458
	KRR	0.933	46.721	36.709
	GPR	0.947	41.621	32.880
	MARS	0.947	41.679	32.922

Likewise, Table 6 represents the relative error for the testing data computed for the four substations, shown as the ratio of the $RMSE$ and the MAE to the mean value of the target variable. The scores of the Relative Root Mean square Error ($RRMSE$) and Relative Mean Absolute Error ($RMAE$) are therefore consistent with those in Table 5 that show the superior performance of the proposed FNET model.

In respect to the Skill Score (SS) presented in Fig. 10, we note that the proposed FNET model has achieved the highest SS value, followed by a relatively lower value for the BILSTM and LSTMCNN models for all tested substations while the values for the KRR and GPR model are the lowest particularly at the Annerley substation. Interestingly, all models based on the SS metric also registered comparative performance for the Heathwood substation. It should also be noted that the persistence model is used as the benchmark model for the computation of SS.

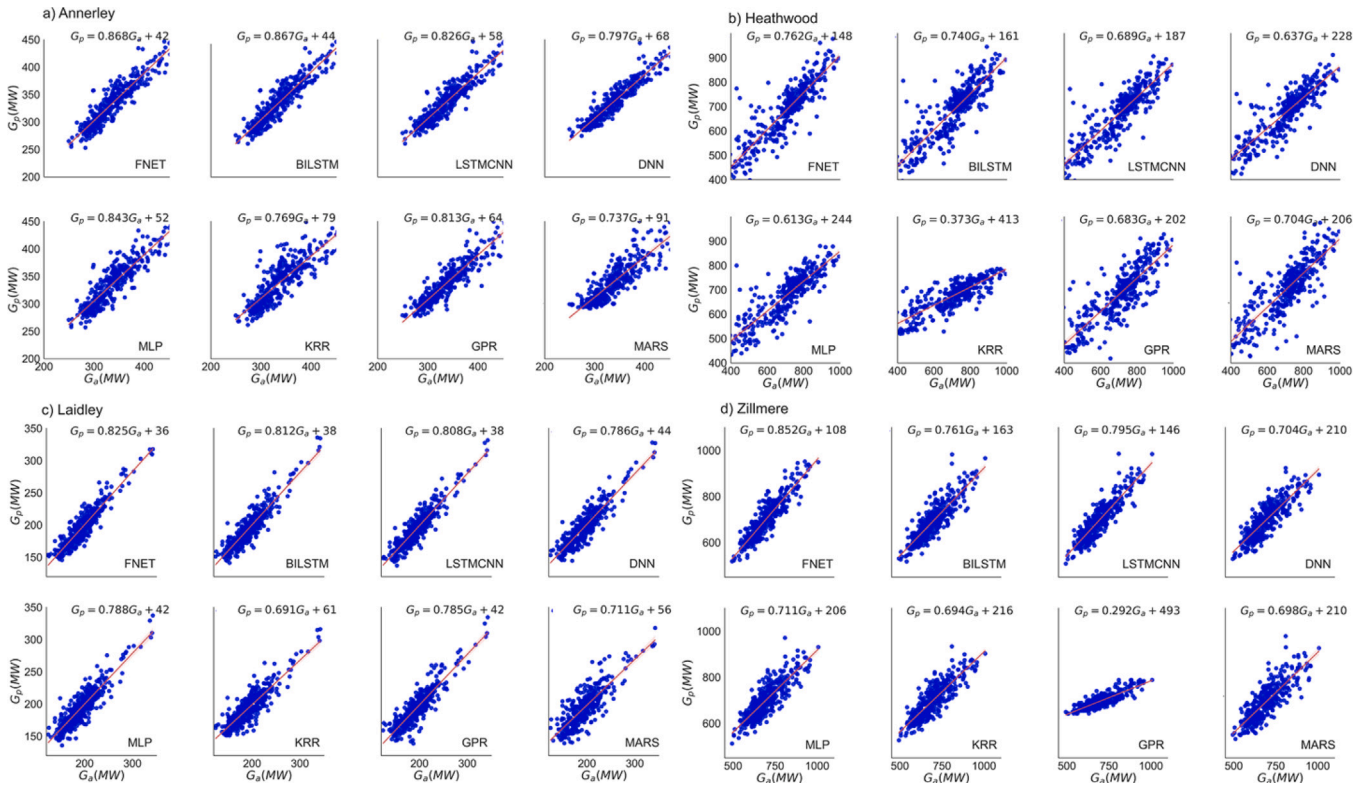


Fig. 9. Variation presentation in the form of Scatter plots for the simulated daily G at all the modelled stations. The red line shows least-square regression $y = mx + c$, where y is the $G_p(predicted)$, x is the $Ga(observed)$, and r is the correlation coefficient. The name of each model is provided in Table 3.

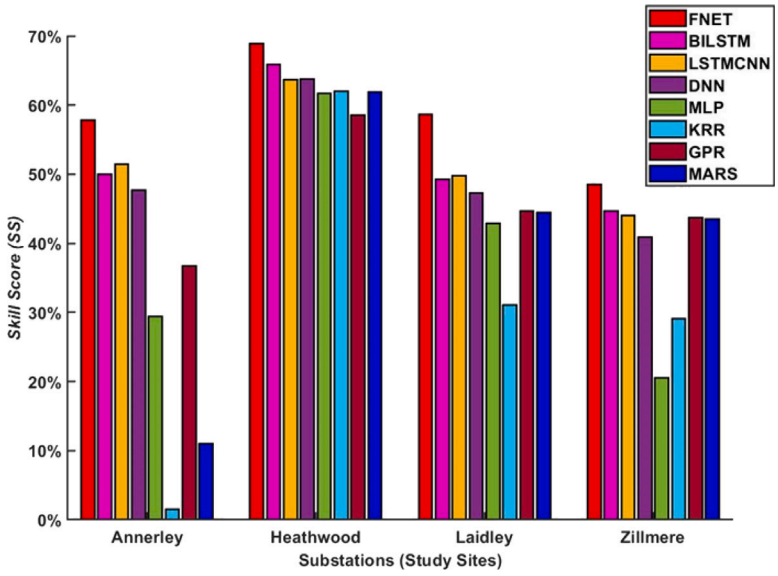


Fig. 10. Bar chart showing Skill Score Metric (SS) of the proposed FNET and the alternative benchmark models. The persistence model considers that G at t equals the G at $t+1$ and assumes that electricity use patterns are stationary.

Therefore the persistence model assumes that the G at a particular time will be the same as measured one day before for lead periods up to one day, one week before for lead times of one week, and one year before for lead times of one year.

To further explore the efficacy of the proposed FNET model, we refer to Table 7 that represents the Standard Deviation of the Relative Error ($STDRE$) and the Explained Variance (E_{var}) computed for the testing phase of experiment. As expected, the proposed FNET model has produced the best scores in terms of both metrics. For example, for the Laidley substation we note that $STDRE \approx 4.369$ and $Evar \approx 0.860$, compared with the KRR model with $STDRE \approx 5.742$ and $E_{var} \approx 0.764$ that appear to indicate the worst model. The distribution of absolute prediction error ($|PE|$) were also visually explored further through the box plots represented in Fig. 11 and empirical cumulative distribution function ($ECDF$) in Fig. 12.

In Fig. 11, we note that the proposed FNET model for all substations documented a smaller $|PE|$ division, which is in agreement with Tables 5 and 6. On the other hand, the LSTMCNN model appears to

Table 6
The geographic comparison of the Deep Hybrid Fused Network (FNET)) model vs. other comparative models in terms of the relative errors ($RRMSE, \%$) and ($RMAE, \%$) computed within the test sites. Note that the best model is boldfaced (blue).

Sub-Stations	Predictive Model	Model Performance Metrics	
		RRMSE	RMAE
Annerley	FNET	**4.48%**	**3.41%**
	BILSTM	4.88%	3.67%
	LSTMCNN	4.80%	3.67%
	DNN	4.99%	3.78%
	MLP	5.79%	4.35%
	KRR	6.95%	5.13%
	GPR	5.48%	4.15%
	MARS	6.50%	5.10%
Heathwood	FNET	**9.32%**	**7.19%**
	BILSTM	9.76%	7.75%
	LSTMCNN	10.08%	8.09%
	DNN	10.05%	8.09%
	MLP	10.35%	8.33%
	KRR	10.30%	8.26%
	GPR	10.75%	8.91%
	MARS	10.31%	8.35%
Laidley	FNET	**6.67%**	**5.43%**
	BILSTM	7.39%	5.89%
	LSTMCNN	7.35%	5.81%
	DNN	7.53%	6.19%
	MLP	7.84%	6.35%
	KRR	8.61%	7.00%
	GPR	7.71%	6.14%
	MARS	7.73%	6.01%
Zillmere	FNET	**5.73%**	**4.55%**
	BILSTM	5.94%	4.67%
	LSTMCNN	5.98%	4.57%
	DNN	6.14%	4.84%
	MLP	7.12%	5.69%
	KRR	6.73%	5.29%
	GPR	5.99%	4.77%
	MARS	6.00%	4.77%

Table 7
The testing performance of the Deep Hybrid Fused Network (FNET) model vs. LSTMCNN, DNN, BILSTM, MLP, KRR, GPR and MARS models as measured by Standard Deviation of Relative Error ($STDRE$), and Explained Variance (E_{var}).

Sub-stations	Predictive Model	Model Performance Metrics	
		STDRE	Evar
Annerley	FNET	**2.762**	**0.889**
	BILSTM	3.105	0.868
	LSTMCNN	2.993	0.872
	DNN	3.078	0.865
	MLP	3.632	0.814
	KRR	4.476	0.733
	GPR	3.429	0.834
	MARS	3.923	0.766
Heathwood	FNET	**7.287**	**0.790**
	BILSTM	8.128	0.747
	LSTMCNN	7.811	0.754
	DNN	7.777	0.754
	MLP	7.807	0.750
	KRR	7.583	0.758
	GPR	7.660	0.777
	MARS	7.343	0.776
Laidley	FNET	**4.369**	**0.860**
	BILSTM	4.571	0.839
	LSTMCNN	4.579	0.850
	DNN	5.234	0.823
	MLP	5.411	0.805
	KRR	5.742	0.764
	GPR	5.201	0.810
	MARS	4.849	0.830
Zillmere	FNET	**3.555**	**0.824**
	BILSTM	3.576	0.808
	LSTMCNN	3.570	0.813
	DNN	3.764	0.795
	MLP	4.202	0.727
	KRR	4.109	0.756
	GPR	3.669	0.806
	MARS	3.646	0.804

have an immediate performance compared to the FNET model followed by the BILSTM, DNN, GPR, MLP, KRR, and the MARS model. On the contrary, because the distributions created by the proposed FNET model were evenly dispersed with a limited number of outliers points for all four substations, the box plots show a clear distinction in the model performance. In particular, the $ECDF$ line plots representing the benchmark models showed a very close profile for all of the four substations. The $ECDF$ profile of the proposed FNET, on the other hand, revealed a remarkably narrow profile constrained within the smallest range at all of the four substations.

We now show Fig. 13 that depict a detailed account of the predictive skill of the proposed FNET model where the frequency distribution of $|PE|$ caused by the FNET *vs.* the alternative models is shown. Notably, the value of $|PE|$ achieved by the proposed FNET model was within the lowest range for all of the four substations. Consequently, for all four substations, the box plots in Fig. 11, together with the $ECDF$ plots in Fig. 12, and cumulative frequency plot in Fig. 13 further indicate the proposed FNET model's superiority in daily G prediction when compared with the competing benchmark models.

The efficacy of the proposed FNET model was also evaluated using the Willmott's Index (E_{WI}), Nash–Sutcliffe Coefficient (E_{NS}) and the Legates & McCabe's (E_{LM}) index (Table 8). It should be noted that E_{WI} is an improved metric over $RMSE$ and MAE which aims to overcome the insensitivity issues when differences between observed and predicted G values are not squared. Considering all four substations, the proposed FNET model seems to perform the best to attain the highest E_{WI}, E_{NS}, and E_{LM} except for the case of Zillmere sub-station with $E_{WI} \approx 0.878$ to fall just after the LSTMCNN model with $E_{WI} \approx 0.884$. For example, at the Laidley study site, the proposed FNET model seems to yield $E_{WI} \approx 0.891$, $E_{NS} \approx 0.859$, and $E_{LM} \approx 0.603$ followed by the LSTMCNN model with $E_{WI} \approx 0.891$, $E_{NS} \approx 0.832$, and $E_{LM} \approx 0.558$ and BILSTM with $E_{WI} \approx 0.885$, $E_{NS} \approx 0.828$, and $E_{LM} \approx 0.556$. These metrics when computed for the MLP and KRR model appear to be the lowest with $E_{WI} \approx 0.870$ and 0.869, $E_{NS} \approx 0.804$ and 0.764, and $E_{LM} \approx 0.536$ and 0.483, respectively. In corroboration with the previous findings, the E_{WI}, E_{NS}, and the E_{LM} values yield consistent results and therefore indicate that the deep hybrid FNET model is able to predict the G values more correctly than the benchmark models.

We now revert to Absolute Percentage Bias ($APB, \%$) and Kling–Gupta Efficiency (KGE), as per Fig. 14(a, and the global performance indicator (GPI), as per Fig. 14(b). With the lowest APB and the highest KGE and GPI, we note that the proposed FNET model outperformed all benchmark models. According to the GPI, the lowest performing model were the KRR and the MARS model for daily G predictions. Fig. 15 represents the performance comparison using Combined Performance Index (CPI), where a lower percentage of CPI could imply a more robust model. While the results reconfirmed the superiority of the proposed FNET model across all substations, it is interesting that the KRR model, which is the worst model according to the other metrics, yielded the second-best percentage of CPI, which lies just after the value for the proposed FNET model for the Laidley study site. These findings also reaffirm that the proposed FNET model outperforms the benchmark models for daily prediction of electricity demand.

Although various error indicators so far showed the differences in predictive accuracy of models, these results need further careful

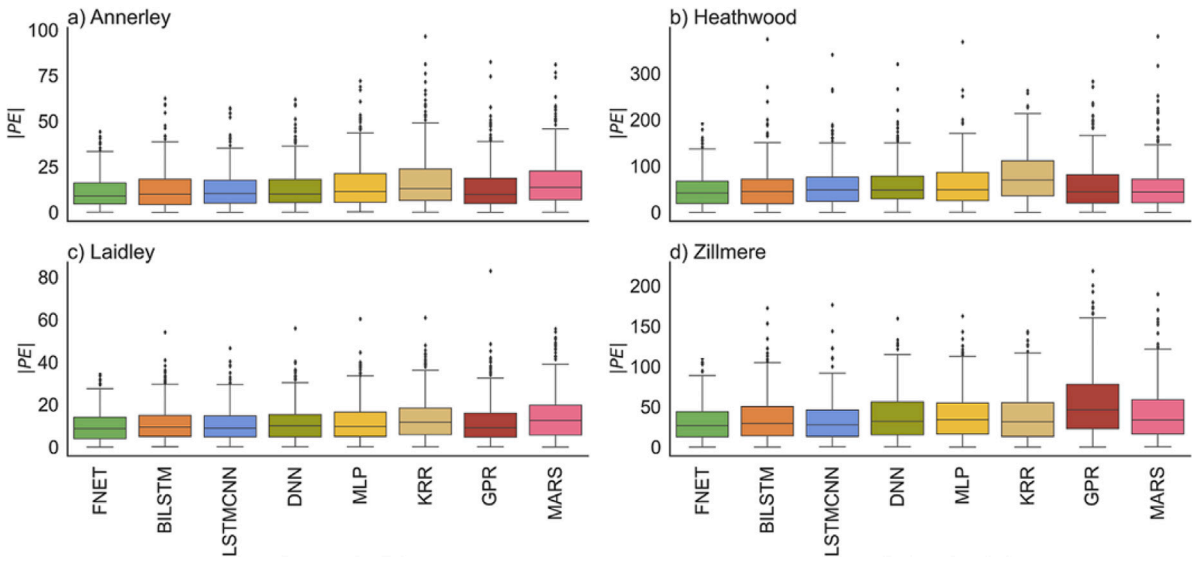


Fig. 11. Box plot exemplifying the veracity of the proposed FNET model in terms of the overall distribution of the Prediction Error ($|PE|$ (MW)) computed against the alternative models.(a) Annerley, (b) Heathwood, (c) Laidley, (d) Zillmere sub-station.

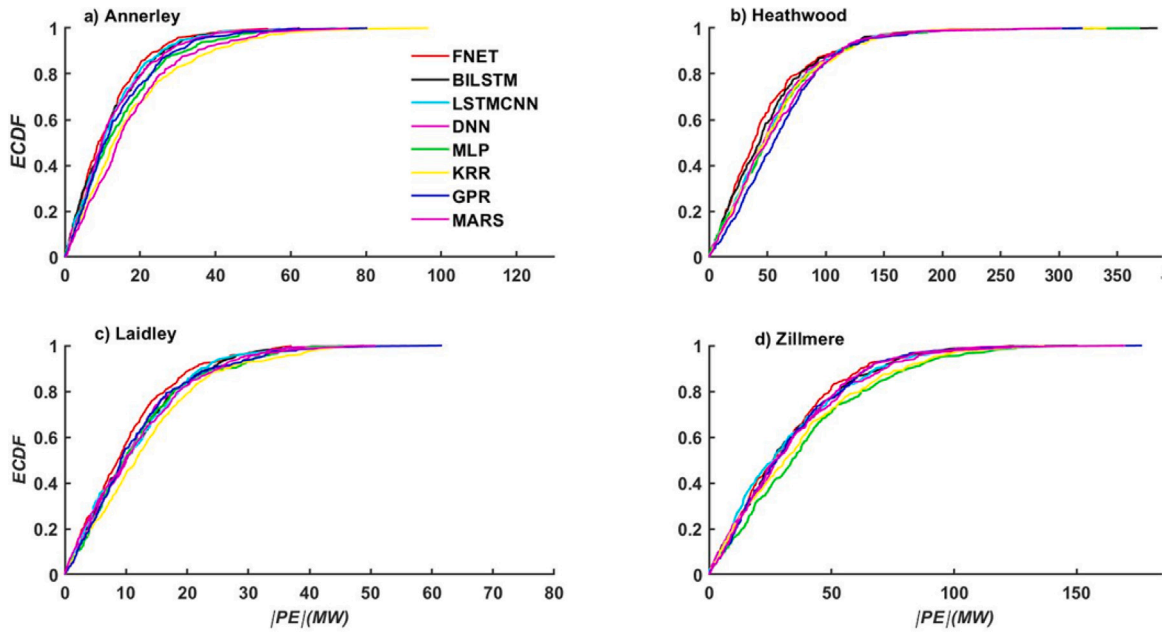


Fig. 12. Empirical cumulative distribution function ($ECDF$) for $|PE|$ (MW) of the G predicted by the LSTMCNN, DNN, BILSTM, MLP, KRR, GPR and MARS models against the proposed FNET model.

consideration as the variations in model accuracy could be driven by the nature of the data and its features. To address this issue, we jointly apply the Diebold–Mariano (DM) and Harvey–Leybourne–Newbold (HLN) statistical test to quantify the differences in accuracy between these models, aiming to determine whether two predictions are significantly different.

Tables 9 and 10 show DM, HLN and λ. Importantly, both test statistics indicate superior performance of the proposed FNET model relative to the benchmark models, certainly depicts the improvements made on the LSTMCNN and BILSTM models in accordance with the absolute values of DM being larger than 1.96 - the $z-score$ of 5% significance level. The observed differences between LSTMCNN and FNET models are also quite significant with the absolute value of the $DM \approx 2.9547 > 1.96$. Similarly, between the BILSTM and the proposed FNET model, the absolute value of the $DM \approx 2.1264 > 1.96$.

In terms of the λ shown Table 10, when the LSTMCNN model is compared with the FNET model, the model improvement is evident in $RMSE$, APB and KGE as being $\approx 7.25\%, \approx 7.82\%, \approx 7.23\%$, respectively (Annerley substation), $\approx 8.13\%, \approx 1.68\%, \approx 11.71\%$, respectively (Heathwood substation), $\approx 10.22\%, \approx 6.34\%, \approx 11.29\%$, respectively (Laidley substation) and $\approx .22\%, \approx 3.71\%, \approx 2.49\%$ (Zillmere substation). Therefore the DM, HLN and Promoting Percentages further ascertain that the predictive capability of the FNET model is considerably better than the benchmark models.

Fig. 16 shows the directional symmetry (DS) criteria whereby the proposed FNET model scored the highest $DS \approx 87.74\%$, and this

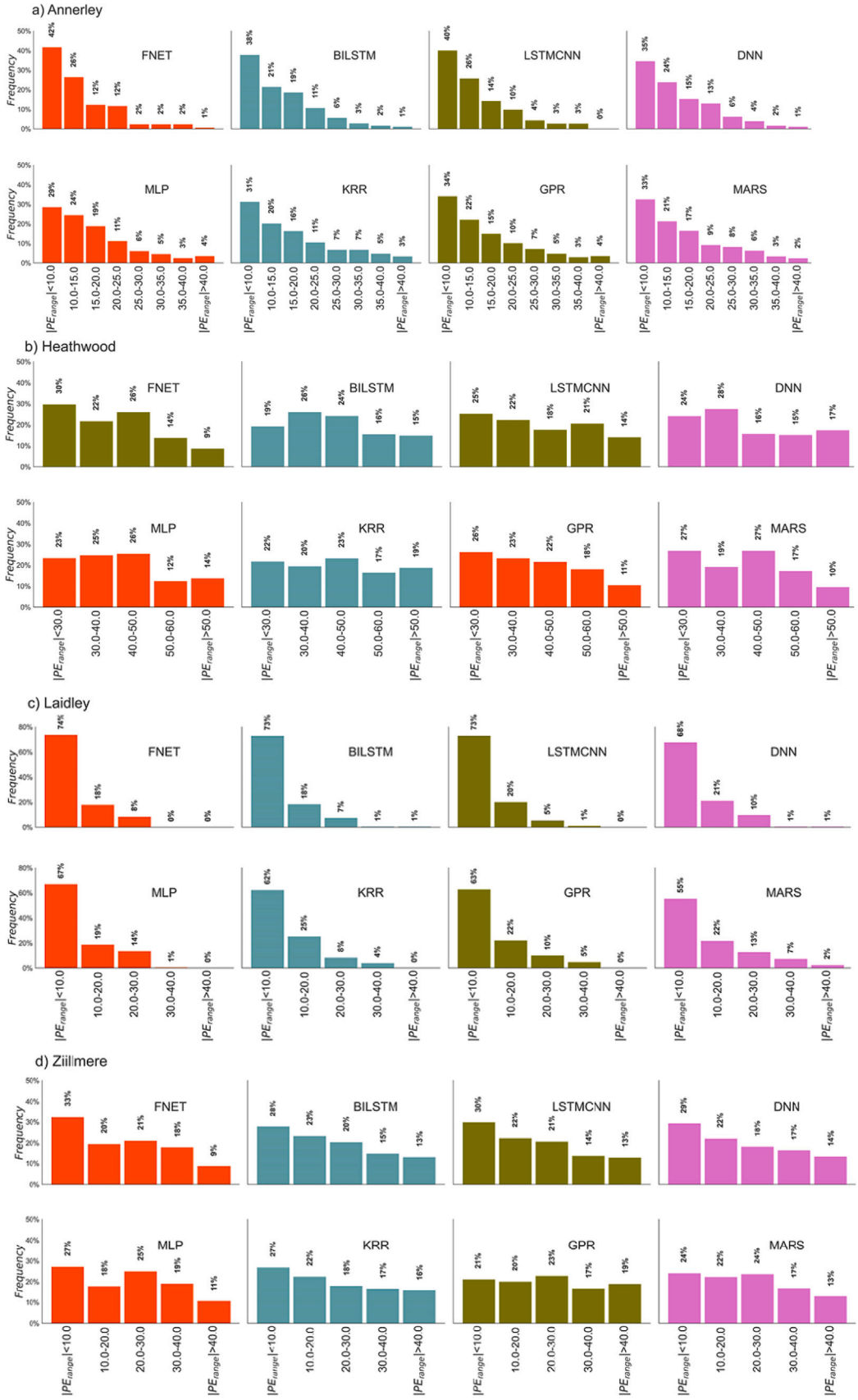


Fig. 13. Cumulative frequency of the Prediction Error($|PE|$(MW)) for four substations at (a) Annerley, (b) Heathwood, (c) Laidley and (d) Zillmere.

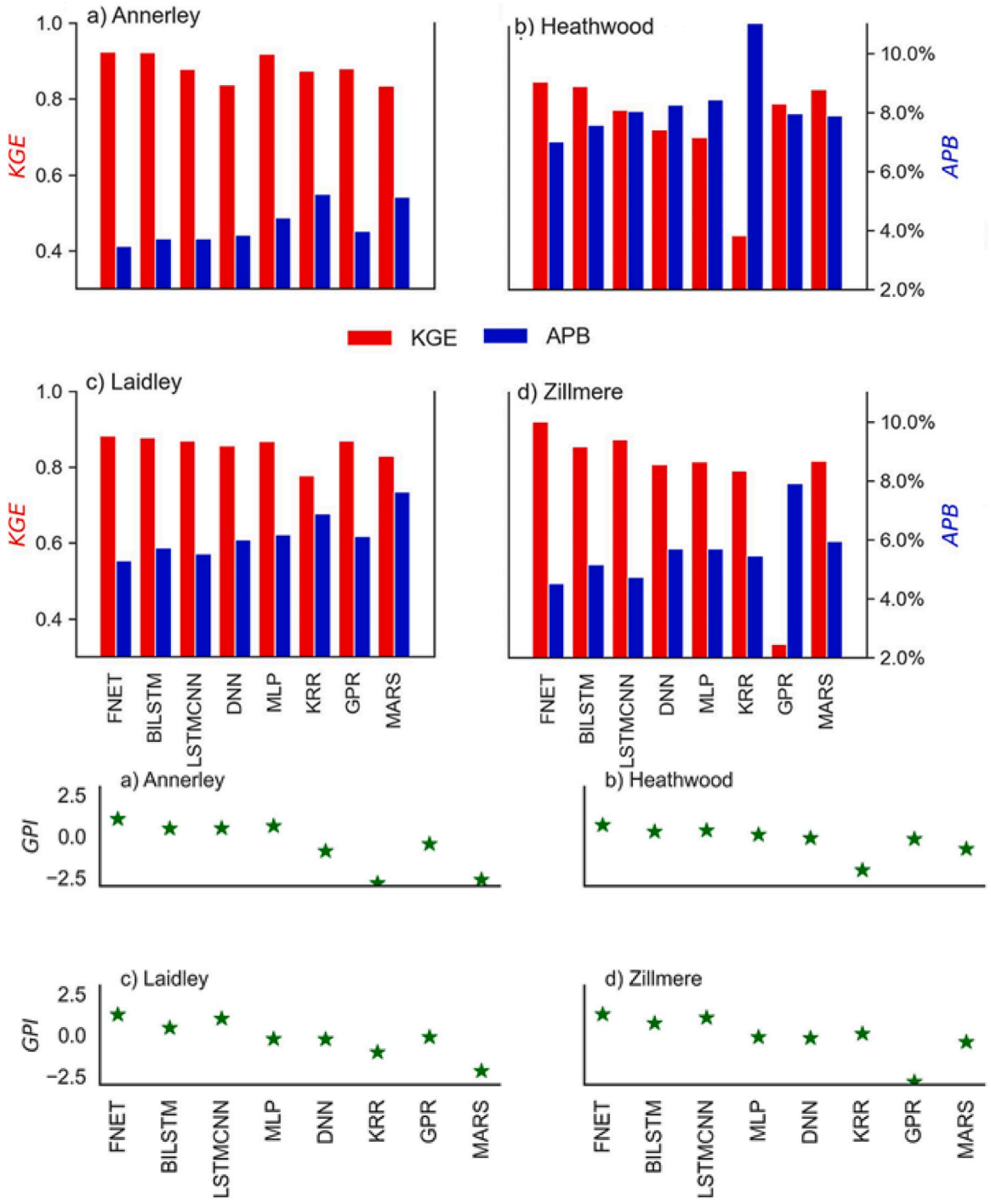


Fig. 14. (a) Absolute Percentage Bias ($APB, \%$) and Kling–Gupta Efficiency (KGE), (b) Global Performance Indicator (GPI) used to evaluate the proposed FNET model in respect to several benchmark models.

value remained considerably higher than those of seven other models by magnitude of 23–26%. Fig. 17 provides additional information on the performance of the FNET and benchmark models by using Taylor diagrams. In particular, the Taylor diagram depicts the three complementary model performance that comprises of the Standard Deviation, Centralized Root Mean Square Error ($CRMSE$), and the Correlation between predicted and observed electricity demand in the testing phase. The diagrams also indicate that the simulated point by the FNET model is closer to the observation (OBS) compared with other benchmark models and implies that the predictions derived from FNET and the observations have a similar standard deviation and higher correlation (≈ 0.93), and $CRMSE$ is closer to zero. In congruence with Table 9, Table 10, Figs. 16 and 17, we can confirm that the proposed FNET model demonstrated better and reliable prediction capability at all four substations.

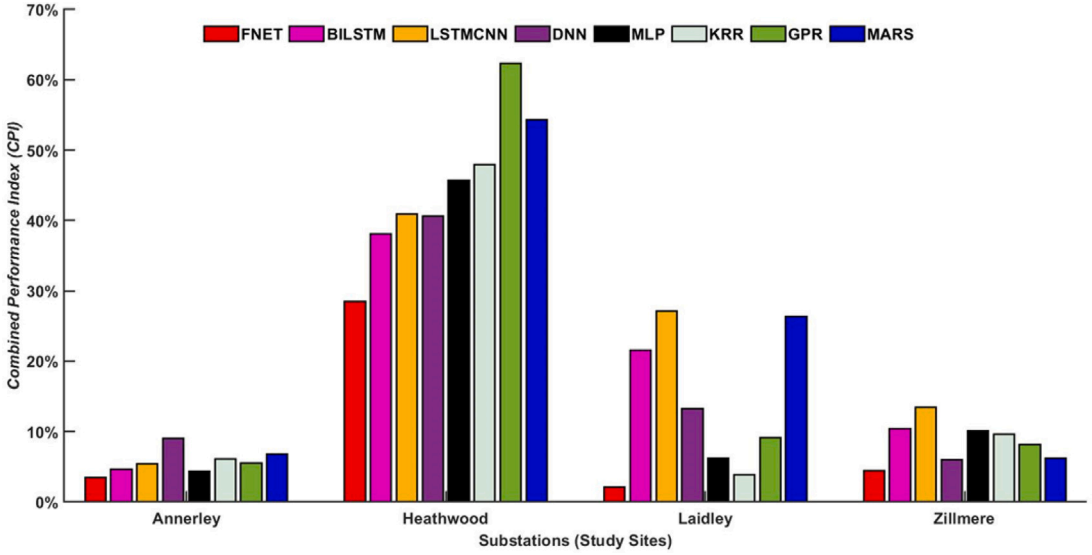


Fig. 15. Bar chart showing the efficacy of the proposed FNET model in terms of Combined Performance Index (CPI, %) for four substations. (a) Annerley, (b) Heathwood, (c) Laidley, (d) Zillmere.

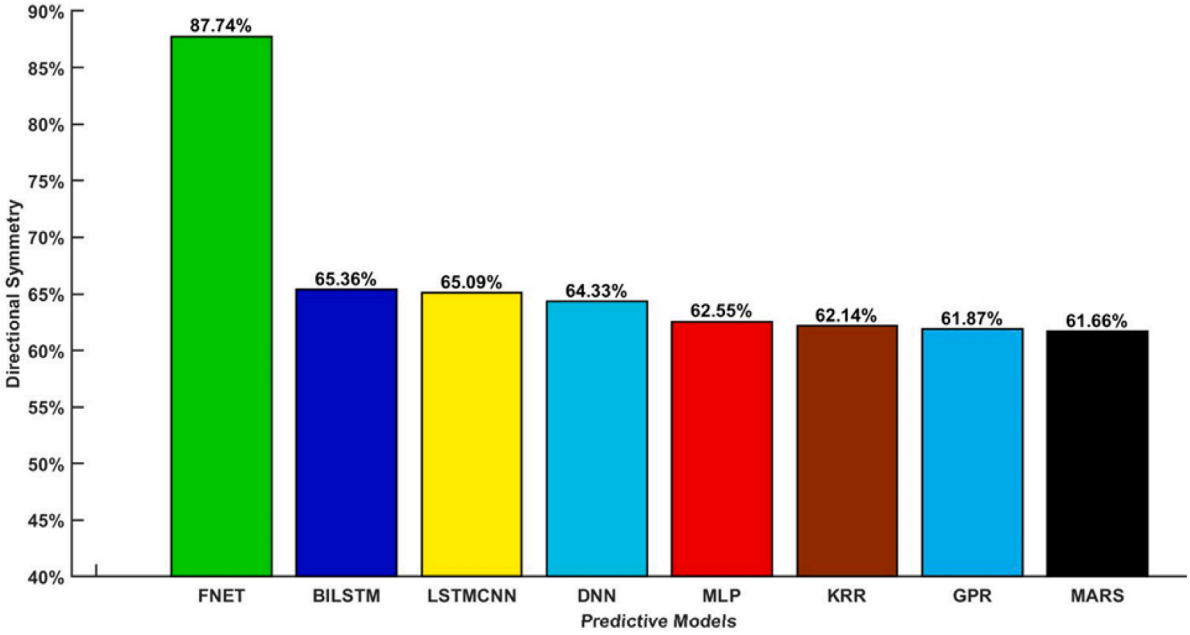


Fig. 16. The criteria of directional symmetry (DS) assessment for the introduced model (*i.e.*, FNET) and the benchmark models.

4.1.1. Uncertainty evaluation

To explore the errors encountered in point-based predictions of daily electricity demand dataset, we now quantify the inherent uncertainties generated by the proposed FNET and benchmark models. We express this uncertainty as the prediction interval PIs of the underlying distribution of the predictive model errors in the testing phase. While the PIs can provide a lower bound and an upper bound for these predictions, the modelling process described earlier can only provide a point-based prediction. Therefore, as explained in Section 3.1.4, the residual bootstrap approach was able to compute the uncertainties for each model, as shown in Tables 11 and 12.

It is evident that the $PICP$ of the proposed FNET model was not significantly different from that of the benchmark models evaluated at the 95% confidence level. However, the $MPIW$ values are relatively higher, and the F values are relatively lower. In particular, the $MPIW$ value for the daily prediction of G emulated by the proposed FNET model is $\approx$ 57.97 for the Annerley substation compared with a value of $\approx$ 64.92, $\approx$ 64.90, $\approx$ 67.44, $\approx$ 82.83, $\approx$ 101.24, $\approx$ 70.65, and $\approx$ 94.06 for the BILSTM, LSTMCNN, DNN, MLP, KRR, GPR and the MARS models, respectively. Among the benchmark models, we note that the KRR model has attained the highest value of $MPIW$ compared with the other predictive models.

It is important to note that there was an $\approx$ 11% reduction in the magnitude of $MPIW$ when comparing the proposed FNET model with the BILSTM and LSTMCNN models. Similarly, for the case of DNN, MLP, GPR, KRR and MARS models, we noted a reduction in the $MPIW$ value of $\approx$ 14%, $\approx$ 30%, $\approx$ 18%, $\approx$ 38%, and $\approx$ 42%, respectively, for the Annerley substation. A similar trend could also be seen for the Heathwood, Laidley and Zillmere substations.

When referred to the more comprehensive index, which is actually the weighted harmonic average of the $PICP$ and $1/MPIW$ metrics used to evaluate the quality of the $PIs \cdot F$ value, the proposed FNET

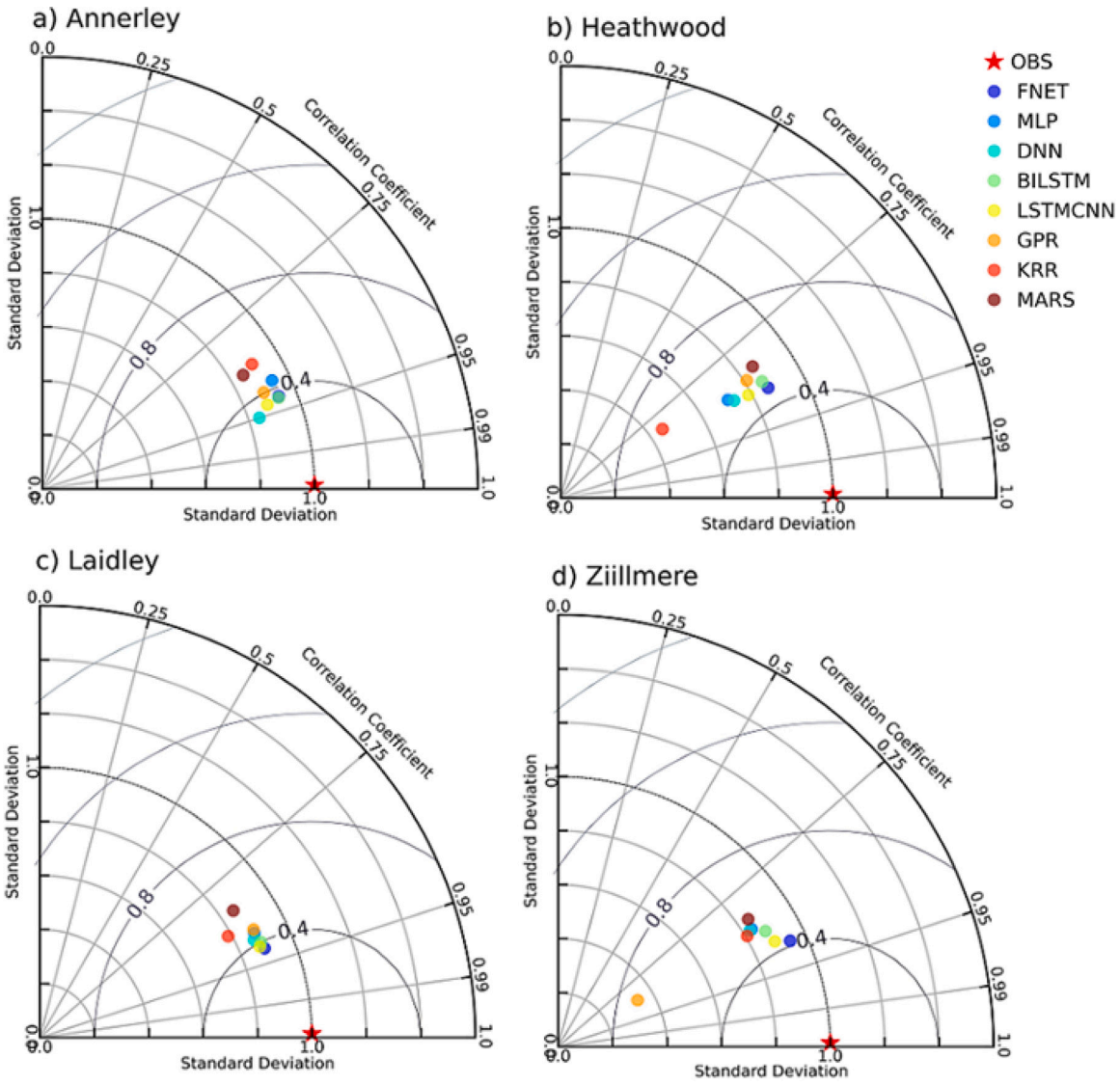


Fig. 17. Two dimension graphical presentation "Taylor diagram" for the predictive FNET model evaluation with the benchmark models of daily G over the testing phase.

model registered larger values than that of the benchmark models for all tested substations. In respect to the sharpness of the generated PI, denoted as the Winkler Score (WS), the proposed FNET model also appeared quite superior. In fact, the WS tends to reward the narrow PI values and penalizes them if the targets are not successfully captured by the PI value whereas a good quality PI is expected to have a lower absolute value of the WS for a given confidence level. Table 12 presents the WS at the 95% confidence level and the Average Relative Interval Width $ARIL$ value derived from the PIs. Evidently, the proposed FNET model had a smaller magnitude of WS and $ARIL$ compared with the benchmark models, for example, the Annerley site where the FNET model generated $WS \approx 68.251$ and $ARIL \approx 0.174$ compared with a value of $WS \approx 74.54$ and $ARIL \approx 0.195$ noted for the second best model (*i.e.*, the LSTMCNN) and a value of $WS \approx 112.162$ and $ARIL \approx 0.305$ noted for the worst performing (*i.e.*, the KRR) model.

Fig. 18 is a visual representation of the quality of PIs for eight models used in daily prediction of G in testing phase. It is observable that the predicted G falls within the lower bound and the upper bound, as shown by the grey area, and thus provides a good probability of the predicted value - a factor that is significantly beneficial to decision-makers in the energy industry. In Fig. 18, we also show the Continuous Ranked Probability Score ($CRPS$) and the PI Normalized Average Width ($PINAW$) for all models. Importantly, the proposed FNET model has produced the lowest value of $CRPS$ and $PINAW$ ($CRPS \approx 14.536, PINAW \approx 0.301$ for Annerley, $CRPS \approx 65.302, PINAW \approx 0.387$ for Heathwood, $CRPS \approx 13.415, PINAW \approx 0.251$ for Laidley, and $CRPS \approx 38.646, PINAW \approx 0.244$ for Zillmere stations) relative to all benchmark models. Overall, these results ascertain that the proposed FNET model has superior performance in terms of both confidence intervals and point-based prediction of daily electricity demand.

4.1.2. SHAP interpretation of the FNET model

The SHAP violin summary plots (Fig. 19) for the four stations illustrate the impact of various features on the model's predictions (global explanation). The $x-axis$ represents the SHAP value, indicating the magnitude and direction of each feature's impact on the model's output. Positive SHAP values suggest that the feature contributes to an increase in the predicted value, while negative SHAP values indicate a decrease. The $y-axis$ lists the features in descending order of their importance, from top to bottom. Each point on the plot corresponds to an instance from the dataset, with the colour gradient from blue to red representing the feature value: blue for low values and red for high values. This colour-coding helps to visualize the relationship between feature values and their corresponding SHAP values, revealing patterns and insights about how each feature influences the model's predictions across different instances.

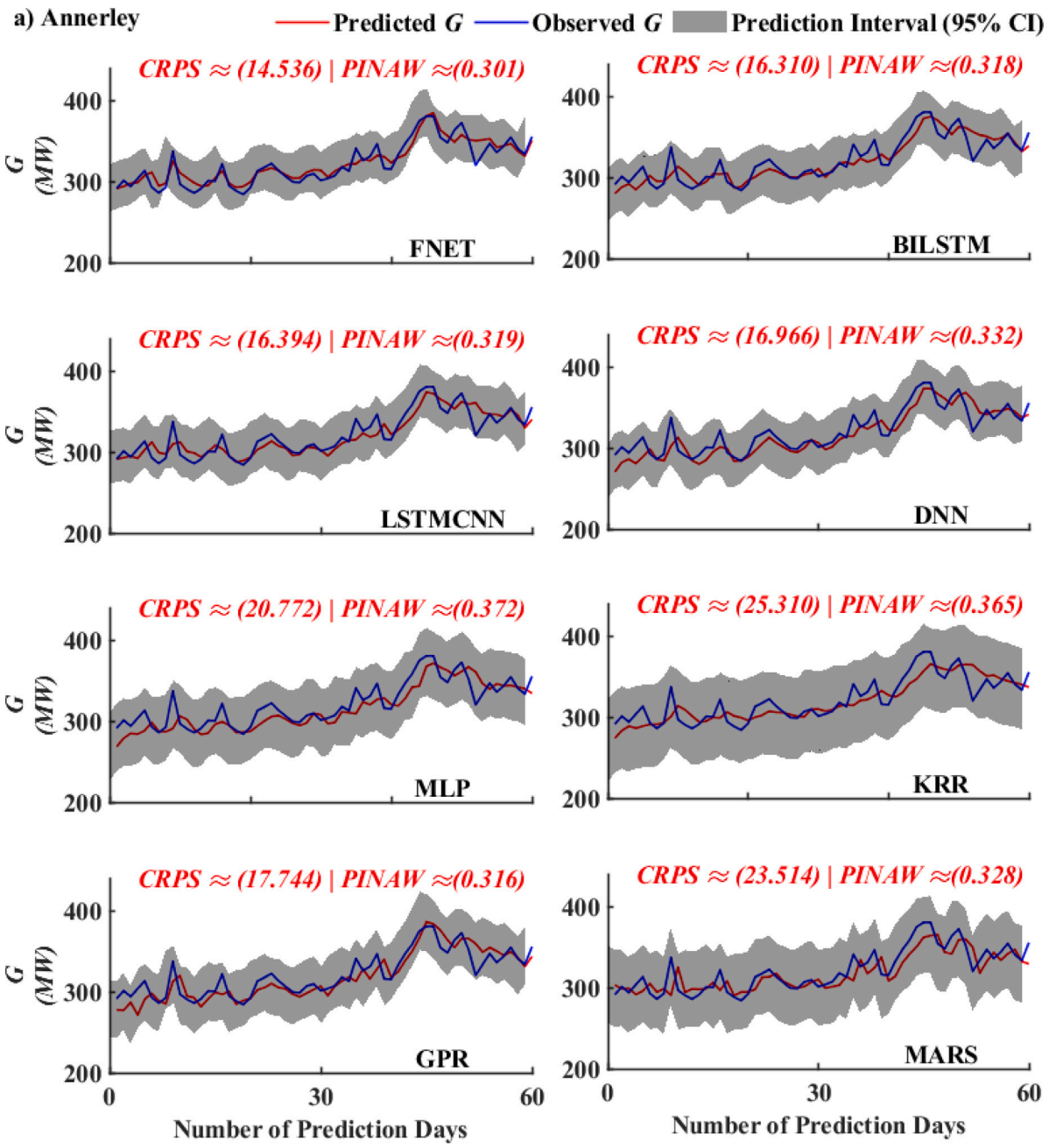


Fig. 18. Daily predicted G and PIs at the 95% confidence level. Continuously Ranked Probability Score ($CRPS$) and Prediction Interval Normalized Average Width ($PINAW$) are shown. (a) Annerley, (b) Heathwood, (c) Laidley, (d) Zillmere substations. For conciseness, only the last 60 days of the predicted G values are shown.

- **Heathwood Substation**: The most influential features are $G_{(t-1)}$, $G_{(t-6)}$, $G_{(t-2)}$, $G_{(t-5)}$, and $Etn_{(t-1)}$. The plot indicates that $G_{(t-1)}$ and $G_{(t-6)}$ have the largest positive impact on the model output when their values are high (red points), whereas lower values (blue points) of these features have a negative impact. $Tmax_{(t-1)}$ also shows a notable impact but to a lesser extent.
- **Annerley Substation**: Similar to Heathwood, $G_{(t-1)}$, $G_{(t-5)}$, and $Etn_{(t-1)}$ are key predictors. The $Esyn_{(t-1)}$ and $VP_{(t-1)}$ features also play significant roles. The plot shows that high values of $G_{(t-1)}$ and $G_{(t-5)}$ contribute positively to the model output, whereas high values of $Etn_{(t-1)}$ and $VPd_{(t-1)}$ contribute negatively.
- **Laidley Substation**: Here, $G_{(t-1)}$, $G_{(t-6)}$, and $Etn_{(t-1)}$ are the most impactful features. The $Tmax_{(t-1)}$ and $Tmin_{(t-5)}$ also show considerable influence. High values of $G_{(t-1)}$ and $G_{(t-6)}$ are associated with a positive impact on the model output, while high values of $Etn_{(t-1)}$ tend to have a negative impact.
- **Zillmere Substation**: In this location, $G_{(t-1)}$, $GSR_{(t-1)}$, and $Tmin_{(t-1)}$ are prominent features. The $VP_{(t-1)}$ and $VPd_{(t-1)}$ are also significant. The plot reveals that high values of $G_{(t-1)}$ and $GSR_{(t-1)}$ contribute positively to the model output, while high values of $Tmin_{(t-1)}$ and $VP_{(t-1)}$ contribute negatively.

For each location, the importance of features such as historical load values G, temperature $Tmin$, $Tmax$, and humidity $Rhmax$, $Rhmin$ varies, but generally, recent historical load $G_{(t-1)}$ and evapotranspiration Etn have significant impacts. Features like $G_{(t-1)}$ and Etn usually have positive impacts, meaning higher values of these features increase the model output. Other features like VP and VPd can have mixed impacts depending on their values. The impact of each feature can vary

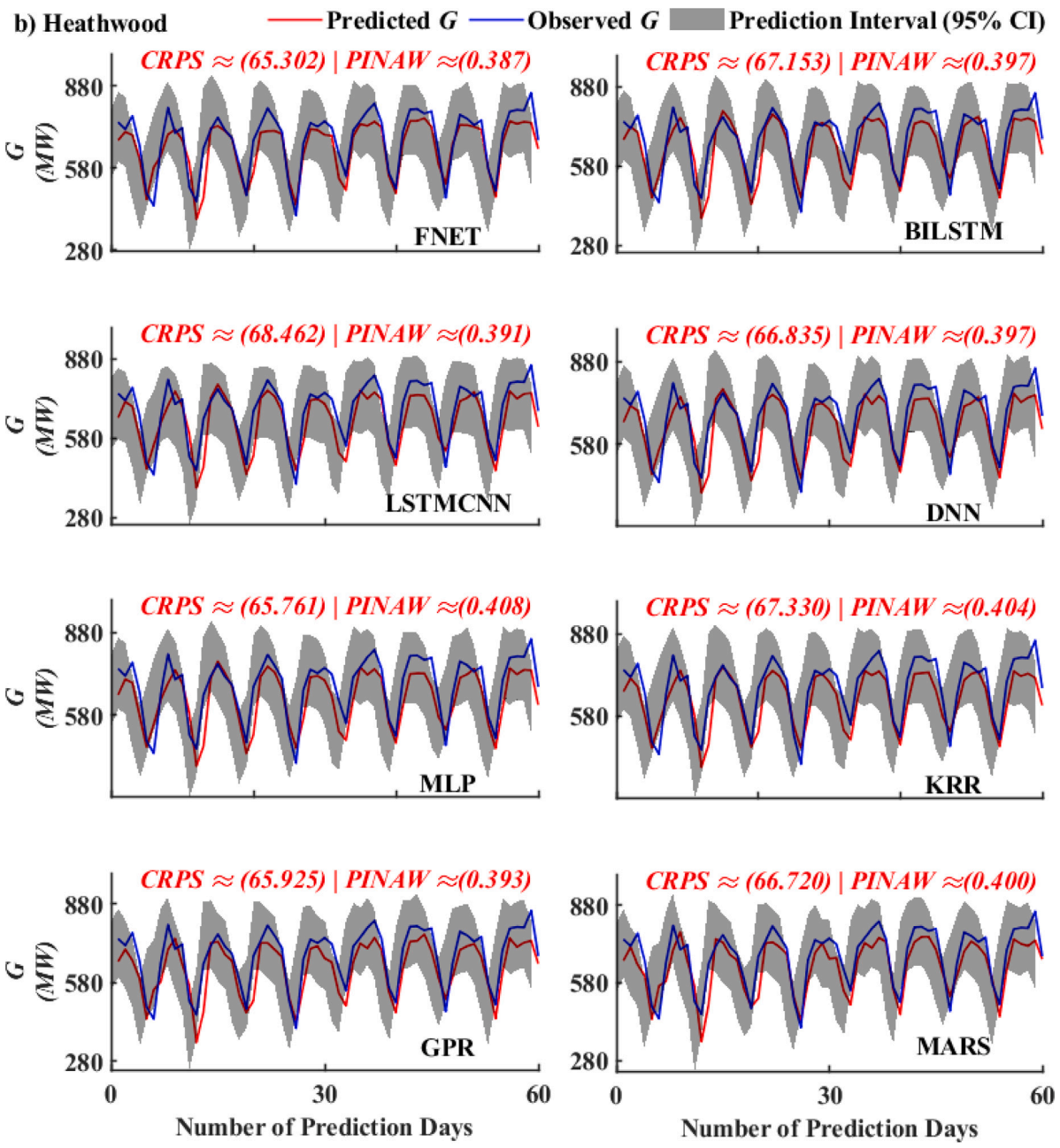


Fig. 18. (*continued*).

significantly across different locations, indicating that local conditions and historical patterns play a crucial role in the model's predictions.

The SHAP bar plots in Fig. 20 for instance 50 across Heathwood, Zillmere, Annerley, and Laidley substation provide detailed insights into the importance and impact of individual features for specific predictions (local explanation). The analysis is summarized below:

- For Heathwood substation, the most impactful features include $G_{(t-1)}$, $GSR_{(t-1)}$, and $G_{(t-2)}$. Among these, $G_{(t-1)}$ and $GSR_{(t-1)}$ have positive impacts. This finding is consistent with the global summary, where historical load values and meteorological variables such as GSR and $Tmax$ are consistently important predictors.
- In Zillmere substation, the major contributing features are $G_{(t-1)}$, $G_{(t-6)}$, and $GSR_{(t-1)}$. Here, $G_{(t-1)}$ and $G_{(t-6)}$ show significant negative impacts. This observation aligns with the global importance of historical load values and meteorological features like Etn and $Tmax$.
- For Annerley substation, the dominant features include $G_{(t-1)}$, $Etn_{(t-1)}$, and $G_{(t-5)}$. In this instance, $G_{(t-1)}$ and $Etn_{(t-1)}$ exhibit substantial negative impacts, reflecting the global significance of historical load values and meteorological variables such as $Tmin$ and $Rhmax$.
- In Laidley substation, the key features are $G_{(t-1)}$, $Tmin$, and VP. Here, $G_{(t-1)}$ shows a strong positive impact, consistent with the global importance of historical load values and meteorological variables like $Tmin$ and GSR.

This comparative analysis highlights the consistency of historical load values (G) as significant predictors across locations in both instance-specific and global explanations. The exact impact of features varies based on local conditions and specific data patterns. The detailed instance-specific SHAP values provide insights into how the model makes predictions for specific instances, while the global summary plots offer an overarching view of feature

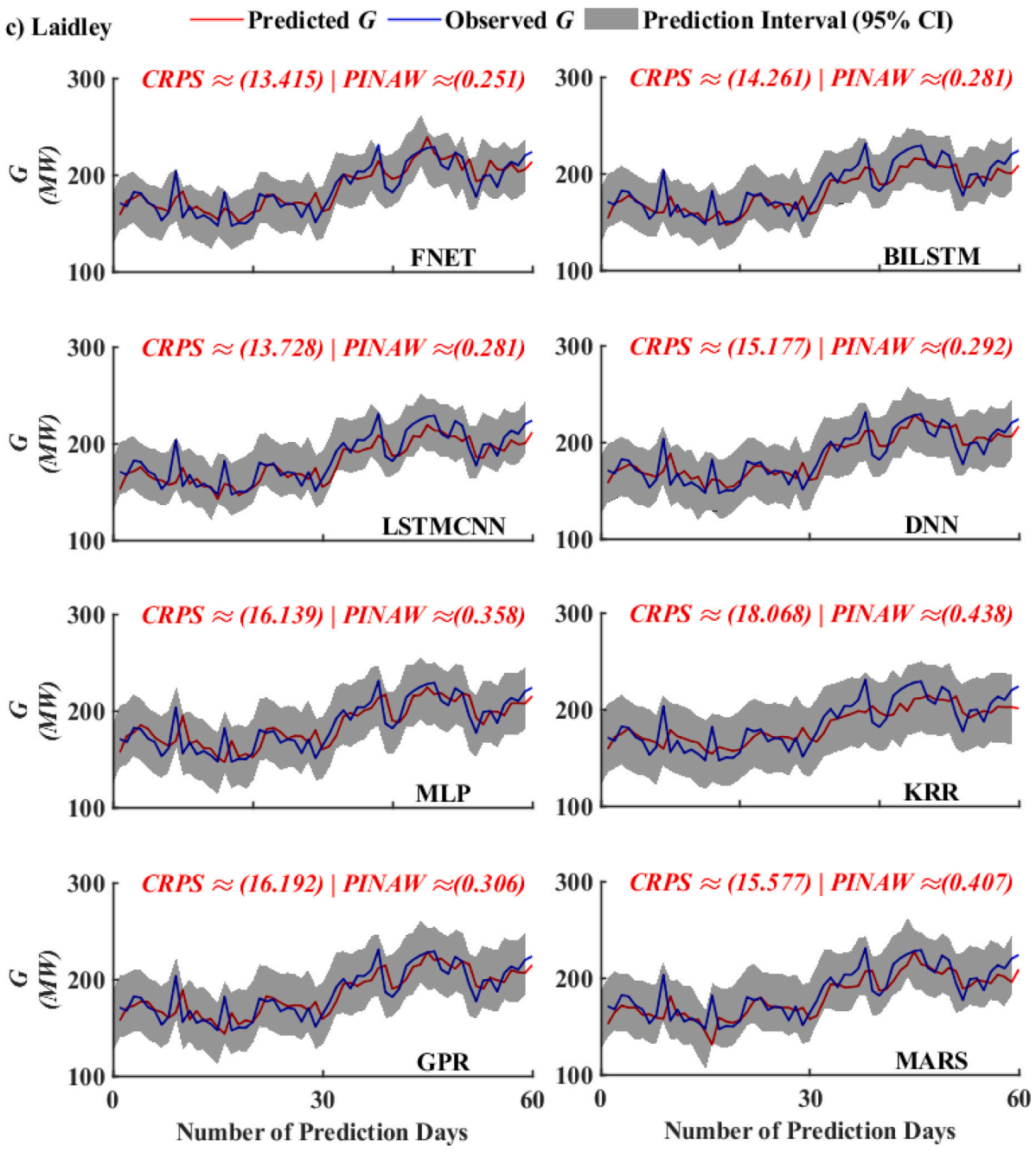


Fig. 18. (*continued*).

importance across the dataset. Emphasizing both perspectives is crucial for comprehensive model interpretability.

The Fig. 21 presents twelve SHAP (SHapley Additive exPlanations) dependence plots, each illustrating the relationship between various predictors and their influence on the model's predictions for the Zillmere sub-station. These plots are organized in a grid layout to allow for a detailed comparison of multiple features. The analysed features include lagged values of G (e.g., $G_{(t-1)}$, $G_{(t-2)}$, and $G_{(t-3)}$) as well as meteorological variables such as maximum temperature, vapour pressure, and vapour pressure deficit. A key observation is that the lagged values of G show a consistently strong, predominantly positive impact on the model's output, with the influence gradually weakening as the lag increases. For example, the first two rows display dependence plots for G lagged values (e.g., $G_{(t-1)}$, $G_{(t-2)}$, and $G_{(t-3)}$), revealing a clear pattern where higher G values correspond to larger SHAP values, which suggests that higher values of G enhance the model's predictions. This effect is particularly noticeable in the first few lags ($G_{(t-1)}$ to $G_{(t-4)}$), where SHAP values increase almost linearly with higher G values. In contrast, the meteorological features—such as maximum temperature, vapour pressure, and vapour pressure deficit—exhibit more complex, nonlinear contributions to the model's predictions, largely depending on their interactions with other variables. The third row introduces these environmental variables, including $Tmax_{(t-1)}$, $VP_{(t-1)}$, and $VPd_{(t-1)}$, and reveals intricate, nonlinear relationships between these features and the SHAP values. For instance, $Tmax_{(t-1)}$ generally displays an increasing trend, indicating that higher temperatures tend to positively contribute to the model's predictions. However, the spread in SHAP values suggests that this relationship is modulated by interactions with other features. Moreover, $VP_{(t-1)}$ and $VPd_{(t-1)}$ show varying contributions, implying that their influence can shift between positive and negative, depending on interactions

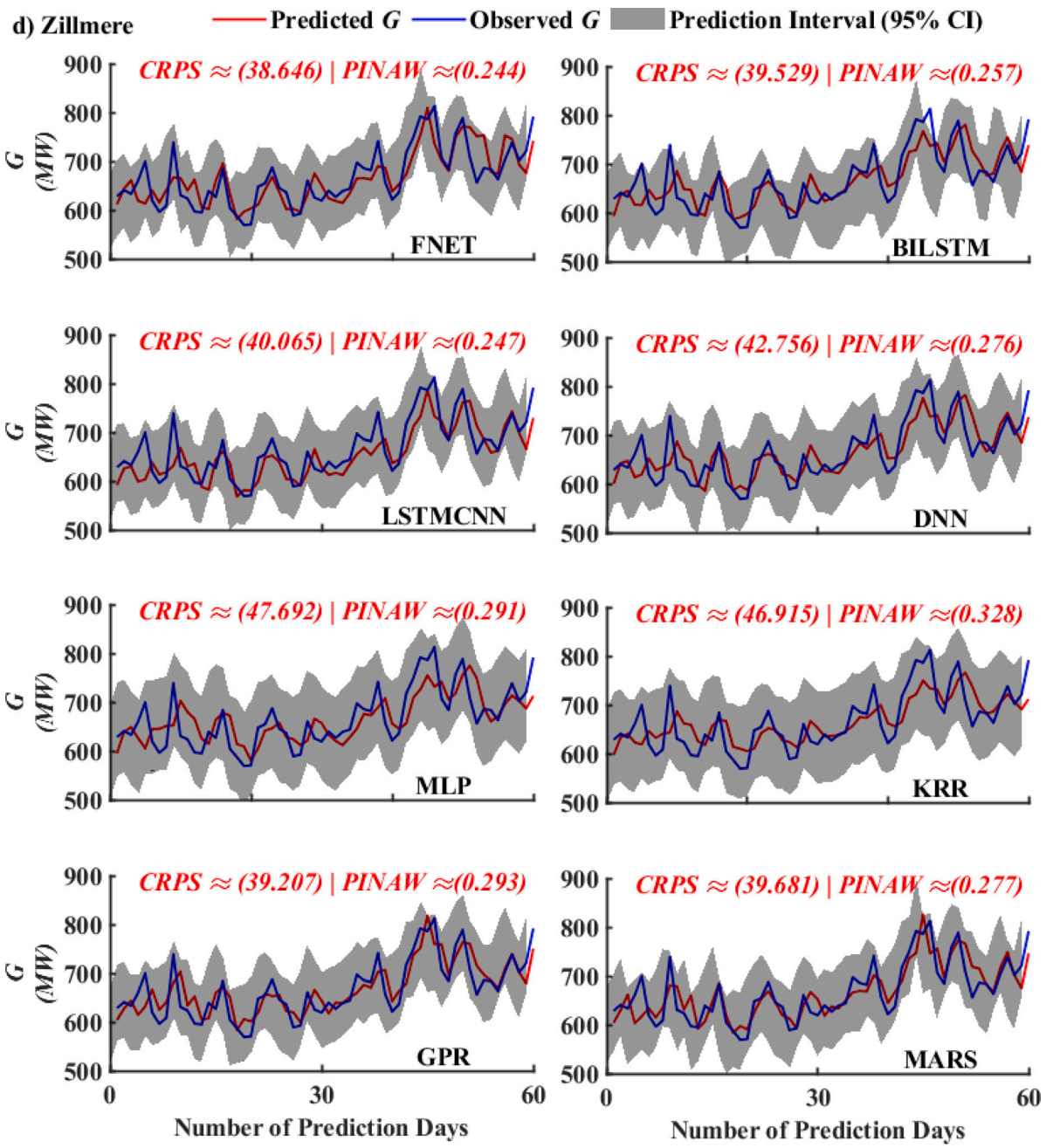


Fig. 18. (*continued*).

with other variables. The colour gradients in the plots help visualize these interactions, highlighting that some features, like $G_{(t-2)}$ and $G_{(t-3)}$, display SHAP values that shift based on the influence of another feature, as shown by the colour scale. Several features, such as $G_{(t-5)}$ and $Rhmax_{(t-1)}$, exhibit more scattered, nonlinear relationships with the target variable, indicating that their contributions to the model's predictions are less straightforward and depend heavily on their specific values and interactions with other variables.

The SHAP dependence plot for $Etn_{(t-1)}$ in Fig. 21 (last row first column) provides a detailed view of the interaction between $Etn_{(t-1)}$ and $G_{(t-1)}$, revealing important insights into how these two features jointly influence the model's predictions. As the values of $Etn_{(t-1)}$ increase, the corresponding SHAP values also rise, signifying a generally positive contribution of $Etn_{(t-1)}$ to the predictions. This trend indicates that higher levels of $Etn_{(t-1)}$ consistently push the model's output upward, making $Etn_{(t-1)}$ a key feature in determining the prediction accuracy. What makes this relationship particularly interesting is the interaction with $G_{(t-1)}$, which is represented by the colour gradient in the plot. As the colour shifts from blue to red, corresponding to increasing values of $G_{(t-1)}$, it becomes evident that the effect of $Etn_{(t-1)}$ on the model's predictions intensifies when $G_{(t-1)}$ is high. Specifically, when $G_{(t-1)}$ takes on higher values, the SHAP values for $Etn_{(t-1)}$ rise more steeply, indicating a strong positive interaction between these two variables. This suggests that the impact of $Etn_{(t-1)}$ is not independent; rather, it is amplified in scenarios where $G_{(t-1)}$ is elevated, meaning the two features work together to enhance the model's predictions. Additionally, the relationship between $Etn_{(t-1)}$ and its SHAP values is not purely linear. There is a noticeable spread in SHAP values, particularly at intermediate levels of $Etn_{(t-1)}$, which reflects a more complex interaction. This spread suggests that the influence of $Etn_{(t-1)}$ on the model's output is modulated by its interaction with $G_{(t-1)}$, as well as potentially other features. In particular, as $G_{(t-1)}$ changes, the

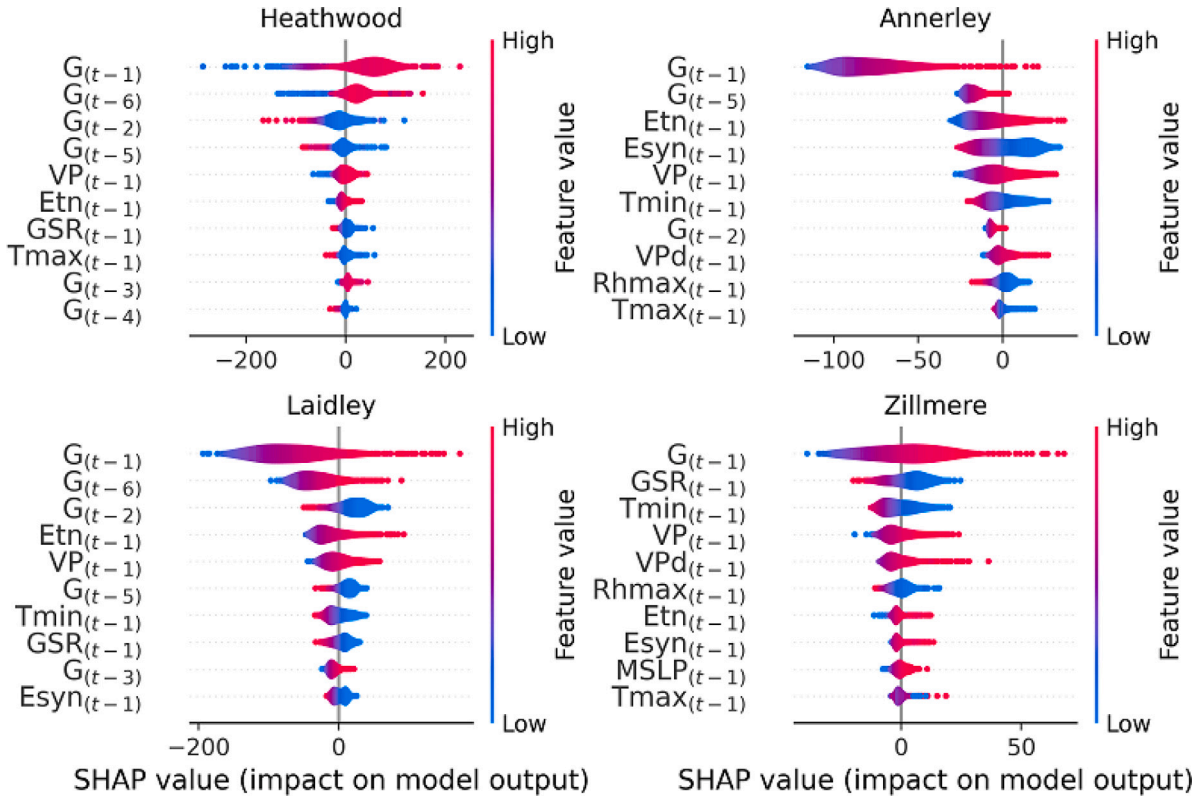


Fig. 19. Violin plots of the SHAP values (Global explanation) computed for each feature in the FNET model trained on different substations dataset. The colours vary from blue (low feature value) and red (high feature value). The SHAP values indicate the influence of each feature on model prediction. Negative SHAP values indicate that a specific feature value reduces the model output, while positive ones increase the model output.

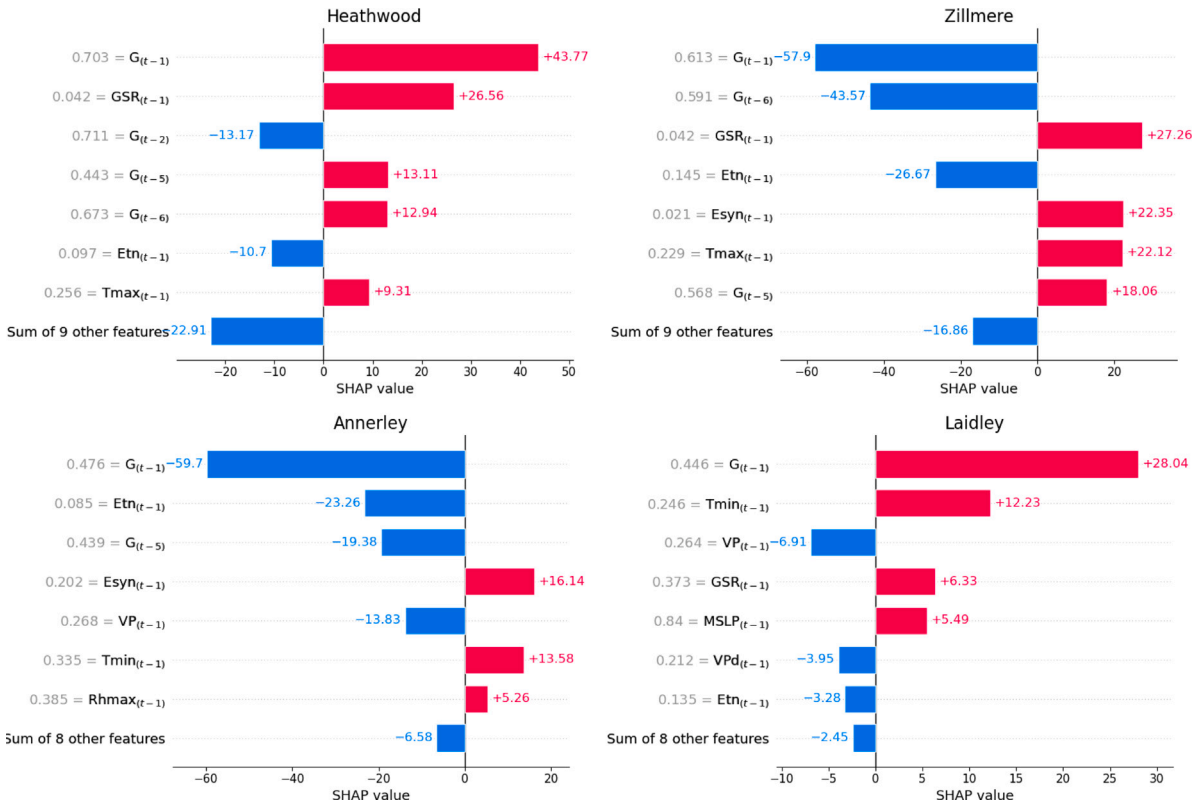


Fig. 20. Local explanation bar plot of FNET model showing the contribution of important features for instance 50 (*i.e.* 50th prediction from FNET model).

Table 8
The performance of the Deep Hybrid Fused Network (FNET) model vs. LSTMCNN, DNN, BILSTM, MLP, KRR, GPR and MARS models using the Willmott's Index (E_{WI}), Nash–Sutcliffe Coefficient (E_{NS}), and the Legates & McCabe's (E_{LM}) Index of Agreement. Note that the best model is boldfaced (blue).

Sub-Station	Predictive Model	Model Performance Metrics		
		E_{WI}	E_{NS}	E_{LM}
Annerley	FNET	**0.919**	**0.889**	**0.676**
	BILSTM	0.914	0.868	0.653
	LSTMCNN	0.910	0.872	0.653
	DNN	0.915	0.862	0.641
	MLP	0.884	0.814	0.586
	KRR	0.827	0.732	0.511
	GPR	0.900	0.833	0.608
	MARS	0.819	0.765	0.521
Heathwood	FNET	**0.889**	**0.760**	**0.517**
	BILSTM	0.879	0.732	0.492
	LSTMCNN	0.876	0.721	0.461
	DNN	0.875	0.722	0.461
	MLP	0.870	0.709	0.442
	KRR	0.869	0.713	0.443
	GPR	0.873	0.700	0.396
	MARS	0.878	0.718	0.435
Laidley	FNET	**0.891**	**0.859**	**0.603**
	BILSTM	0.885	0.828	0.556
	LSTMCNN	0.891	0.832	0.558
	DNN	0.854	0.820	0.549
	MLP	0.850	0.804	0.536
	KRR	0.806	0.764	0.483
	GPR	0.863	0.810	0.546
	MARS	0.879	0.813	0.545
Zillmere	FNET	0.878	**0.821**	**0.580**
	BILSTM	0.872	0.807	0.565
	LSTMCNN	**0.884**	0.807	0.570
	DNN	0.863	0.794	0.551
	MLP	0.798	0.724	0.472
	KRR	0.822	0.754	0.509
	GPR	0.862	0.804	0.560
	MARS	0.869	0.804	0.560

predictive power of $Etn_{(t-1)}$ shifts, resulting in varying levels of SHAP values across different points in the plot. In conclusion, the SHAP dependence plot underscores the synergistic relationship between $Etn_{(t-1)}$ and $G_{(t-1)}$. While $Etn_{(t-1)}$ generally contributes positively to the model's predictions, its impact is magnified in the presence of higher $G_{(t-1)}$ values. This interaction points to a more intricate and dynamic interplay between these two features, suggesting that both need to be considered together for a fuller understanding of the model's behaviour and predictions.

4.1.3. Computational resource requirements

The computational time of a prediction model is critical for utility companies, especially in scenarios involving online training. In such cases, daily electricity demand observations are continually incorporated into the training dataset for model retraining, making computational time a key factor. The time required for electricity demand prediction is influenced by factors such as the length of the moving window, the number of predictors, and, most significantly, the choice of prediction model. Table 13 compares the computation times of the proposed FNET model with seven benchmark models. The results indicate that the proposed model is less computationally efficient than the others. However, once the model is trained, it remains operational for an extended period. Additionally, the testing time is under one minute, making the proposed model suitable for practical applications. The simulations were performed on an Intel® Core™ i9 10th Generation processor, operating at 3.8 GHz with 32 GB of memory.

5. Conclusions and future research directions

Based on the historical electricity demand (G) and a set of local climate data for several substations in Queensland, Australia, a deep learning-based hybrid Deeply Fused Network (*i.e.*, the FNET model) has been proposed and evaluated its efficacy for daily electricity demand (G) point-based as well as confidence interval predictions. Using different statistical evaluation methods, the proposed FNET model was compared with BILSTM, LSTMCNN, DNN, MARS, MLP, KRR, and GPR models to determine its ability in predicting the daily G. According to the results, the proposed FNET model achieved high accuracy among all compared models. The main reason behind this is that the FNET model has high ability in capturing the non-linearity of electricity demand, local climate data and the long-term temporal dependencies between the data points. The other models simply could not match the predictive power of FNET. Furthermore, by employing SHAP analysis, this study delved into the inner workings of the black-box machine learning and deep learning models. This method also illuminated the intricate relationships between variables and their impact on model predictions. The results underscored the pivotal role of historical load values (G) and evapotranspiration (Etn) in shaping electricity demand prediction.

Based on contributions of this study, we aver that there may be significant advantages in adopting the proposed FNET model by current energy industries. The debates in the energy sector are emphasizing a need for decarbonization of the global economy [90–92]. Therefore, a greater proportion of renewable energies is becoming the norm in future electricity supply systems. Energy usage in buildings as well as emissions from vehicles in particular are showing the most significant potential in cost-effective emissions reductions. For energy use in buildings, the adoption of the proposed FNET model for electricity demand management and including key predictor variables such as power consumption by building appliances is a crucial factor that can be included in re-training the proposed FNET model. Likewise, for the transport sector, the emissions reductions can be met effectively through a promotion of electric vehicles (EVs) and utilizing solar (or other forms of renewable energies) for EV charging [93,94]. In order to create a low carbon roadmap and future a carbon neutral pathway of the building sector and especially tackling the carbon emission mitigation in building operations, rooftop solar systems, and large-scale solar farms supporting both the energy requirements in buildings as well as that of the transport sector, could a potential solution. These have been clearly outlined in recent reviews 2024challenges where synergizing technical innovation, developing advanced building technologies and renewable energy solutions have already been outlined. Therefore, the proposed FNET model may be a contributory automation technology further investigated for modelling energy efficiency in buildings, predicting demand and supply of solar (or other renewables) and including weather variables for short-term and climate variables for long-term demand modelling.

In the proposed FNET model, uncertainty values associated with G can be addressed statistically by generating interval predictions that take into account the variability of data features. In order to evaluate the nature of electricity supply mix, requirements for installed storage capacities, or financial planning of energy prices or system costs, it is essential to gain a better understanding of these predicted uncertainties in electricity demand patterns. Furthermore, the proposed FNET model offers an indication of the extent to which G values are underestimated or overestimated, which can be extremely useful when scheduling energy supply reserves, implementing energy policy, managing operational demands of the energy sector, etc. The model uses a combination of machine learning algorithms to identify and quantify the discrepancies between the observed and predicted energy consumption. This information can be used to inform decisions about energy supply, policy, and operational management, helping to ensure

Table 9
The values of the improvement percentages λ of the proposed and benchmark models over the testing modelling phase. λ_{RMSE} indicates the Root Mean Square Error, λ_{KGE} indicates the Kling Gupta Efficiency, and λ_{APB} indicates the Absolute Percentage Bias.

Predictive models	Annerley			Heathwood			Laidley			Zillmere		
	λ_{RMSE}	λ_{APB}	λ_{KGE}	λ_{RMSE}	λ_{APB}	λ_{KGE}	λ_{RMSE}	λ_{APB}	λ_{KGE}	λ_{RMSE}	λ_{APB}	λ_{KGE}
BILSTM	8.89	7.24	3.24	4.73	5.28	9.54	10.82	11.90	5.31	3.63	3.55	4.77
LSTMCNN	7.25	7.23	7.82	8.13	11.71	1.68	10.22	11.29	6.34	4.22	2.49	3.71
DNN	11.35	10.81	3.10	7.90	11.67	2.78	12.98	13.52	7.41	7.16	6.94	4.39
MLP	29.35	27.91	3.79	11.02	15.56	3.56	17.54	16.93	6.34	24.25	25.69	11.43
KRR	55.13	51.13	8.38	10.55	15.29	9.08	29.18	30.29	15.97	17.37	16.93	11.61
GPR	22.44	21.21	12.17	15.39	25.18	1.37	15.69	14.22	4.57	4.55	4.74	5.19
MARS	45.24	47.92	12.72	10.66	16.97	2.15	15.92	14.62	5.24	4.70	4.87	2.55

Table 10
Prediction modelling evaluation based on Diebold–Mariano (DM) Harvey–Leybourne–Newbold (HLN) over the testing phase. For the case of positive results, it indicates that rows superior results to the column. However, if it is negative, then otherwise. Boldfaced blue indicates the best results.

a)

	FNET	BILSTM	LSTMCNN	DNN	MLP	KRR	GPR	MARS
FNET		**2.1264**	**2.9547**	**3.4756**	**5.3419**	**6.4246**	**4.109**	**3.8052**
BILSTM			2.8684	3.4618	6.8822	5.0853	3.4455	2.3604
LSTMCNN				1.5211	5.6053	3.8864	3.2022	1.7243
DNN					5.3677	3.6333	2.9941	1.5141
MLP						−0.4155	−0.9883	−1.7821
KRR							−0.6437	−1.4779
GPR								1.8058

b)

	FNET	**BILSTM**	**LSTMCNN**	**DNN**	**MLP**	**KRR**	**GPR**	**MARS**
FNET		**2.1602**	**3.0017**	**3.5309**	**5.4269**	**6.5268**	**4.1744**	**3.8657**
BILSTM			2.9141	3.5169	6.9917	5.1662	3.5004	2.3979
LSTMCNN				1.5453	5.6945	3.9482	3.2531	1.7518
DNN					5.4531	3.6911	3.0418	1.5382
MLP						−0.4221	−1.0041	−1.8105
KRR							−0.6539	−1.5014
GPR								−1.8345

Table 11
The attained prediction results of 95% Probabilistic confidence with respect to prediction interval coverage probability ($PICP$), mean prediction interval width ($MPIW$) and F index for four substations. F index is defined as the weighted harmonic average of $PICP$ and $1/MPIW$ and evaluates the quality of interval prediction. Boldfaced blue indicates the best modelling results.

Predictive models	Annerley			Heathwood			Laidley			Zillmere		
	$PICP$	$MPIW$	$F \times 10^{-2}$	$PICP$	$MPIW$	$F \times 10^{-2}$	$PICP$	$MPIW$	$F \times 10^{-2}$	$PICP$	$MPIW$	$F \times 10^{-2}$
FNET	95.07	**57.97**	**3.449**	95.05	**256.63**	**0.779**	95.07	**53.42**	**3.74**	94.79	**154.58**	**1.294**
BILSTM	95.07	64.92	3.080	95.05	263.14	0.760	95.07	56.21	3.56	95.07	155.69	1.285
LSTMCNN	94.52	64.90	3.081	95.05	258.88	0.773	95.07	54.03	3.70	95.07	158.67	1.260
DNN	95.34	67.44	2.965	95.33	263.10	0.760	95.07	60.40	3.31	94.79	170.18	1.175
MLP	94.79	82.83	2.414	95.05	270.26	0.740	95.07	63.64	3.14	95.07	190.76	1.048
KRR	95.07	101.24	1.975	95.05	267.77	0.747	95.07	71.66	2.79	95.07	187.38	1.067
GPR	95.07	70.65	2.830	95.05	260.40	0.768	95.07	63.97	3.13	94.79	155.09	1.289
MARS	95.07	94.06	2.126	94.78	264.81	0.755	95.07	60.57	3.30	94.79	158.20	1.264

that the energy sector can meet its demand in a cost-effective and reliable manner. By pairing machine learning algorithms with data from sources such as weather forecasts, population estimates, and energy market trends, the model can accurately predict energy consumption and identify areas where energy supply and demand are mismatched. The obtained information could be utilized as essential input for decisions regarding energy supply, policy, and operational management, thereby helping to ensure the reliability and cost-effectiveness of energy supplies. Therefore, improved demand forecasting can lead to better resource allocation, reduced costs, and enhanced grid stability.

By using the Deeply Fusion Network (FNET) modelling approaches presented in this paper, decision-makers can also gain clearer understanding of the future G required for integrating electricity demand and renewable energy supply, as well as associated uncertainty factors, in a highly stochastic environment, enabling more informed business decisions in terms of capacity and quality. Future research work may focus on the effect of integrated human behaviour on point and confidence interval predictions, for example the effects of social gatherings and seasonal effects on the prediction of electricity demand and what changes may be required in the proposed FNET model in order to account for these influences. Furthermore, evapotranspiration (Etn), evaporation (Esyn) and vapour pressure (VP) for Annerley and Laidley had a major influence on FNET model output, but not for Heathwood and Zillmere (see Fig. 19). Despite the fact that the exact reason for this is not clear from the present study, we note the significant effect of evapotranspiration on electricity demand in this area could inform

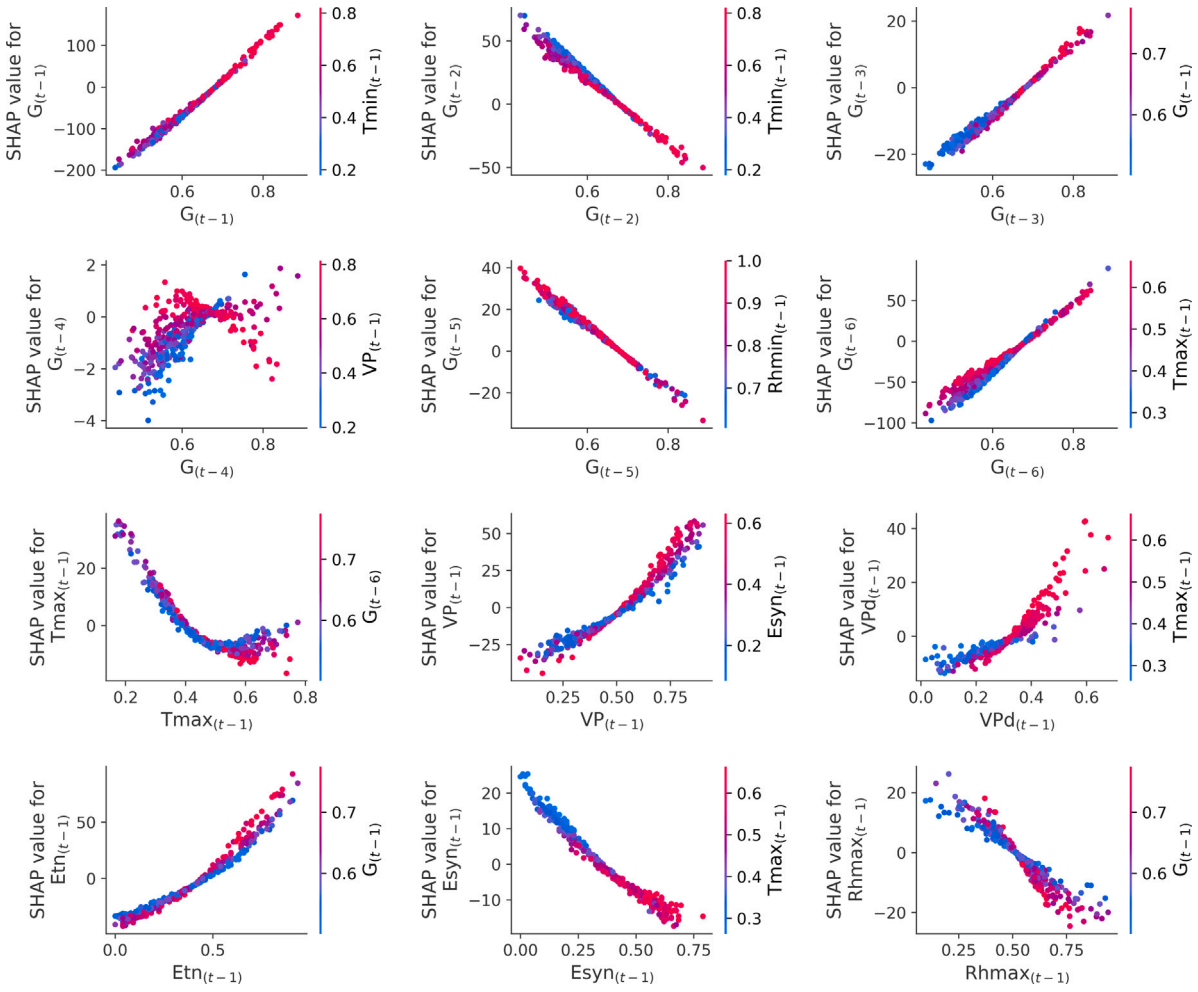


Fig. 21. SHAP dependence plot. Interaction effects of predictors for G prediction for Zillmere sub-station.

energy policies and strategies. The reasons for this are that differential effects of climate change on average and peak demand for heating and cooling have already been noted, for example, across the contiguous USA [95]. In one study [96], researchers found that humidity plays a crucial role in predicting summer electricity demand. According to their study, air temperature factor was necessary but not sufficient to characterize residential space cooling demand during summer months, but humidity levels played a critical role in capturing true heat sensations. The use of air conditioning may therefore be affected by such a sensation, for example, in Annerley and Laidley where Etn and VP were pivotal indicators. Therefore, not taking humidity into account when modelling electricity demand can lead to underestimation of climate sensitivity and have an impact on key decisions. Further studies are warranted to specifically examine why evaporation (and vapour pressure) had a significant impact on electricity demand while solar radiation and maximum temperature did not. As a result, we acknowledge these limitations, so a future study could provide insights that could be utilized by utilities and market regulators to help them make informed decisions in these regions where evaporation (and vapour pressure) are significant determinants of electricity demand.

For future studies related to expanding the practicality of the model and methods presented, one may also argue that the proposed FNET model should be improved with more diverse input from local renewable energy platforms to create a more dynamic and responsive model for predictive modelling in a mixed grid. Thus, the FNET model can offer significant scientific evidence to help energy market workers achieve high-performance quality with sound energy policy decisions. Likewise, exploring the use of other climate variables, extending the FNET model to other geographical regions and incorporating additional machine learning techniques, would provide valuable insights for further advancing the field of electricity demand prediction.

In this study, we have used 11 different climate-based predictors to build the FNET model. However, the model's effectiveness might heavily rely on the quality and availability of climate data especially for other regions not tested in this study. Therefore, addressing the potential data scarcity or variability in different regions in future studies would strengthen the application of the study. This could include capturing wider satellite data for short-term demand predictions, particularly, using Himawari satellite variables (at 10-minute scales), medium-scale (hourly or daily) variables from the European Centre for Medium Range Weather Forecasting (ECMWF) and other ground measurement sites where available. Another limitation of this study is the restricted testing of the proposed FNET model to specifically 4 study sites (i.e., Annerley, Heathwood, Laidley and Zillmere). Therefore, testing the model on additional datasets from various geographical locations could demonstrate its generalizability and robustness across different energy markets. In this study, we have already presented comparisons of FNET model with 7 different models. However, given the rapidly evolving area of artificial intelligence, in future studies, one may test the proposed FNET with several models, including newer or alternative models for a more comprehensive benchmark.

The proposed FNET model uses residual bootstrapping for uncertainty estimation, however exploring other techniques might provide additional insights into model confidence and reliability. Among these methods are jackknife resampling, in which one observation is systematically left out and the model is calculated every time, Bayesian

Table 12
The attained prediction results of the Probabilistic for 95% confidence based on the Winkler score (WS) and the average relative interval length ($ARIL$) for the modelled substations. Boldfaced blue indicate the best results.

Study site	Predictive model	Model Performance Metrics	
		WS	ARIL
Annerley	FNET	**68.251**	**0.174**
	BILSTM	75.246	0.195
	LSTMCNN	74.540	0.195
	DNN	77.278	0.203
	MLP	92.966	0.249
	KRR	112.162	0.305
	GPR	84.084	0.213
	MARS	103.254	0.283
Heathwood	FNET	**307.252**	**0.378**
	BILSTM	312.481	0.387
	LSTMCNN	308.814	0.381
	DNN	311.791	0.387
	MLP	310.630	0.398
	KRR	312.499	0.394
	GPR	309.803	0.383
	MARS	309.732	0.390
Laidley	FNET	**58.103**	**0.054**
	BILSTM	61.553	0.067
	LSTMCNN	59.321	0.059
	DNN	64.916	0.057
	MLP	69.660	0.062
	KRR	77.522	0.061
	GPR	70.506	0.058
	MARS	65.386	0.080
Zillmere	FNET	**173.923**	**0.227**
	BILSTM	174.667	0.228
	LSTMCNN	176.950	0.233
	DNN	184.750	0.249
	MLP	212.497	0.280
	KRR	202.654	0.275
	GPR	178.459	0.229
	MARS	181.202	0.232

Table 13
Average of computation time.

Model	Construction time (Training and Validation)	Testing
FNET	89 min	54 s
BILSTM	17 min	42 s
LSTMCNN	18 min	44 s
DNN	10 min	17 s
MLP	8 min	14 s
KRR	7 min	14 s
GPR	15 min	25 s
MARS	8 min	18 s

methods with a probabilistic framework for estimating uncertainty, and ensemble methods in which multiple models (ensembles) are generated by training on different subsets of data or with a variety of random initializations, in order to estimate uncertainties based on a spread of predictions. It is also possible to estimate the conditional quantiles of the response variable instead of estimating a single value. Other potential candidates include conformal prediction where a non-parametric approach is used to estimate uncertainty, variance estimates from gradient boosting, and Bayesian Neural Networks that can introduce uncertainty into the model weights themselves by placing prior distributions on the weights are useful candidates for future tests on uncertainty estimation of the proposed FNET model. Finally, from a practical point of view, the integration of CNN and BILSTM may be practically difficult although study has demonstrated its efficacy for 4 of the study sites in Queensland. To address this limitation, we need to further test the integrated CNN-BILSTM models for a wider range if study sites, to fully ascertain their practical deployment.

CRediT authorship contribution statement

Sujan Ghimire: Writing – original draft, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mohanad S. AL-Musaylh:** Writing – review & editing, Visualization, Validation. **Thong Nguyen-Huy:** Writing – review & editing, Visualization, Validation, Investigation. **Ravinesh C. Deo:** Writing – review & editing, Visualization, Supervision, Resources, Project administration. **Rajendra Acharya:** Writing – review & editing. **David Casillas-Pérez:** Writing – review & editing. **Zaher Mundher Yaseen:** Writing – review & editing. **Sancho Salcedo-Sanz:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data were acquired from ENERGEX (https://www.energex.com.au).

References

[1] Assembly UNG, et al. Resolution adopted by the general assembly on 25 september 2015. 2015, Washington: United Nations.
[2] Al-Musaylh MS, Deo RC, Adamowski JF, Li Y. Short-term electricity demand forecasting with MARS, SVR and ARIMA models using aggregated demand data in Queensland, Australia. Adv Eng Inform 2018;35:1–16.
[3] Al-Musaylh MS, Deo RC, Adamowski JF, Li Y. Short-term electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF reanalysis atmospheric predictors in southeast Queensland, Australia. Renew Sustain Energy Rev 2019;113:109293.
[4] Balalla DT, Nguyen-Huy T, Deo R. MARS model for prediction of short-and long-term global solar radiation. In: Predictive modelling for energy management and power systems engineering. Elsevier; 2021, p. 391–436.
[5] Mohanad SA-M, Ravinesh CD, Yan L. Particle swarm optimized–support vector regression hybrid model for daily horizon electricity demand forecasting using climate dataset. In: E3S web of conferences. Vol. 64, EDP Sciences; 2018, p. 08001.
[6] Al-Musaylh MS, Deo RC, Li Y, Adamowski JF. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. Appl Energy 2018;217:422–39.
[7] Al-Musaylh MS, Deo RC, Li Y. Electrical energy demand forecasting model development and evaluation with maximum overlap discrete wavelet transform-online sequential extreme learning machines algorithms. Energies 2020;13(9):2307.
[8] Ghimire S, Bhandari B, Casillas-Pérez D, Deo RC, Salcedo-Sanz S. Hybrid deep CNN-SVR algorithm for solar radiation prediction problems in Queensland, Australia. Eng Appl Artif Intell 2022;112:104860.
[9] Jamei M, Ahmadianfar I, Olumegbon IA, Karbasi M, Asadi A. On the assessment of specific heat capacity of nanofluids for solar energy applications: Application of Gaussian process regression (GPR) approach. J Energy Storage 2021;33:102067.
[10] Adewuyi SA, Aina S, Oluwaranti AI. A deep learning model for electricity demand forecasting based on a tropical data. Appl Comput Sci 2020;16(1).
[11] Zhang M, Li J, Li Y, Xu R. Deep learning for short-term voltage stability assessment of power systems. IEEE Access 2021;9:29711–8.
[12] Wang K, Qi X, Liu H. Photovoltaic power forecasting based LSTM-convolutional network. Energy 2019;189:116225.
[13] Qu J, Qian Z, Pei Y. Day-ahead hourly photovoltaic power forecasting using attention-based CNN-LSTM neural network embedded with multiple relevant and target variables prediction pattern. Energy 2021;232:120996.
[14] Hafeez G, Alimgeer KS, Khan I. Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid. Appl Energy 2020;269:114915.

[15] del Real AJ, Dorado F, Durán J. Energy demand forecasting using deep learning: applications for the french grid. Energies 2020;13(9):2242.

[16] Zhang L, Liu P, Zhao L, Wang G, Zhang W, Liu J. Air quality predictions with a semi-supervised bidirectional LSTM neural network. Atmos Pollut Res 2021;12(1):328–39.

[17] Dolatabadi A, Abdeltawab H, Mohamed YA-RI. Hybrid deep learning-based model for wind speed forecasting based on DWPT and bidirectional LSTM network. IEEE Access 2020;8:229219–32.

[18] Cheng H, Xie Z, Wu L, Yu Z, Li R. Data prediction model in wireless sensor networks based on bidirectional LSTM. EURASIP J Wireless Commun Networking 2019;2019(1):1–12.

[19] Wang S, Wang X, Wang S, Wang D. Bi-directional long short-term memory method based on attention mechanism and rolling update for short-term load forecasting. Int J Electr Power Energy Syst 2019;109:470–9.

[20] Shao Z, Chao F, Yang S-L, Zhou K-L. A review of the decomposition methodology for extracting and identifying the fluctuation characteristics in electricity demand forecasting. Renew Sustain Energy Rev 2017;75:123–36.

[21] Rafi SH, Deeba SR, Hossain E, et al. A short-term load forecasting method using integrated CNN and LSTM network. IEEE Access 2021;9:32436–48.

[22] Zhang J, Wei Y-M, Li D, Tan Z, Zhou J. Short term electricity load forecasting using a hybrid model. Energy 2018;158:774–81.

[23] Li C, Chen Z, Liu J, Li D, Gao X, Di F, et al. Power load forecasting based on the combined model of LSTM and XGBoost. In: Proceedings of the 2019 the international conference on pattern recognition and artificial intelligence. 2019, p. 46–51.

[24] Wu F, Cattani C, Song W, Zio E. Fractional ARIMA with an improved cuckoo search optimization for the efficient short-term power load forecasting. Alex Eng J 2020;59(5):3111–8.

[25] Wang J, Wei Z, Zhang T, Zeng W. Deeply-fused nets. 2016, arXiv preprint arXiv:1605.07716.

[26] Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z. Deeply-supervised nets. In: Artificial intelligence and statistics. Pmlr; 2015, p. 562–70.

[27] Srivastava RK, Greff K, Schmidhuber J. Training very deep networks. Adv Neural Inf Process Syst 2015;28.

[28] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.

[29] Zhao L, Wang J, Li X, Tu Z, Zeng W. On the connection of deep fusion to ensembling. 2016, arXiv preprint arXiv:1611.07718.

[30] Li J, Li X, Jing X. Deeply-fused human motion recognition network in radar for in-home monitoring. In: 2019 IEEE symposium series on computational intelligence. SSCI, IEEE; 2019, p. 584–7.

[31] Ghimire S, Nguyen-Huy T, Deo RC, Casillas-Perez D, Salcedo-Sanz S. Efficient daily solar radiation prediction with deep learning 4-phase convolutional neural network, dual stage stacked regression and support vector machine CNN-REGST hybrid model. Sustain Mater Technol 2022;32:e00429.

[32] Ghimire S, Nguyen-Huy T, Prasad R, Deo RC, Casillas-Pérez D, Salcedo-Sanz S, et al. Hybrid convolutional neural network-multilayer perceptron model for solar radiation prediction. Cogn Comput 2022;1–27.

[33] Ghimire S, Deo RC, Wang H, Al-Musaylh MS, Casillas-Pérez D, Salcedo-Sanz S. Stacked LSTM sequence-to-sequence autoencoder with feature selection for daily solar radiation prediction: A review and new modeling results. Energies 2022;15(3):1061.

[34] Ghimire S, Yaseen ZM, Farooque AA, Deo RC, Zhang J, Tao X. Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. Sci Rep 2021;11(1):1–26.

[35] Ghimire S. Predictive modelling of global solar radiation with artificial intelligence approaches using MODIS satellites and atmospheric reanalysis data for Australia (Ph.D. thesis), University of Southern Queensland; 2019.

[36] Ghimire S, Deo RC, Casillas-Pérez D, Salcedo-Sanz S. Improved complete ensemble empirical mode decomposition with adaptive noise deep residual model for short-term multi-step solar radiation prediction. Renew Energy 2022;190:408–24.

[37] Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional neural networks. Pattern Recognit 2018;77:354–77.

[38] Ghimire S, Nguyen-Huy T, AL-Musaylh MS, Deo RC, Casillas-Pérez D, Salcedo-Sanz S. A novel approach based on integration of convolutional neural networks and echo state network for daily electricity demand prediction. Energy 2023;127430. http://dx.doi.org/10.1016/j.energy.2023.127430, URL https://www.sciencedirect.com/science/article/pii/S0360544223008241.

[39] Pedamonti D. Comparison of non-linear activation functions for deep neural networks on MNIST classification task. 2018, arXiv preprint arXiv:1804.02763.

[40] Nwankpa C, Ijomah W, Gachagan A, Marshall S. Activation functions: Comparison of trends in practice and research for deep learning. 2018, arXiv preprint arXiv:1811.03378.

[41] Klambauer G, Unterthiner T, Mayr A, Hochreiter S. Self-normalizing neural networks. Adv Neural Inf Process Syst 2017;30.

[42] Ghimire S, Nguyen-Huy T, Prasad R, Deo RC, Casillas-Pérez D, Salcedo-Sanz S, et al. Hybrid convolutional neural network-multilayer perceptron model for solar radiation prediction. Cogn Comput 2022;1–27.

[43] Ghimire S, Deo RC, Casillas-Pérez D, Salcedo-Sanz S, Sharma E, Ali M. Deep learning CNN-LSTM-MLP hybrid fusion model for feature optimizations and daily solar radiation prediction. Measurement 2022;111759.

[44] Ghimire S, Deo RC, Casillas-Pérez D, Salcedo-Sanz S. Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms. Appl Energy 2022;316:119063.

[45] Mansouri I, Ozbakkaloglu T, Kisi O, Xie T. Predicting behavior of FRP-confined concrete using neuro fuzzy, neural network, multivariate adaptive regression splines and M5 model tree techniques. Mater Struct 2016;49(10):4319–34.

[46] Friedman JH. Multivariate adaptive regression splines. Ann Stat 1991;19(1):1–67.

[47] Meyer H, Kühnlein M, Appelhans T, Nauss T. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. Atmos Res 2016;169:424–33.

[48] An J, Yin F, Wu M, She J, Chen X. Multisource wind speed fusion method for short-term wind power prediction. IEEE Trans Ind Inf 2020;17(9):5927–37.

[49] Ahmed AM, Sharma E, Jui SJJ, Deo RC, Nguyen-Huy T, Ali M. Kernel ridge regression hybrid method for wheat yield prediction with satellite-derived predictors. Remote Sens 2022;14(5):1136.

[50] Esfe MH, Motallebi SM, Toghraie D. Investigation of thermophysical properties of MWCNT-MgO (1: 1)/10W40 hybrid nanofluid by focusing on the rheological behavior: Sensitivity analysis and price-performance investigation. Powder Technol 2022;405:117472.

[51] Esfe MH. The dual behavior of the dynamic viscosity of multiwalled carbon nanotubes–Al2O3 (3: 7)/ethylene glycol hybrid nanofluids: an experimental study. Eur Phys J Plus 2022;137(6):1–13.

[52] Prasad SMM, Nguyen-Huy T, Deo R. Support vector machine model for multistep wind speed forecasting. In: Predictive modelling for energy management and power systems engineering. Elsevier; 2021, p. 335–89.

[53] Shahsavar A, Jamei M, Karbasi M. Experimental evaluation and development of predictive models for rheological behavior of aqueous Fe3O4 ferrofluid in the presence of an external magnetic field by introducing a novel grid optimization based-Kernel ridge regression supported by sensitivity analysis. Powder Technol 2021;393:1–11.

[54] Sheng H, Xiao J, Cheng Y, Ni Q, Wang S. Short-term solar power forecasting based on weighted Gaussian process regression. IEEE Trans Ind Electron 2017;65(1):300–8.

[55] Castillo-Botón C, Casillas-Pérez D, Casanova-Mateo C, Ghimire S, Cerro-Prada E, Gutierrez P, et al. Machine learning regression and classification methods for fog events prediction. Atmos Res 2022;272:106157.

[56] Rohani A, Taki M, Abdollahpour M. A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (part: I). Renew Energy 2018;115:411–22.

[57] Ghimire S, Deo RC, Downs NJ, Raj N. Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. J Clean Prod 2019;216:288–310.

[58] de Almeida MR, Correa DN, Rocha WF, Scafi FJ, Poppi RJ. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. Microchem J 2013;109:170–7.

[59] Pullanagari RR, Li M. Uncertainty assessment for firmness and total soluble solids of sweet cherries using hyperspectral imaging and multivariate statistics. J Food Eng 2021;289:110177.

[60] Hwang E. Prediction intervals of the COVID-19 cases by HAR models with growth rates and vaccination rates in top eight affected countries: Bootstrap improvement. Chaos Solitons Fractals 2022;155:111789.

[61] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.

[62] Wu C, Chau K-W. Data-driven models for monthly streamflow time series prediction. Eng Appl Artif Intell 2010;23(8):1350–67.

[63] Durdu ÖF. Application of linear stochastic models for drought forecasting in the Büyük Menderes river basin, western Turkey. Stoch Environ Res Risk Assess 2010;24(8):1145–62.

[64] Sivakumar B, Woldemeskel FM, Puente CE. Nonlinear analysis of rainfall variability in Australia. Stoch Environ Res Risk Assess 2014;28(1):17–27.

[65] Sivakumar B. Chaos theory in hydrology: important issues and interpretations. J Hydrol 2000;227(1–4):1–20.

[66] Jeffrey SJ, Carter JO, Moodie KB, Beswick AR. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ Model Softw 2001;16(4):309–30.

[67] Sanner MF, et al. Python: a programming language for software integration and development. J Mol Graph Model 1999;17(1):57–61.

[68] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. {TensorFlow}: a system for {large-scale} machine learning. In: 12th USeNIX symposium on operating systems design and implementation (OSDI 16). 2016, p. 265–83.

[69] Ketkar N. Introduction to keras. In: Deep learning with python. Springer; 2017, p. 97–111.

[70] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[71] Brownlee J. Gentle introduction to the adam optimization algorithm for deep learning. Machine Learning Mastery 2017;3.

[72] Bergstra J, Yamins D, Cox DD, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In: Proceedings of the 12th python in science conference. Vol. 13, Citeseer; 2013, p. 20.
[73] Han Y, Lee K. Convolutional neural network with multiple-width frequency-delta data augmentation for acoustic scene classification. In: IEEE AASP challenge on detection and classification of acoustic scenes and events. 2016.
[74] Wang J, Cao J. Data-driven S-wave velocity prediction method via a deep-learning-based deep convolutional gated recurrent unit fusion network. Geophysics 2021;86(6):M185–96.
[75] Deo RC, Ghimire S, Downs NJ, Raj N. Optimization of windspeed prediction using an artificial neural network compared with a genetic programming model. In: Research anthology on multi-industry uses of genetic programming and algorithms. IGI Global; 2021, p. 116–47.
[76] Marsaglia G, Tsang WW, Wang J. Evaluating Kolmogorov's distribution. J Stat Softw 2003;8:1–4.
[77] Alothman T, Alsaif SA, Alfakhri A, Alfadda A. Performance assessment of 25 global horizontal irradiance clear sky models in riyadh. In: 2022 IEEE international conference on environment and electrical engineering and 2022 IEEE industrial and commercial power systems europe (EEEIC/I&CPS europe). IEEE; 2022, p. 1–6.
[78] Li M-F, Tang X-P, Wu W, Liu H-B. General models for estimating daily global solar radiation for different solar radiation zones in mainland China. Energy Convers Manag 2013;70:139–48.
[79] Willmott CJ. On the evaluation of model performance in physical geography. In: Spatial statistics and models. Springer; 1984, p. 443–60.
[80] Legates DR, McCabe Jr. GJ. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour Res 1999;35(1):233–41.
[81] Dawson CW, Abrahart RJ, See LM. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ Model Softw 2007;22(7):1034–52.
[82] Gueymard CA. A review of validation methodologies and statistical performance indicators for modeled solar radiation data: Towards a better bankability of solar projects. Renew Sustain Energy Rev 2014;39:1024–34.
[83] Despotovic M, Nedic V, Despotovic D, Cvetanovic S. Review and statistical analysis of different global solar radiation sunshine models. Renew Sustain Energy Rev 2015;52:1869–80.
[84] Mariano RS, Preve D. Statistical tests for multiple forecast comparison. J Econometrics 2012;169(1):123–30.
[85] Liu H, Mi X, Li Y. Smart deep learning based wind speed prediction model using wavelet packet decomposition, convolutional neural network and convolutional long short term memory network. Energy Convers Manage 2018;166:120–31.
[86] Bottieau J, Wang Y, De Grève Z, Vallée F, Toubeau J-F. Interpretable transformer model for capturing regime switching effects of real-time electricity prices. IEEE Trans Power Syst 2022.
[87] Dogulu N, López López P, Solomatine D, Weerts A, Shrestha D. Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrol Earth Syst Sci 2015;19(7):3181–201.
[88] Ni Q, Zhuang S, Sheng H, Kang G, Xiao J. An ensemble prediction intervals approach for short-term PV power forecasting. Sol Energy 2017;155:1072–83.
[89] Singla P, Duhan M, Saroha S. Review of different error metrics: A case of solar forecasting. AIUB J Sci Eng (AJSE) 2021;20(4):158–65.
[90] Yan R, Ma M, Zhou N, Feng W, Xiang X, Mao C. Towards COP27: Decarbonization patterns of residential building in China and India. Appl Energy 2023;352:122003.
[91] Zhang S, Ma M, Zhou N, Yan J, Feng W, Yan R, et al. Estimation of global building stocks by 2070: Unlocking renovation potential. Nexus 2024.
[92] Ma M, Zhou N, Feng W, Yan J. Challenges and opportunities in the global net-zero building sector. Cell Reports Sustain 2024.
[93] Deng Y, Ma M, Zhou N, Ma Z, Yan R, Ma X. Chinas plug-in hybrid electric vehicle transition: An operational carbon perspective. Energy Convers Manage 2024;320:119011.
[94] Yuan H, Ma M, Zhou N, Xie H, Ma Z, Xiang X, et al. Battery electric vehicle charging in China: Energy demand and emissions trends in the 2020s. Appl Energy 2024;365:123153.
[95] Amonkar Y, Doss-Gollin J, Farnham DJ, Modi V, Lall U. Differential effects of climate change on average and peak demand for heating and cooling across the contiguous USA. Commun Earth Environ 2023;4(1):402.
[96] Maia-Silva D, Kumar R, Nateghi R. The critical role of humidity in modeling summer electricity demand across the United States. Nat Commun 2020;11(1):1686.