Advancing Predictive Modeling in Conventional Solar Stills: A Deep Learning Approach Leveraging Data Augmentation and Convolutional Neural Networks.

Hashim H. Migaybil [a,b], Bhushan Gopaluni [b]

[a] Department of Chemical and Materials Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

[b] Department of Chemical and Biological Engineering, Faculty of Applied Science, The University of British Columbia, Vancouver, BC V6T1Z3, Canada

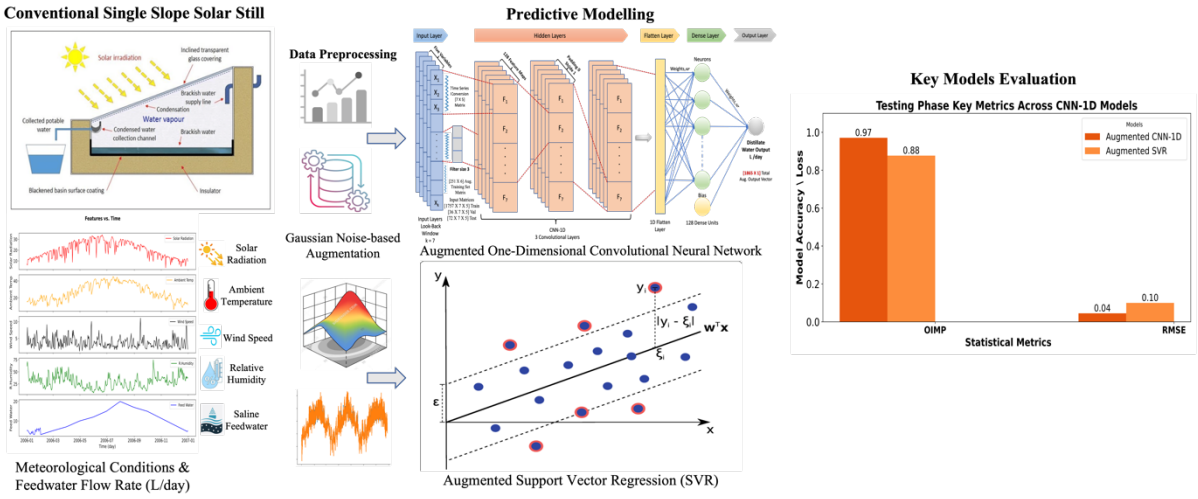
[a] Corresponding author, Hashim H. Migaybil (e-mail: hmigaybil@kau.edu.sa).

Abstract

Accurate forecasting of freshwater productivity from conventional single-slope solar stills is crucial for enhancing operational efficiency and minimizing capital costs. A persistent challenge in this domain is the scarcity of experimental data, which limits the training of reliable predictive models. This study proposes a data-efficient forecasting framework that integrates a one-dimensional convolutional neural network (CNN-1D) with time-series data augmentation. Gaussian noise sampled from $\mathcal{N}(0, 0.01^2)$ was applied exclusively to the training set, generating six augmented samples per instance. Both the augmentation factor (six) and the look-back window (seven days) were selected through systematic optimization, ensuring preservation of temporal dependencies. The CNN-1D architecture comprised three convolutional layers with 128 filters, ReLU activations, a flattening stage, and a dense regression output layer. Hyperparameters—including learning rate, batch size, kernel size, and regularization strength—were fine-tuned using Tree-structured Parzen Estimator (TPE) optimization with a maximum of 50 trials, where the best-performing configuration achieved the lowest loss. Model training employed a feed-forward backpropagation algorithm with 365 daily observations to predict freshwater yield (P_{std}, L/day). Benchmarking against an optimized support vector regression (SVR) model with a radial basis function kernel revealed that the augmented CNN-1D achieved superior performance (RMSE = 0.04, MAE = 0.03, OIMP = 0.97), consistently outperforming both the baseline CNN-1D and the optimized SVR. Residual analyses confirmed its robustness, minimal bias, and strong generalization across unseen data. These findings demonstrate that combining augmentation with hierarchical feature extraction enables a scalable and computationally efficient predictive tool for solar still performance, offering significant potential for sustainable freshwater management in arid and data-constrained regions.

Keywords: Solar still, Convolutional neural networks (CNNs), Data Augmentation, Look-back window, Support vector regression (SVR), Distillate water

Graphical Abstract



1. Introduction

The escalating global water crisis, intrinsically linked to climate shifts and unsustainable resource management practices, poses a significant threat to ecological integrity, food security, and socioeconomic development in the 21st century [1]. With the world's population projected to reach 9.7 billion by 2050, the demand for freshwater is expected to surge by over 55%, exacerbating the existing pressure on already strained water resources [2]. Over 2.2 billion people globally lack access to safely managed drinking water services, while nearly 4 billion experience severe water scarcity for at least one month annually [3]. The available freshwater constitutes only about 2.5% of global water resources, with less than 1% readily accessible for human consumption [4]. This predicament is particularly acute in arid and semi-arid regions, such as substantial portions of the Middle East and North Africa (MENA). In response, these regions are pioneering ambitious initiatives. For instance, Saudi Arabia's NEOM project aims to become a paradigm of sustainable living by developing advanced, large-scale water solutions predominantly powered by renewable energy sources, thereby establishing new benchmarks for efficiency and environmental stewardship in water production [5].

To mitigate the freshwater deficit, various desalination technologies have been developed and widely implemented worldwide. The most established methodologies for large-scale applications include Multi-Stage Flash (MSF) distillation, Multi-Effect Distillation (MED), and Seawater Reverse Osmosis (SWRO) [6]. MSF and MED, as mature thermal processes, have historically exhibited high energy consumption. The specific energy consumption (SEC) for older MSF desalination plants typically ranges from 10 to 16 kWh/m³. In contrast, modern MED plants, particularly those equipped with waste heat recovery systems, consume approximately 2.5 to 5 kWh/m³ of electrical energy or 60 to 70 kWh/m³ of thermal energy [7]. Older thermal facilities can have a considerable carbon footprint, with CO_2 emissions exceeding 10 to 15 kg CO_2 equivalent per cubic meter of water produced [8]. SWRO offers a more energy-efficient alternative, operating at less than 3 kWh/m³ and achieving a 50-70% decrease in energy consumption compared to older methods. However, SWRO encounters challenges such as extensive pre-treatment, issues with membrane fouling, and environmental concerns related to

brine discharge. While innovations like energy recovery devices have enhanced efficiency, global energy demand for desalination continues to rise, underscoring the need for further advancements in sustainable water purification technologies [9] [10].

In this framework, solar distillation, which harnesses solar irradiance to purify water, represents a compelling and environmentally sustainable alternative, particularly in regions abundant in solar resources [11]. Technologies like single-slope solar stills are noteworthy for their operational simplicity, low capital and maintenance costs, and inherent suitability for decentralized, off-grid water production in remote or resource-limited communities [12]. Despite these advantages, a primary obstacle to their widespread adoption is their relatively modest freshwater productivity, which typically ranges from 2 to 5 L/m². day under average climatic conditions [13]. This output, while valuable for specific applications, contrasts sharply with the notable throughput capacities of conventional desalination plants. Therefore, enhancing solar stills' efficiency and predictive management is essential to augmenting their contribution to sustainable freshwater generation, rendering them a more viable component within diverse and resilient water security frameworks.

Researchers have explored various experimental modifications to enhance the performance of solar stills. Hammoodi et al. [14] conducted a study on the performance of pyramid solar stills enhanced with wick and reflective materials under the climatic conditions of Iraq. Incorporating wick materials increased evaporation and heat absorption, resulting in a remarkable 122% productivity boost and achieving a thermal efficiency of 53%. Introducing reflectors also augmented productivity by 170%, although the efficiency remained constant at 48%. Mohammed et al. investigated the use of phase change materials (PCMs) to enhance the productivity of a single-slope solar still in Qena. The study assessed the energy storage behavior of RT42 PCM placed inside the basin area in quantities of 2 kg, 4 kg, and 6 kg to improve performance during both day and night. The findings indicated that the configuration with 4 kg of PCM outperformed the other setups, resulting in productivity increases of approximately 18.6% compared to the 2 kg configuration and 27.7% compared to the 6 kg configuration. Additionally, this configuration achieved an average daily efficiency of 66.7%, representing a 29.7% improvement in freshwater yield over a conventional solar still without PCM. Using PCMs raised the basin water temperature and supported increased water production during nighttime due to enhanced thermal energy storage. The produced water met acceptable quality parameters, making it suitable for domestic applications [15].

Machine learning (ML) integration has introduced new possibilities for modeling, performance prediction, and real-time control of solar stills. Wang et al. developed and evaluated machine learning models to accurately predict the hourly production performance of tubular solar stills using experimental data. The models compared included artificial neural networks (ANN) and random forests (RF), both with and without Bayesian optimization, in addition to traditional multilinear regression. Before optimization, the RF model outperformed the ANN and multilinear regression, achieving a coefficient of determination (R^2) of 0.9758 and a mean absolute percentage error (MAPE) of 5.21%. In contrast, the ANN and multilinear regression recorded R^2 values of 0.9614 and 0.9267, respectively, with MAPE values of 7.697% and 10.911%. Notably, applying Bayesian optimization significantly improved the prediction accuracy of the ANN by 35%. However, the RF model demonstrated superior robustness and less sensitivity to hyperparameter

tuning. Overall, the study concluded that the RF model, owing to its accuracy and stability, is a highly effective and practical tool for forecasting the performance of tubular solar stills [16].
Ghandourah et al. investigated the thermal performance of two solar still designs, utilizing a hybrid artificial intelligence model optimized via the Golden Jackal Optimizer (GJO). The study compared an aluminum-based solar still (ALSS) with a polycarbonate-based solar still (PCSS), which included modified absorber plates featuring air cavities. The proposed ANN-GJO model demonstrated superior performance in predicting overall heat transfer coefficients, energy efficiency, exergy efficiency, and distillate yield compared to conventional ANN, ANN-GA, and ANN-PSO models. Experimental results revealed that ALSS performed exceptionally well, achieving a maximum distillate output of 3.80 L/m²/day and an energy efficiency of 48.8%, in contrast to PCSS, which produced 3.40 L/m²/day and had an energy efficiency of 42.4%. These findings underscore the effectiveness of ANN-GJO in thermal modeling and confirm that the choice of material has a significant impact on solar still productivity [17].

Recent advancements in solar desalination technologies have increasingly focused on hybridization and intelligent modeling approaches to overcome the intrinsic limitations of conventional systems. Shoeibi et al. explored the enhancement of solar still performance by integrating porous media, nano-enhanced phase change materials (NEPCMs), and nano-coated absorbers. Their system utilized anthracite as a porous medium to enhance solar absorption, while paraffin wax infused with CuO and Al_2O_3 nanoparticles (at concentrations of 0.1% and 0.3% by weight) served as the NEPCM to improve thermal energy storage. These NEPCMs were situated within copper pipes above the anthracite layer, which were further coated with CuO-dispersed black paint to enhance thermal conductivity. The study revealed that the NEPCMs with 0.3% CuO and Al_2O_3 increased freshwater productivity by 55.8% and 49.5%, respectively. Additionally, including nano-coated pipes alone contributed to a 5.7% increase in productivity. Moreover, the incorporation of CuO and Al_2O_3 nanoparticles effectively lowered the melting point of the PCM by 2.1 °C and 1.8 °C, respectively, thereby enhancing its thermal responsiveness. The cost per liter (CPL) for distilled water was reduced to $0.10 and $0.104, confirming the economic viability of these enhancements. These findings highlight the significant impact of improved thermal conductivity, enhanced latent heat storage, and optimized absorption on the efficiency and cost-effectiveness of solar stills [18]. Collectively, these studies highlight the potential of combining experimental enhancements and machine learning techniques, mainly when guided by evolutionary optimization algorithms, to maximize freshwater productivity, enhance modeling reliability, and improve operational efficiency in solar still systems.

This body of research underscores the need for advanced predictive modeling techniques to address the shortcomings of conventional solar stills. In this context, the present study introduces a one-dimensional convolutional neural network (CNN 1D) model, augmented with Gaussian noise-based time series data expansion, to enhance forecasting accuracy under limited data conditions. The proposed model, trained on key meteorological variables and feedwater flow, aims to surpass traditional nonlinear regression methods in predictive precision and generalization, thereby facilitating intelligent design and operational planning for solar desalination systems. This research paper highlights multiple novel and significant contributions:

- This research presents a novel integration of Gaussian noise-based data augmentation (jittering) with CNN-1D and SVR models, significantly enhancing prediction accuracy for

freshwater productivity in solar stills. This methodology effectively addresses data scarcity, improving model robustness and generalization compared to conventional techniques.

- Comprehensive hyperparameter tuning resulted in the meticulous development of an optimized CNN-1D model with the 5-$(128)^3$-1 architecture. This carefully structured model effectively captures the spatial and temporal patterns inherent in solar still data, leading to significantly improved prediction accuracy compared to baseline models.
- The study extensively benchmarks the augmented CNN and SVR models against their baseline counterparts, clearly demonstrating the superior performance of the augmented CNN-1D model across several key statistical metrics. This thorough validation underscores the practical advantages and enhances predictive accuracy achieved through data augmentation.
- A pioneering threshold-based classification method was introduced to reliably categorize freshwater outputs based on meaningful productivity thresholds. This technique mitigates the influence of outliers, providing enhanced interpretability and stability compared to traditional quartile-based approaches.
- The research presents practical insights with immediate relevance to real-world solar desalination systems. By illustrating the clear benefits of merging data augmentation with CNN-1D modeling, this study lays the groundwork for cost-effective and reliable forecasting in solar desalination technologies, potentially shaping future developments toward sustainability.

The subsequent sections of this paper are structured as follows. Section 2 outlines the proposed data augmentation framework, along with the equations incorporated into the CNN 1D and SVR models, emphasizing essential elements such as time-series processing, model architecture, and hyperparameter tuning. Section 3 provides a comprehensive description of the implementation procedures, model evaluation, and a comparative analysis of the augmented and baseline models. In Section 4, recent advancements and insights were explored, addressing existing challenges and conducting performance analyses within the field. Finally, this article culminates with a discussion of closing remarks in section 5, exploring potential avenues for future research.

2. Methodology and Framework

In this study, a novel framework for modeling and predicting the freshwater productivity of a conventional single-slope solar still unit was established using a one-dimensional Convolutional Neural Network (CNN-1D). The performance of the baseline CNN-1D model, trained on the original un-augmented dataset, was assessed using various statistical metrics. Then, it was compared to the performance of the augmented CNN-1D model that benefited from augmented datasets through a noise-based data augmentation technique known as jittering. The latter model leveraged the increased diversity and robustness of the augmented dataset, resulting in improved prediction accuracy. The preprocessing step involved transforming the conventional solar still dataset into time series data using a seven-day look-back window. The data was then split into training, validation, and testing sets while maintaining consistent scaling. The augmentation process was then applied to the training data only by introducing random Gaussian noise to the input features, generating multiple variations of each training sample.

Moreover, the Support Vector Regression (SVR) was employed as a benchmark to compare the predictive performance of both the baseline and augmented CNN-1D models. The impact of data augmentation on predictive accuracy was analyzed by evaluating statistical metrics across these models. This study investigates the predictive capabilities of the CNN-1D architecture in estimating freshwater productivity by analyzing time-series data derived from a solar still system. Notably, this investigation deliberately omits a focus on the physical or geometric characteristics of the system.

2.1 Developing Baseline CNN-1D model:

To assess the long-term performance of the conventional single-slope solar still, a total of 365 numerical continuous datasets were manually extracted from previous experimental findings documented in the literature, specifically those conducted by Santos[19]. As shown in Table 1, the compiled datasets encompassed direct measurements of feedwater volumes and daily average distillate production, recorded between February 2006 and August 2007, using a rectangular basin still with an area of 0.976 m² [20]. Data accompanied these datasets, corresponding to daily meteorological conditions, including solar incidents obtained from NASA power webs[21], which served as input variables.

The freshwater productivity dataset was digitized from Figure 1 ('Solar still A' long-term production) in Santos [19] using WebPlotDigitize, with repeated digitization confirming an error margin <2%. Four meteorological drivers (solar irradiance, ambient temperature, wind speed, and relative humidity) were obtained from the NASA POWER database. The daily feedwater flow rate, not explicitly reported by Santos, was estimated using the formula feedwater flow rate = product / daily efficiency. While Santos reported that daily efficiency varies substantially with weather conditions (e.g., a 35% increase in production at higher wind speeds under the same irradiance), reconstructing fluctuating daily efficiency would add unmodeled noise and reduce reproducibility. Instead, a representative average efficiency of 35% was adopted as a conservative assumption, guided by Santos' own discussion of variability. This ensured that feedwater remained linearly proportional to productivity, while enabling the models to capture non-linear meteorological influences on efficiency. The resulting dataset (productivity plus five inputs) is therefore complete, reproducible, and has been validated in prior work [32].

Table 1: Input and output parameters. Summary of meteorological drivers, feedwater flow rate, and freshwater productivity used for modeling the conventional single-slope solar still.

Parameter	Symbol	Unit	Source / Estimation	Role
Solar irradiance	Is	$MJ.m^{-2}.day^{-1}$	NASA POWER database [21]	Input
Ambient temperature	T_{amb}	°C	NASA POWER database [21]	Input
Relative humidity	RH	%	NASA POWER database [21]	Input
Wind speed	V_w	$m.s^{-1}$	NASA POWER database [21]	Input
Feedwater flowrate	P_{st}	$L.day^{-1}$	Estimated from solar still productivity and efficiency [19]	Input
Freshwater productivity	P_{std}	$L.day^{-1}$	(Solar Still A, experimental data) [19]	Output

Although clearness index and day length are recognized determinants of solar system performance, they were not explicitly included as input features in this study. The clearness index is directly derived from solar irradiance, which was already included as a primary driver, and day length is strongly correlated with seasonal irradiance cycles captured in the daily dataset. With only 365 daily samples, introducing additional features would increase dimensionality without improving predictive accuracy and could risk overfitting. Instead, the model focused on five non-redundant inputs that represent the dominant meteorological and operational influences on freshwater output.

A baseline one-dimensional Convolutional Neural Network (CNN-1D) model was developed with feedwater volumes and daily average weather data as inputs. At the same time, the total freshwater production was designated as the output variable (P_{std}, L/day). The earlier study [19] aimed to predict the distillate production of single-basin solar stills, considering various physical properties of both the feedwater and the distillate; for instance, key factors such as density and specific heat capacity were dependent on feed salinity and initial water temperature, as these elements are crucial for evaluating the water safety criteria necessary for safe human consumption[22].

To develop the baseline CNN-1D model, the original dataset, which comprised 365 samples with five input features and one output related to freshwater productivity, was transformed/reshaped into a time series format using a look-back window of 7. This transformation yielded a structured dataset suitable for time-series forecasting. The data was then divided into training (70%), testing (20%), and validation (10%) sets using a hold-out strategy with a fixed random state of 42 to ensure reproducibility. The baseline CNN-1D model was trained solely on the original dataset, omitting any augmentation. The resulting input data matrix had a shape of [359 × 7 × 5], derived from applying a look-back window of size seven across the 365-day sequence. As shown in Table 3, the target output was formatted as a [359 × 1] vector, representing the daily freshwater yield. The training set was scaled to normalize the feature values, and the same scaling parameters were applied to the validation and testing sets to maintain consistency.

Before model training, the dataset underwent systematic preprocessing. Freshwater productivity values were digitized from Santos [19] (solar still A). Meteorological variables (irradiance, ambient temperature, wind speed, and relative humidity) were obtained from NASA POWER, and the feedwater flow rate was estimated from reported productivity and efficiency correlations. The dataset of 365 daily samples was restructured into a sequential format using a 7-day look-back window. Data augmentation was applied exclusively to training inputs, where Gaussian noise sampled from $\mathcal{N}(0, 0.01^2)$ generated six augmented samples per original observation. All input features were normalized to the [0, 1] range using Min–Max scaling, while the output variable was left unscaled. Finally, the dataset was split into training (70%), validation (10%), and testing (20%) sets, ensuring no overlap between them. These preprocessing steps ensured standardized and temporally consistent inputs for the CNN-1D architecture.

A series of rounds involving random selection and partitioning was conducted to enhance the generalizability and robustness of our findings. This process was repeated ten times to minimize potential biases associated with a single partitioning. Following each iteration, separate training and evaluation for the models were performed, and the results from these iterations were averaged to provide a comprehensive assessment of the models' effectiveness. This methodology enhances reliability by considering the variations in data splits.

In the training phase, the CNN-1D was trained to predict an output, precisely, distilled water in a conventional single-slope solar still system. During the validation phase, the model's performance on unseen data was assessed to monitor for overfitting and to ensure its generalizability. The testing phase involved evaluating the model's capability to forecast a single output value (P_{std}, L/day), which facilitated the evaluation of the model's validity and reliability and helped determine whether training should continue or be halted.
Our analysis employed a supervised machine learning approach, specifically the CNN-1D model. This model was compared against a Support Vector Regressor (SVR) using labeled datasets throughout the training, testing, and validation phases. This comparison aimed to identify the optimal modeling approach for superior prediction outcomes. The CNN-1D and SVR models were constructed using a hold-out strategy that involved 42 random states, which regulated the shuffling process during the train-test split. The baseline CNN-1D model was developed using a Collab notebook and Python software, employing the widely utilized feed-forward/backpropagation algorithm.

Solar stills experience various energy gains and losses influenced by specific weather variables. Our study used the recorded daily operational feedwater volume as input for the CNN-1D model. The feedwater flow rate was included in the analysis due to its relationship with the constant factor of the basin area. Users typically reference the feedwater volume when determining the necessary amounts for flushing and operating solar stills.

The selected input variables—daily solar radiation (Is), ambient temperature (T_{amb}), wind speed (V_w), relative humidity (RH), and saline water feed flow rate (P_{st})—were identified as the key factors affecting solar still performance. These inputs were correlated with the dependent variable, water productivity (P_{std}), which is crucial to our machine learning analysis.

It is essential to highlight that these input parameters were selected due to their significant impact on solar still performance and their effectiveness in accurately predicting freshwater productivity. It was decided to exclude the inlet water temperature for several reasons. The inlet temperature of the basin exhibited a consistently stable pattern, resulting in minimal fluctuations. Preliminary investigations indicated that this parameter did not significantly enhance the model's predictive accuracy concerning water productivity. Incorporating it would complicate the model without providing substantial advantages. Therefore, the five most influential operational variables were focused to optimize the model, thereby enhancing its efficiency and interpretability.

Throughout the model training process, several pivotal decisions were made to enhance the efficacy of the CNN-1D. A total of 50 iterations were selected, employing a batch size of 8 samples, which was determined to be optimal through rigorous hyperparameter tuning. The Adam optimizer was implemented, with a regularization strength of 0.1 and a learning rate of 0.0001, to promote robust training while mitigating the risk of overfitting. To further enhance the model's generalization capabilities, early stopping was employed, incorporating a patience value of 20; this mechanism halted training if no improvement in validation loss was observed over 20 consecutive epochs.

It is pertinent to define an epoch as a complete pass of the entire dataset through the convolutional neural network, encapsulating both forward and backward propagation processes. This systematic methodology for dataset preparation and model training underlines the development of a reliable and accurate CNN-1D model tailored for predicting the performance of solar stills. Furthermore, Table 2 elucidates the sources of input and output parameter data relevant to both configurations of solar still systems, thereby providing a comprehensive framework for understanding the variables involved in this study.

Table 2: Parameter descriptions and data sources. Description of input and output parameters with sources: meteorological drivers from NASA POWER, productivity digitized from [19], and feed water flow rate estimated from efficiency correlations.

Solar Still Design Configurations	**Input Parameters**	**Data Source**	**Output Parameters**	**Data Source**
Conventional Single-Slope Solar Still System (Baseline model)	*Meteorological Conditions:* Solar radiation, MJ/m^2.day Ambient temperature, ^{o}C Wind speed, m/s Relative humidity, %	NASA Power [21].	Solar still productivity Water distillate (L/day)	Literature/ experimental datasets [19].
	Operational Variable: Saline water feed flowrate, L/day	Literature/ experimental datasets [23].		

2.2 Criteria of evaluation for the Augmented CNN-1D model:

In this study, the predictive performance of an augmented CNN-1D model was investigated to forecast the distillate water productivity of a conventional solar still. The augmentation process was applied exclusively to the training dataset to enhance model robustness by introducing variability in the input features. This approach ensured the integrity of the validation and testing datasets, enabling an unbiased assessment of the model's predictive capabilities. The dataset, comprising numerical continuous values, was preprocessed to separate the input features—solar radiation, ambient temperature, wind speed, relative humidity, and feedwater flow rate—from the target output, distillate water productivity. Subsequently, the data was transformed/reshaped into a time-series format using a look-back window of 7, which was determined as optimal through iterative experiments during preprocessing using Python code. This configuration effectively captured the temporal dependencies critical for accurate modeling.

To ensure a reliable evaluation of performance and prevent data leakage, the data augmentation process—specifically, the injection of Gaussian noise—was applied exclusively to the training set. This strategy enhances the model's generalization by introducing additional variability during training, enabling the model to learn stronger and more resilient representations. Notably, the ground truth labels corresponding to the training samples, denoted as $\tilde{y}_{\text{train}}[\text{i}, \text{j}]$, were preserved without modification throughout the augmentation process to maintain accurate supervision. Meanwhile, both validation and testing sets were intentionally excluded from augmentation, preserving their integrity as unbiased benchmarks. Altering these sets could lead to artificially inflated performance metrics by exposing the model to synthetic patterns that it might encounter during training. Therefore, restricting augmentation to the training phase upholds methodological rigor and supports a valid assessment of the model's true predictive capabilities.

For the augmented CNN-1D model, the training dataset, representing 70% of the curated data, was expanded by generating six additional augmented samples for each original instance through a jittering process. To enhance model generalization and address data scarcity, Gaussian noise statistically defined as $\mathcal{N}(0, \sigma^2)$, was systematically injected into the input features during the augmentation phase. As illustrated in Table 3, this process generated 1506 synthetic training samples, each with a dimensionality of [7 × 5], representing seven consecutive days and five input variables. These augmented samples were subsequently concatenated with the original 251 unaltered sequences, yielding a total input training dataset of shape [1757 × 7 × 5]. Correspondingly, the output vector comprising the distillate water productivity values was structured as a [1865 × 1] matrix, inclusive of both augmented and non-augmented labels. Each dataset instance was further structured into time-series submatrices of dimensions [7 × 6], wherein each row captured the temporal progression of six key parameters. The selection of the look-back window (7 days) and the augmentation factor (sixfold increase applied to the training set) was guided by preprocessing experiments and empirical evaluations, aiming to maximize the model's ability to capture short-term dependencies in continuous numerical datasets.

To address dataset limitations, Gaussian noise sampled from $\mathcal{N}(0, 0.01^2)$ was added to the input features during training, where the standard deviation $\sigma = 0.01$ was chosen to represent natural variability within the normalized meteorological and feedwater parameters. Gaussian perturbation was selected because it closely resembles natural variability in meteorological drivers and ensures statistical tractability. The variance parameter ($\sigma^2 = 0.01^2$) was determined through sensitivity analyses, where higher values (>0.02) resulted in unrealistic deviations and smaller values (<0.005) offered negligible benefits. The $\sigma = 0.01$ configuration provided the most consistent improvements in predictive accuracy and generalization, while preserving the physical fidelity of the dataset. An augmentation factor of six was selected after performing sensitivity analyses across ×3, ×6, and ×10 configurations. The ×3 setup yielded only limited improvements relative to the baseline dataset, whereas the ×10 setup introduced redundancy and a slight tendency toward overfitting. By contrast, ×6 consistently provided the most favorable balance, delivering the lowest RMSE and MAE, the highest R^2, and stable generalization across validation and test sets. Notably, augmentation was applied exclusively to the training set to preserve the integrity of the validation and testing phases. This evidence-based selection confirms that the factor of six is both methodologically justified and empirically optimal for the dataset under consideration.
It is worth mentioning that the selection of the look-back window size (7 days) and the number of augmentations per training instance (6) was established through extensive empirical experimentation. Multiple trials were conducted to evaluate various window lengths and augmentation scales, with performance assessed using validation metrics such as RMSE and R^2. A 7-day look-back window was ultimately chosen, as it consistently captured the temporal dependencies in the daily freshwater productivity data without introducing excessive redundancy or overfitting. Similarly, generating six Gaussian noise-augmented samples for each original training point provided a significant increase in data diversity, thereby enhancing model generalization while maintaining computational efficiency. Consequently, these hyperparameters were adopted as the optimal configuration for all augmented models in this study.

Noise in meteorological inputs is an inherent challenge for predictive modeling. In this study, Gaussian noise $\mathcal{N}(0, 0.01^2)$ was deliberately added during augmentation, enabling CNN-1D to learn robust features that generalize well despite variability. The convolutional architecture further

mitigates the effect of random perturbations by emphasizing local spatial–temporal patterns rather than isolated fluctuations. As a result, the augmented CNN-1D consistently outperformed both its baseline counterpart and SVR, highlighting its relative robustness to noisy inputs. Nonetheless, systematic noise from faulty measurements remains a limitation, and future research will investigate the integration of denoising pipelines and physics-informed neural networks to enhance resilience under such conditions.

An important observation is that data augmentation substantially improved the CNN-1D model while providing negligible benefit to SVR. This can be attributed to fundamental structural differences. CNN-1D architectures rely on convolutional filters that extract hierarchical spatial–temporal patterns from sequential data. Exposure to augmented inputs containing Gaussian perturbations encourages the CNN to refine these filters, enhancing robustness to variability and reducing overfitting. By contrast, SVR employs kernel-based mappings with decision boundaries determined by support vectors. Augmented samples generated by small perturbations rarely introduce new support vectors, limiting the capacity of SVR to leverage the additional variability. Consequently, augmentation is more impactful for deep learning architectures capable of adaptive feature learning than for margin-based models with fixed kernel transformations. This distinction underscores the advantage of CNN-based models in data-scarce time-series applications such as solar still productivity forecasting.

The CNN-1D model architecture was refined using hyperparameter optimization. This process adjusted critical parameters such as the number of convolutional layers, dense units, kernel size, batch size, learning rate, and regularization strength to achieve optimal predictive performance. Early stopping with a patience threshold of 20 epochs was implemented during training to prevent overfitting. While the training data was augmented to improve learning diversity, the validation dataset (10% of the curated data) and the testing dataset (20%) were left unaltered to ensure unbiased evaluation. The testing dataset was used to predict known freshwater productivity values to verify accuracy, while the validation dataset was used to monitor generalization performance.

The augmented CNN-1D model was assessed using statistical metrics, including the coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), coefficient of variation (CV), efficiency coefficient (EC), overall index of model performance (OIMP), and residual as defined in Equations (1–7) [24] [17]. Additionally, the predictive performance of the CNN-1D model was benchmarked against an SVR model using identical datasets and evaluation criteria, demonstrating the advantages of data augmentation in enhancing accuracy and robustness.

$$R^2 = 1 - \frac{\sum_{i=0}^{n} \left(y_{o,i} - y_{p,i}\right)^2}{\sum_{i=0}^{n} \left(y_{o,i} - \bar{y}_o\right)^2} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=0}^{n} \left(y_{o,i} - y_{p,i}\right)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=0}^{n} \left|y_{o,i} - y_{p,i}\right| \quad (1)$$

$$CV = \frac{RMSE}{\bar{y}_o} \times 100 \quad (4)$$

$$EC = 1 - \frac{\sum_{i=0}^{n} \left(y_{o,i} - y_{p,i}\right)^2}{\sum_{i=0}^{n} \left(y_{o,i} - \bar{y}_o\right)^2} \quad (5)$$

$$OIMP = 1 - \left(\frac{RMSE}{y_{max} - y_{min}}\right) \cdot (1 - EC) \quad (6)$$

$$Residual = y_{o,i} - y_{p,i} \quad (7)$$

Knowing that the observed value is represented as $y_{o,i}$ the predicted value generated by the machine learning models is denoted as $y_{p,i}$. The average observed value is indicated by $\bar{y}_o$, and the average predicted value, obtained through theoretical estimation, is referred to as $\bar{y}_p$. The minimum and maximum observed values denoted as y_{min} and y_{max} , respectively, and n represents the number of observations or dataset size. Lower values of RMSE and MAE signify more accurate predictions, as both metrics range from 0 to ∞. In contrast, a higher R^2 value indicates a more remarkable similarity between the trendlines of the observed and predicted samples, achieving an identical pattern when $R^2 = 1$ [34].

A perfect fit between observed and forecasted results occurs when the EC value equals 1. Conversely, an EC value of zero indicates that a mean value would yield the same level of accuracy [26] [27]. The latter metric can sometimes even result in a negative value (−∞) if the mean observed value $\bar{y}_o$ is a better predictor than the simulated value, indicating unacceptable performance and extremely poor model predictions. The OIMP value 1 denotes an optimal fit between observed and predicted outputs. The RMSE values should ideally be close to ±1, while better model accuracy is achieved when the RMSE value approaches zero. For effective data modeling, the RMSE, MAE, and CV measures derived from the different trained models should be as close to zero as possible. Simultaneously, metrics such as R^2, EC, and OIMP should converge towards one, indicating improved performance [28]. A CV, expressed as a percentage, serves as a normalized measure of data dispersion, representing the variability of the prediction error relative to the average observed value $\bar{y}_o$. Lower CV values (closer to 0%) reflect less mean variability, which is advantageous for achieving accurate predictions [17].

Figure 1 presents a flow chart depicting the selection and development process of the augmented CNN-1D model, specifically designed to predict distillate water productivity precisely, denoted as P_{std}. Using the statistical metrics expressed in equations (1-7), the optimal CNN-1D architecture

was systematically determined through a trial-and-error methodology, ensuring that the R^2 value exceeded 0.75 when predicting P_{std} using the testing datasets. Only those models meeting this criterion were retained for further comparative analysis, focusing on the predicted response patterns and residuals derived from the testing datasets, in comparison with the actual experimental results obtained from the training datasets.

The augmented CNN-1D model was formulated by implementing a data augmentation technique with a look-back window of 7, following the transformation and reshaping of the dataset into a time-series format. Six augmented instances were generated by introducing Gaussian noise for each original training sample, thereby significantly expanding the training dataset. The augmented CNN-1D model was subsequently trained on this enriched dataset to enhance predictive performance by learning from a more diverse representation of input data variability.

As illustrated in Figure 2, the streamlined architecture and optimized hyperparameters of the augmented CNN-1D model, including the number of convolutional layers, filters, kernel size, learning rate, dense units, and regularization strength, were meticulously fine-tuned to improve generalization and performance across all modelling phases. Additionally, the model incorporated non-linear transfer functions, such as ReLU, to effectively capture the complex relationships inherent in continuous numerical data associated with solar still freshwater productivity, P_{std}. It is essential to note that graphical and statistical comparisons were used to rigorously evaluate the performance of both the baseline and augmented CNN-1D models. Metrics such as R^2, RMSE, MAE, and the OIMP were used to assess the predictive accuracy and reliability of these models. The augmented CNN-1D model framework was implemented through a structured pipeline, as summarized in the corresponding pseudocode Table 4, aiming to enhance predictive accuracy and generalization in modeling solar still productivity.

This study ultimately underscores the effectiveness of the augmented CNN-1D model, trained on the expanded training dataset, in delivering superior predictions and achieving lower statistical errors compared to the baseline model. Figure 3 demonstrates the implementation of a time-series data augmentation framework for developing and evaluating baseline and augmented CNN-1D models. By employing a 7-day look-back window on the original dataset, temporal dependencies were effectively captured. The augmented model was trained on data enhanced with Gaussian noise, aiming to improve prediction robustness and generalization compared to the baseline model, which was trained exclusively on the original samples. This methodology highlights the significant role of data augmentation in enhancing the predictive capabilities of CNN-1D models while concurrently addressing the challenges associated with limited experimental data in solar still studies.

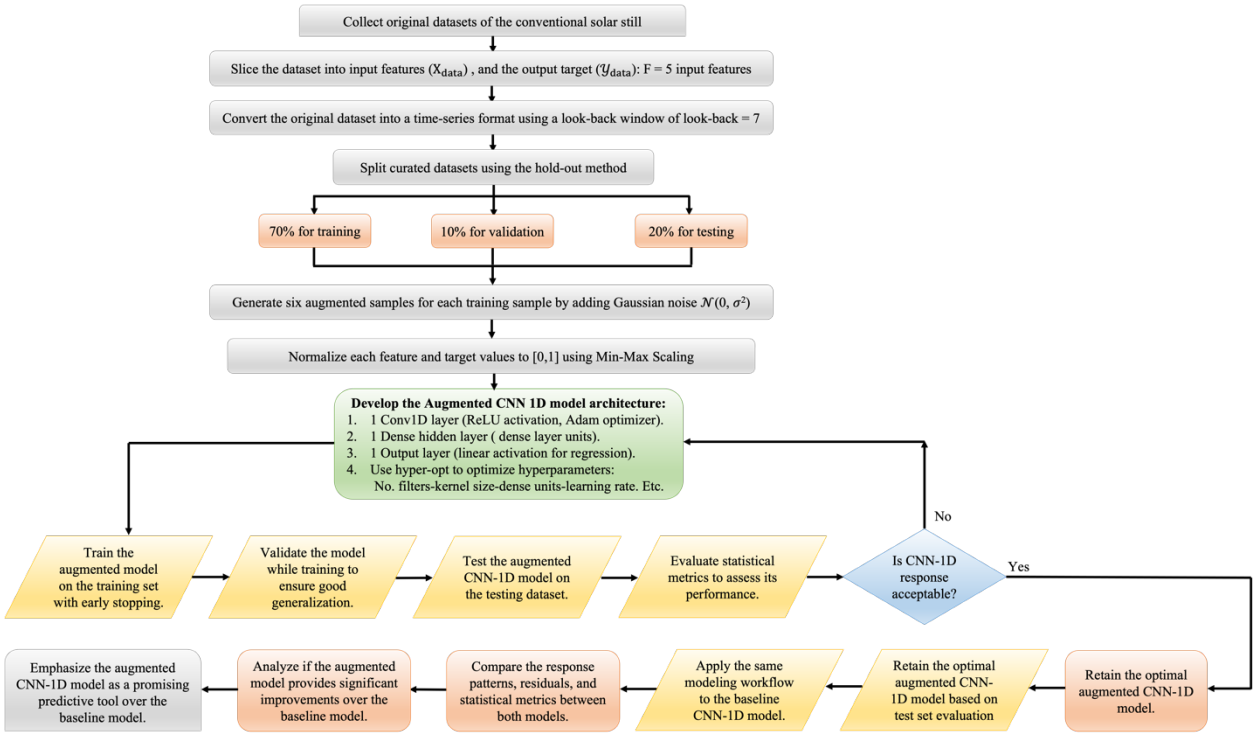


Figure 1: Workflow of the augmented CNN-1D framework. End-to-end process including 7-day look-back, Gaussian noise augmentation, scaling, hyperparameter optimization, training, and evaluation, benchmarked against CNN-1D and SVR.

Table 3 presents the structured input and output shapes for all developed models, utilized for time-series prediction of freshwater productivity in a solar still system. Each model was trained on sequences derived from a fixed 7-day look-back window, transforming the original 365 daily samples into 359 tailored time-series instances. For the augmented CNN-1D and SVR models, the 70% training set was expanded sixfold through Gaussian noise-based data augmentation, significantly increasing sample diversity and size. The CNN models maintained the 3D temporal-spatial structure of the input as samples × time steps × features, while the SVR models required each time-series sequence to be flattened into a 1D vector. This uniform framework ensured consistent evaluation across models, facilitating robust benchmarking. The comprehensive structure outlined in Table 4 validates the consistency of the preprocessing procedures. It supports a comparative performance analysis of the models in forecasting solar distillation outputs under varying environmental conditions.

Table 3: Model matrix dimensions. Dimensions of baseline and augmented CNN-1D and SVR models after 7-day look-back conversion, where augmentation produced six additional training samples per instance and SVR inputs were flattened to (7 × 5) features.

Model	Input Shape	Output Shape	Remarks
Baseline CNN 1D	359 × 7 × 5	**[359 × 1]**	Sliding window; standard time-series sequence
Augmented CNN 1D	[1757 × 7 × 5] Train [36 × 7 × 5] Val [72 × 7 × 5] Test	**[1865 × 1]**	251× 6 = 1506 Augmented 251 original set (36 Val + 72 Test) = 108
Baseline SVR	[359 × 35]	[359 × 1]	Time-series sequences flattened 7 × 5 = 35
Augmented SVR	[1757 × 35] Train [36 × 35] Val [72 × 35] Test	[1865 × 1]	Same logic as CNN 1D flattening input after 6× augmentation

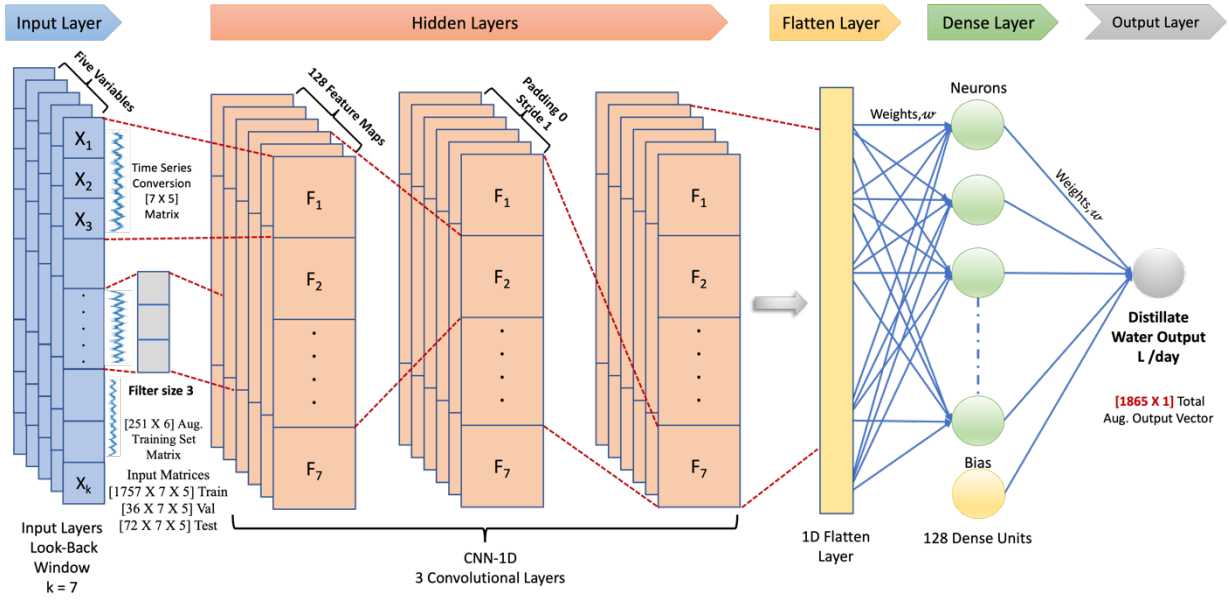


Figure 2: Augmented CNN-1D architecture. Schematic of the optimal 5-(128) ^{3}-1 structure with three convolutional layers, a dense layer, and a single output predicting daily distillate yield.

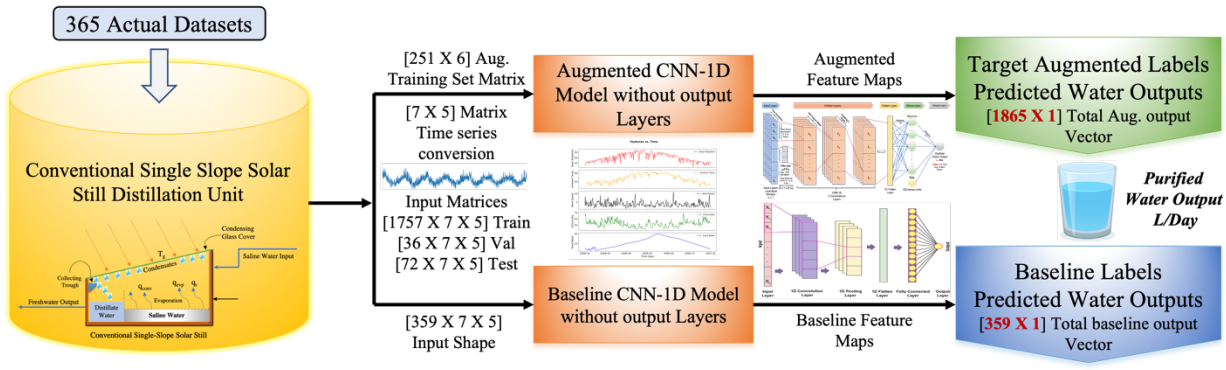


Figure 3: Data augmentation strategy. Block diagram comparing baseline CNN-1D (original data) with augmented CNN-1D (Gaussian noise–enriched data) for solar still forecasting.

This research focused on developing a CNN-1D regression model to forecast continuous numerical variables related to water productivity in solar still systems. Unlike binary classification tasks that rely on discrete labels [0,1], our regression framework leverages non-linear activation functions, specifically ReLU, to effectively capture complex, non-linear dependencies in time series data. This design enables the CNN 1D architecture to learn subtle patterns and relationships across multiple environmental input features, improving predictive accuracy and robustness in modeling continuous distillate water production.

Table 4: Pseudocode of augmented CNN-1D pipeline. Stepwise formulation of preprocessing, augmentation, scaling, model building, hyperparameter tuning, training, and evaluation.

Step	Modelling Name	Pseudocode Explanation/ Mathematical Representation
1	**Import Libraries**	import Pandas, NumPy, train_test_split, MinMax Scaler, Keras layers, Early Stopping, hyperopt

2	**Slice Solar Still Datasets**	$X_{data} = SolarData.iloc[:, :-1].values$ $\mathcal{Y}_{data} = SolarData.iloc[:, -1].values$
3	**Time Series Conversion**	$For\ i \in \{1,2,\ldots, N - look\ back\}$: where, $look\ back = 7$ $X_{ts}[i] = \begin{bmatrix} X\,[i] \\ X\,[i+1] \\ \ldots \\ X\,[i + lookback - 1] \end{bmatrix}$ $\mathcal{Y}_{ts}[i] = \mathcal{Y}[\mathrm{i} + \mathrm{lookback} - 1]$
4	**Datasets Split**	Two stages split into 70% training, 10% validation, 20% testing sets: $(X_{train}, X_{val\ test,}, \mathcal{Y}_{train}, \mathcal{Y}_{val\ test}) = train\_test\_split(X_{ts}, \mathcal{Y}_{ts}, 0.7)$ $(X_{val}, X_{test,}, \mathcal{Y}_{val}, \mathcal{Y}_{test}) = train\_test\_split(X_{val\ test,}\ \mathcal{Y}_{val\ test}, 0.67)$
5	**Augmentation (Training Data Only)**	$For\ i \in \{1,2,\ldots, M_{train}\},\quad j \in \{1, 2, \ldots, 6\}$: $\tilde{X}_{train}[i,j] = X_{train}[i] + \mathcal{N}(0, \sigma^2)$, $\tilde{\mathcal{Y}}_{train}[i,j] = \mathcal{Y}_{train}[i]$
6	**Scaling**	$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad \mathcal{Y}_{scaled} = \frac{\mathcal{Y} - \mathcal{Y}_{min}}{\mathcal{Y}_{max} - \mathcal{Y}_{min}}$
7	**Define CNN-1D Model (Conv Layers)**	CNN1D Model = Conv1D ⟶ Flatten ⟶ Dense Units ⟶ Output For $l = 1, 2, 3$: $Z^{(l)}[i,j] = \sum_{k=0}^{K-1} w^{(l)}[k,j] \cdot A^{(l-1)}[i+k] + b^{(l)}[j]$, $A^{(l)}[i,j] = \max(0,\ Z^{(l)}[i,j])$ and $A^{(0)} = x$, ReLU activation $P = \left[\frac{K-1}{2}\right]$ $(For\ 'padding = same')$ $Ensures\ L_{out} = L_{in}, S = 1$
8	**Flatten Layer**	$f = Flatten\left(A^{(3)}\right) \epsilon\ \mathbb{R}^M, where\ M = T \times 128$
9	**Dense Layer**	$h_m = \max\left(0, \sum_{i=1}^{M} W_{mn} \cdot f_n + b_m\right),\ m = 1, \ldots, 128$
10	**Output Layer**	$\hat{\mathcal{Y}} = \sum_{i=1}^{M} w_m^{(0)} \cdot h_m + b^{(0)},\ \hat{\mathcal{Y}} \epsilon\ \mathbb{R}^M$
11	**Early Stopping**	$EarlyStopping\ (monitor = "val_{loss}", \mathrm{patience} = 20)$
12	**Hyperparameter Optimization**	$Optimize$: $filters \in \{8, 16, 32, 64, 128\}$, $kernel\ size \in \{2, 3, 5\}$, $dense\ units \in \{8, 16, 32, 64, 128\}$, $learning\ rate, \eta \in [10^{-4}, 10^{-2}]$, $batch\ size \in \{8, 16, 32, 64\}$, $reg\ strength, \lambda \in [10^{-4}, 10^{-2}]$
13	**Model Training**	$Minimize\ loss$: Loss: $MSE = \frac{1}{M} \sum_{i=1}^{M} (\mathcal{Y}_{o,i}[i] - \mathcal{Y}_{p,i}[i])^2$

		$Train$: $Adam\ optimizer, MSE\ loss\ , MAE\ metric$
14	**Model Evaluation (Key Metrics)**	$Statistical\ Metrics\ Evaluation\ on\ test\ set$: $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{o,i}[i] - y_{p,i}[i])^2}$ $R^2 = 1 - \frac{\sum_{i=0}^{n}\left(y_{o,i} - y_{p,i}\right)^2}{\sum_{i=0}^{n}\left(y_{o,i} - \bar{y}_o\right)^2}$ $MAE = \frac{1}{n}\sum_{i=0}^{n}\lvert y_{o,i} - y_{p,i}\rvert$

Although conventional solar stills have been extensively studied experimentally, their predictive modeling remains underdeveloped due to limited datasets and the nonlinear nature of their governing variables. This study directly addresses energy system performance by developing a robust forecasting framework that enables accurate prediction of freshwater productivity under variable meteorological conditions. The proposed modeling approach complements experimental investigations by transforming scarce field measurements into actionable predictive tools, thereby supporting operational optimization and resource planning for solar desalination systems.

2.3 Developing Baseline and Augmented SVR Models:

During the data preprocessing stage for the baseline and augmented SVR models, the datasets underwent a systematic transformation designed explicitly for time-series regression. For the augmented SVR model, which employed the same preprocessing pipeline as the augmented CNN-1D models, the original training data (70% of the complete dataset) was enhanced through a Gaussian noise-based augmentation strategy. This approach introduced zero-mean Gaussian noise, represented as $\mathcal{N}(0, \sigma^2)$, to each of the five input features, generating six synthetic variations for every original instance. As a result, the augmented SVR training samples were subsequently concatenated with the original 251 unaltered sequences to total input training dataset dimensions [1757 × 35], with each row containing five predictors—solar radiation, ambient temperature, wind speed, relative humidity, and feedwater flow rate—and one output variable, distillate water productivity. Correspondingly, the output vector comprising the distillate water productivity values was structured as a [1865 × 1] matrix, inclusive of both augmented and non-augmented labels. In contrast, the baseline SVR model was trained solely on the original dataset, omitting any augmentation. The resulting input data matrix had a shape of [359 × 7 × 5], derived from applying a look-back window of size seven across the 365-day sequence. The target output was formatted as a [359 × 1] vector, representing the daily freshwater yield.

It is crucial to understand that the input matrix dimension of [359 × 7 × 5] in the baseline CNN-1D and SVR models was derived by employing a sliding look-back window of size seven on the original dataset, which contains 365 consecutive daily observations. This time-series transformation technique segments the data into overlapping sub-sequences, with each input instance consisting of 7 consecutive days used to forecast the output for the next day. As a result, the number of effective sequences is determined by the formula 365 - 7 + 1 = 359, producing an input matrix of shape [359 × 7 × 5] and a corresponding output vector of [359 × 1], representing

the daily freshwater yield. This preprocessing strategy—utilizing a uniform look-back window of 7—was applied uniformly across all four developed models, including both the baseline and augmented variants of the CNN-1D and SVR architectures, thereby ensuring a standardized and equitable framework for comparative evaluation.

To adequately account for temporal dependencies, all instances were reorganized into time-series submatrices of size [7 × 6]. The 7-day look-back window was fine-tuned through various empirical evaluations. This configuration enabled the SVR models to integrate historical context, effectively transforming each time-windowed matrix into a one-dimensional vector of length 42, suitable for the SVR architecture. These preprocessing steps ensured uniformity across baseline and augmented SVR models while facilitating a fair comparison with the CNN-1D model in subsequent predictive performance evaluations.
In addition to developing the CNN model, the present study employed SVR models in both baseline and augmented configurations to predict freshwater productivity derived from conventional solar stills. The initial phase involved constructing a baseline SVR model using the original dataset without augmentation. A comprehensive hyperparameter tuning process was implemented through the TPE algorithm, facilitated by the Hyperopt library. This process aimed to optimize critical model parameters, specifically the regularization parameter (C), the epsilon-insensitive loss parameter (ε), and the selection of the kernel type (linear, polynomial, RBF, or sigmoid). After completing 100 optimization trials, the most effective hyperparameters were determined to construct the final baseline SVR model. Subsequently, an augmented SVR model was devised by training it on a considerably expanded dataset generated through Gaussian noise-based data augmentation (jittering). A parallel hyperparameter optimization procedure comprised 100 trials to thoroughly investigate the parameter space. The optimized augmented SVR model prominently employed the RBF kernel function, facilitating nonlinear mapping of inputs into a higher-dimensional feature space, thereby enhancing predictive accuracy. Both SVR models underwent rigorous evaluation across training, validation, and test datasets, employing consistent statistical metrics that were also utilized to assess the CNN-1D model. A comparative analysis of the two SVR implementations yielded significant insights into the efficacy of data augmentation techniques, highlighting their influence on the predictive reliability and robustness of the SVR modeling framework.

3. Results and Discussion

This section focuses on developing and optimizing a CNN-1D model using data augmentation to enhance prediction accuracy for daily water outputs. Two models were designed and assessed: the baseline CNN-1D, which is non-augmented, and the augmented CNN-1D. A statistical comparison and performance analysis were performed between these models to illustrate the effects of data augmentation on the models' generalization capabilities and accuracy in predicting water outputs. Furthermore, the results included a benchmark of the CNN-1D model against a support vector regression (SVR) model for additional validation.

3.1 Optimization of baseline CNN-1D model architecture:

This study employed CNN-1D and SVR regression models to predict the output of purified water from a conventional single-slope solar still. The performance statistics of the baseline CNN-1D model, which employed various convolutional filters, kernel sizes, dense units, and learning rates, are summarized in Table 5. Through a rigorous optimization process involving a maximum of 50

trials, the optimal architecture for the baseline CNN-1D model was selected. This architecture consists of three convolutional layers, each with 64 filters and a kernel size of 3, and complemented by a dense layer containing eight units. The optimization aimed to find the optimal combination of hyperparameters to balance the model's complexity and predictive accuracy.

During the optimization process, the batch size varied from 8 to 64, and the ReLU activation function was consistently utilized due to its effectiveness in mitigating the vanishing gradient problem and facilitating efficient training. Despite having a seemingly large number of parameters, the chosen architecture demonstrated strong generalization capabilities. Empirical results confirmed that this configuration minimized validation loss and improved predictive accuracy. Techniques such as L2 regularization (with a regularization strength of 0.0001) and early stopping (with a patience of 20 epochs) were effectively employed to address the issue of overfitting. Incorporating three convolutional layers with a kernel size of three was essential for capturing spatial and temporal features from the input data. Using 64 filters in each layer struck an optimal balance between extracting meaningful features and maintaining computational efficiency. The dense layer, containing eight units, further consolidated the learned representations, thereby enhancing the model's ability to accurately predict distillate water output. The choice of filters, kernel size, and dense units within the baseline CNN-1D model was crucial for its learning and generalization capabilities. While fewer filters or reduced kernel sizes might lessen computational demands, they could also hinder the model's capacity to capture underlying patterns in the data.
On the other hand, larger architectures could increase computational complexity without delivering significant improvements in predictive accuracy. Thus, the selected optimal configuration—64 filters, a kernel size of 3, and 8 dense units—effectively balanced these trade-offs, ensuring robust and accurate predictions. The optimization demonstrated the model's ability to generalize effectively across the augmented training and untouched testing datasets. The results indicate that this architecture was particularly effective in capturing the complexities and non-linear relationships in the dataset, ensuring reliable predictions of treated water output for the conventional solar still.
Early stopping was implemented with a patience of 20 epochs to prevent overfitting. The ReLU activation function was applied across all layers due to its simplicity and effectiveness in enhancing the model's predictive performance. Through rigorous optimization, the baseline CNN-1D model achieved the best validation loss of 0.006711 during training and demonstrated substantial predictive accuracy on the testing datasets, with a test RMSE of 0.1, MAE of 0.08, and R^2 of 0.86.
The Adam optimizer was employed in the optimization process due to its computational efficiency and adaptability in adjusting learning rates during training, which enabled smoother and quicker convergence. In contrast to optimizers like SGD, Adagrad, and RMSprop, Adam incorporates momentum and adaptive learning rate adjustments, making it particularly suitable for managing the high-dimensional, non-linear relationships inherent in the freshwater water output dataset. Other gradient descent methods, such as SGD, which relies on a constant learning rate, were excluded because of their slower convergence and higher sensitivity to hyperparameter tuning. While Adagrad can be helpful in sparse datasets, it is prone to diminishing learning rates over time, which is not ideal for this regression task.
The ReLU activation function was exclusively chosen for all layers because it alleviates the vanishing gradient problem, a common issue in deep neural networks. Its simplicity and computational efficiency make it a preferred choice for regression tasks. Alternative activation

functions, such as sigmoid and Tanh, were ruled out due to their slower training performance and tendency to saturate gradients for large input values, which can lead to slower convergence. The combination of ReLU and batch normalization contributed to a more stable training process and improved overall model performance. The optimization process systematically evaluated the number of convolutional filters, kernel sizes, dense units, and batch sizes.

While existing empirical guidelines can provide a starting point for hyperparameter selection, this study employed an empirical assessment tailored to the dataset's characteristics. This approach ensured that the chosen architecture was well-suited to predicting the complexity of the treated water output. By evaluating a range of hyperparameters and leveraging the computational efficiency of the Adam optimizer alongside the robustness of the ReLU activation function, the selected baseline CNN-1D model architecture achieved a balance between computational efficiency and predictive performance.

One key hyperparameter considered during the training process is the number of epochs, which indicates how often the learning algorithm will traverse the entire training dataset. Additionally, the learning rate for the Adam optimizer and the momentum were fine-tuned to maintain fixed values of 0.01 and 0.9, respectively. Having too many epochs can result in an overfit model, while having too few can lead to underfitting of the training dataset. To address this, an early stopping strategy supported by a callback function can be implemented during model training. This approach allows us to define an arbitrarily large number of training epochs (for example, 120 epochs) and terminate the training process whenever the model's performance on a holdout validation dataset stops improving. To attain this technique, the validation dataset was passed to the fit function during training, enabling us to identify a performance indicator to monitor for stopping criteria. The loss on the validation dataset is referred to as "validation loss," which aims to achieve a minimum value. However, stopping training at the first indication of no further improvement is not always the best strategy. During training, the early stopping event may occur if the model reaches a plateau with little progress, or it may even experience slight deterioration before making significant improvements. A delay can be introduced before triggering the stop to accommodate this potential behavior. This is accomplished by specifying the "patience" parameter, which allows us to observe no improvement for a certain number of epochs—in this case, it can be set to 20 [29].

Convolutional filter counts were varied across (16, 32, 64, 128) as part of a global hyperparameter optimization using TPE–Hyperopt (50 maximum evaluations). More minor filter counts (16–32) resulted in underfitting, with reduced predictive accuracy. Larger configurations (>128) increased training duration and introduced mild overfitting, as reflected by divergence between training and validation losses. The 128-filter configuration consistently offered the best trade-off, achieving the lowest error metrics without overfitting when coupled with L2 regularization and data augmentation. The stability of this configuration was further confirmed by its repeated appearance among the top 10 optimization trials, as shown in Tables 5 and 6.

Table 5: CNN-1D hyperparameter optimization results. Top trial results with tuned convolutional layers, dense units, filters, batch size, learning rate, and regularization strength.

Trial number	Conv. layers size	Dense Units	Filters	Batch size	Learning rate	Lambda (Weight decay)	Loss
1	**3**	**8**	**64**	**64**	**0.01**	**0.0001**	**0.006711**
2	4	128	8	8	0.0001	0.001	0.013344
3	1	128	16	8	0.0001	0.001	0.013849

4	3	128	64	64	0.001	1	0.014300
5	1	128	64	32	0.0001	0.1	0.030448
...	...	...	...	...	...	...	...
46	3	16	32	32	0.0001	0.1	0.030752
47	2	64	16	64	0.00001	0.01	0.667131
48	4	8	128	8	0.1	0.001	0.795904
49	4	8	16	8	0.00001	1	1.430148
50	2	128	16	32	0.00001	0.1	5.537681

3.2 Optimization of the augmented CNN-1D model architecture:

The augmented CNN-1D model was developed based on the foundational structure of the baseline CNN 1D model, significantly enhanced through systematic data augmentation and rigorous hyperparameter optimization. Gaussian noise injection was employed to effectively expand the training dataset, which facilitated improved generalization capabilities of the model. The final architecture comprises three convolutional layers, each containing 128 filters with a kernel size of 3, followed by a flatten layer and a fully connected dense layer with 128 neurons. L2 regularization ($\lambda = 0.1$), a batch size of 8, and a learning rate of 0.0001 were implemented to address the overfitting challenges and stabilize the training process. These hyperparameters were optimized through a comprehensive search methodology, as outlined in Table 6, where the optimal configuration, highlighted in the first row, attained the lowest loss value of 0.001359. This architecture demonstrated an exceptional capacity for capturing intricate temporal patterns within the augmented time-series data, resulting in enhanced predictive performance for freshwater output.

Visualizing the validation loss and training loss across the number of epochs (118, 200) for both the baseline and augmented models, respectively, is instrumental in assessing the adequacy of the model's training process. As depicted in Figure 4, an optimal model fit is characterized by the simultaneous reduction and stabilization of validation and training losses at specific points. This methodology is critically essential for mitigating the risks of overfitting the CNN-1D model and preventing underfitting, which can result in the model merely memorizing the training data and subsequently diminishing its predictive accuracy [30].

Through the analysis of the learning curves, it can be observed that the performance of the CNN 1D model is well-balanced, showing no discernible indications of high variance, high bias, or overfitting. The following key observations can be highlighted:

1. Loss Reduction: Training and validation losses consistently decrease as the number of epochs increases, indicating that the model is successfully learning from the data and enhancing its predictive capabilities.

2. Minimum Plateau: The training and validation losses reach a minimum plateau, indicating that the model has converged to an optimal point, as further training yields minimal additional reductions in loss. This behavior provides evidence against the occurrence of overfitting.

3. Convergence of Curves: The convergence and stabilization of both loss curves as the number of epochs progresses suggest that the model's performance on the training and validation datasets has become consistent. This implies that an appropriate balance between fitting the training data and generalizing to unseen data has been achieved.

4. Absence of High Variance or Overfitting: The steady convergence of the curves indicates a lack of significant disparity between training and validation losses, suggesting that overfitting is not a pressing concern. In typical scenarios, overfitting is evidenced by an increasing gap between training and validation losses, as the model excessively adapts to the training data.

5. Absence of Underfitting or High Bias: There is no indication of high bias, which would manifest as sustained high values for training and validation losses without decline. High bias suggests that the ANN model lacks complexity and cannot adequately capture the underlying patterns present within the datasets. This comprehensive analysis indicates that the model has been adequately trained, achieving a predictive performance in equilibrium.

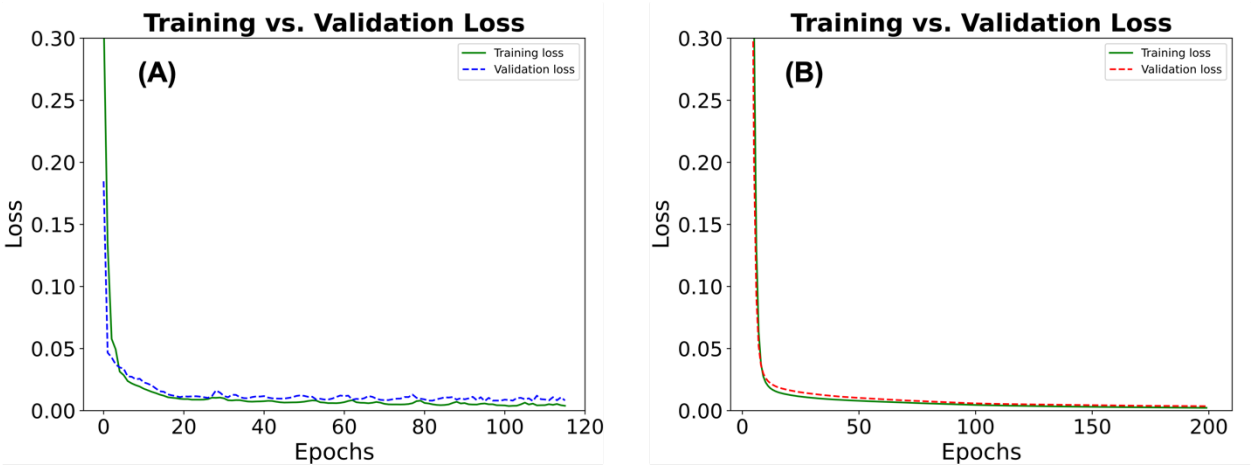


Figure 4: Training and validation loss curves. (A) Baseline CNN-1D: minor fluctuations, moderate generalization. (B) Augmented CNN-1D: stable convergence with overlapping losses, improved robustness.

Table 6: Augmented CNN-1D hyperparameter optimization. Tuned parameters across 50 trials; optimal configuration achieved lowest loss with 3 convolutional layers, 128 filters, and batch size 8.

Trial number	Conv. layers size	Dense Units	Filters	Batch size	Learning rate	Lambda (Weight decay)	Loss
1	**3**	**128**	**128**	**8**	**0.0001**	**0.1**	**0.001359**
2	3	128	128	8	0.0001	0.1	0.001446
3	3	16	128	8	0.0001	0.1	0.001610
4	3	8	128	8	0.0001	0.1	0.001745
5	3	128	64	8	0.0001	0.001	0.001785
...	...	...	...	...	...	...	...
46	3	128	128	8	0.0001	0.001	0.001791
47	3	16	128	8	0.001	0.1	0.001888
48	3	16	128	8	0.1	0.1	0.002163
49	3	128	128	16	0.0001	0.001	0.002238
50	4	128	64	8	0.0001	0.01	0.002324

3.3 Criteria of Optimization for SVR Models

The SVR models underwent thorough hyperparameter optimization using the TPE algorithm. For the augmented SVR model, as shown in Table 7, optimal hyperparameters were identified within 100 optimization trials, yielding a minimal validation loss of 0.01289. The best configuration included a regularization parameter (C) of 1.657, an epsilon (ε) value of 0.135, and the use of an RBF kernel. In contrast, the baseline SVR model required up to 200 optimization trials to achieve its lowest validation loss of 0.01339, as illustrated in Table 8. Its optimal configuration featured a substantially higher regularization parameter (C = 116.801) while maintaining a similar epsilon (ε = 0.135) and kernel type. The extended trial count for the baseline SVR model was empirically driven by its slower convergence behavior, likely due to the limited data variability in the absence

of augmentation. This contrasts with the augmented SVR, which benefited from the enriched training dataset, allowing faster convergence and improved generalization with a lower C value.
The notable disparity in regularization strength (C) highlights the influence of augmentation, as the augmented dataset enabled the SVR model to generalize effectively with a considerably lower regularization parameter, thereby showcasing improved robustness and stability. These trial limits were not arbitrarily selected but reflect the distinct convergence characteristics of the two models, ensuring a fair and robust comparison under consistent optimization protocols. The consistent use of the RBF kernel in both scenarios highlights the inherent nonlinear relationships within the solar still dataset, underscoring the need for nonlinear modeling approaches to accurately predict freshwater productivity. A comparative analysis of the baseline and augmented SVR models reveals consistent yet marginal enhancements in predictive performance attributable to the implementation of the data augmentation technique. Specifically, the augmented SVR model demonstrates a slightly elevated testing R^2 of 0.853 compared to the baseline's 0.848, alongside an improved OIMP of 0.877 versus 0.874.
Additionally, there are marginal reductions in RMSE (0.101 vs. 0.102), MAE (0.087 vs. 0.088), and CV (0.226 vs. 0.228), indicating a decrease in residual dispersion and an improvement in predictive precision. These subtle improvements were similarly observed throughout the validation and training phases. Notably, the augmented model achieved a best validation loss of 0.0129 with a regularization parameter of C = 1.657, which is significantly lower than the baseline model's C = 116.80, resulting in a higher loss of 0.0134. This considerable reduction in regularization strength reflects improved model generalization, suggesting that the augmented SVR model may attain comparable or superior accuracy utilizing a simpler decision function. Collectively, these results highlight the stabilizing effect of data augmentation within the SVR framework, which enhances robustness, reduces model complexity, and ultimately improves the reliability of forecasting freshwater productivity in solar still systems.

Table 7: Augmented SVR hyperparameters. Optimized values of C, ε, and kernel type from 100 trials; best model employed RBF kernel.

Trial Number	Reg. Strength C	Epsilon ε	Kernel	Loss
1	**1.6570**	**0.1350**	**rbf**	**0.01289**
2	23.2988	0.1369	rbf	0.013199
3	19.0919	0.1453	rbf	0.013780
4	14.5340	0.1463	rbf	0.013871
5	55.6788	0.1358	poly	0.014284
...	...	...	...	...
6	136.7895	0.1365	poly	0.014291
7	53.3907	0.1359	poly	0.014295
8	1.5817	0.1510	rbf	0.014308
9	28.4521	0.1404	poly	0.014498
10	16.6311	0.1549	rbf	0.014686

Table 8: Baseline SVR hyperparameters. Top configurations from 200 trials, with RBF kernel consistently selected; best result achieved with a higher C value.

Trial Number	Reg. Strength C	Epsilon ε	Kernel	Loss
1	**116.801**	**0.1350**	**rbf**	**0.013390**

2	20.5278	0.1356	rbf	0.013394
3	135.8546	0.135674	rbf	0.013398
4	352.3807	0.135674	rbf	0.013398
5	102.5012	0.135629	rbf	0.013403
...	...	...	...	...
6	399.6159	0.135630	rbf	0.013403
7	39.7102	0.135768	rbf	0.013407
8	401.6382	0.135870	rbf	0.013421
9	123.0069	0.135994	rbf	0.013424
10	7.0628	0.136426	rbf	0.013448

3.4 Performance analysis of CNN 1D and comparison with SVR:

Figure 5 presents a comprehensive comparative analysis of the baseline CNN 1D and baseline SVR models across the training (A, D, G), validation (B, E, H), and testing (C, F, I) phases. **During the training phase** (Figure 5A), the CNN-1D model demonstrates a notably stronger alignment with observed distillate water output values than the SVR model, as its predictions closely reflect the temporal trends evident in the actual data. This close correlation highlights the CNN model's superior learning capacity and stability. In contrast, the SVR model exhibits significant deviations, particularly in regions of peak output. The CNN-1D model achieves a high **R^2 of 0.907** and **efficiency coefficient (EC) of 0.907**, indicating that the model explains approximately **90.7% of the output variance**. By contrast, the SVR model attained lower values (**R^2 = 0.875**, **EC = 0.875**), reflecting a **3.5% reduction in explained variance**. The CNN-1D model also demonstrated improved stability, with a **coefficient of variation (CV) of 0.177** compared to SVR's higher **CV of 0.205**, denoting a **15.8% increase in prediction variability**. Moreover, the CNN's error metrics were notably lower (**RMSE = 0.081 L/m²·day**, **MAE = 0.065 L/m²·day**) relative to SVR (**RMSE = 0.094**, **MAE = 0.082**), corresponding to **13.8% and 20.7% reductions**, respectively.

Regression plots, illustrated in Figure 5D, further substantiate these observations; the predicted values generated by the CNN model are tightly clustered around the 45° reference line, indicating perfect concordance with the observed outputs. In contrast, the SVR model exhibits a broader distribution of points, especially at both extremes, reflecting diminished predictive accuracy. The residual plot depicted in Figure 5G corroborates this trend, revealing that **~ 90%** of the residuals from the CNN model fall within **±2%.** However, the SVR residuals display better dispersion and skewness, indicative of higher variability in prediction errors.

In the **validation phase** (Figure 5B), the CNN-1D model retains superior performance, consistently aligning closer to actual output values across the entire time series. It achieves an **R^2 = 0.872**, **EC = 0.872**, **OIMP= 0.911** and **CV = 0.219**, while the SVR model recorded **R^2 = 0.809**, **EC = 0.809**, **OIMP= 0.840** and **CV = 0.268**. These values reflect a **7.2% improvement in variance explanation** and an **18.3% reduction in relative variability** for CNN over SVR. Error metrics again favored CNN (**RMSE = 0.095**, **MAE = 0.078**) over SVR (**RMSE = 0.115**, **MAE = 0.098**), indicating reductions of **17.4% and 20.4% in MAE for the CNN model**. The regression plot in Figure 5E confirms that CNN predictions adhere more closely to the 1:1 reference line, whereas SVR predictions oscillate above and below. Residual analysis (Figure 5H) further

supports this observation, with CNN residuals remaining compact and centered around zero, while SVR residuals exhibit more erratic and skewed patterns.

During the testing phase (Figure 5C), the CNN model exhibits superior generalization capabilities, producing predicted outputs that closely align with the observed distillate water values. The model attained an **R^2 = 0.862**, **EC = 0.862**, and **CV = 0.218**, outperforming the SVR model (**R^2 = 0.848**, **EC = 0.848**, **CV = 0.228**). These results represent a **1.4% improvement in explained variance** and a **4.4% reduction in prediction variability**. The CNN also produced lower error metrics (**RMSE = 0.097**, **MAE = 0.077**) than SVR (**RMSE = 0.102**, **MAE = 0.088**), reflecting a reduction of **4.9% and 12.5%**, respectively. As depicted in Figure 5F, CNN predictions aligned more closely with the diagonal reference line. In contrast, SVR predictions fluctuated significantly above and below this line, indicating more variability and inconsistencies in generalization. Residuals in Figure 5I reinforce this distinction, with CNN residuals predominantly contained within ±2%. On the other hand, the SVR residuals demonstrate a broader and less symmetric distribution, indicating a higher degree of uncertainty in predictions.

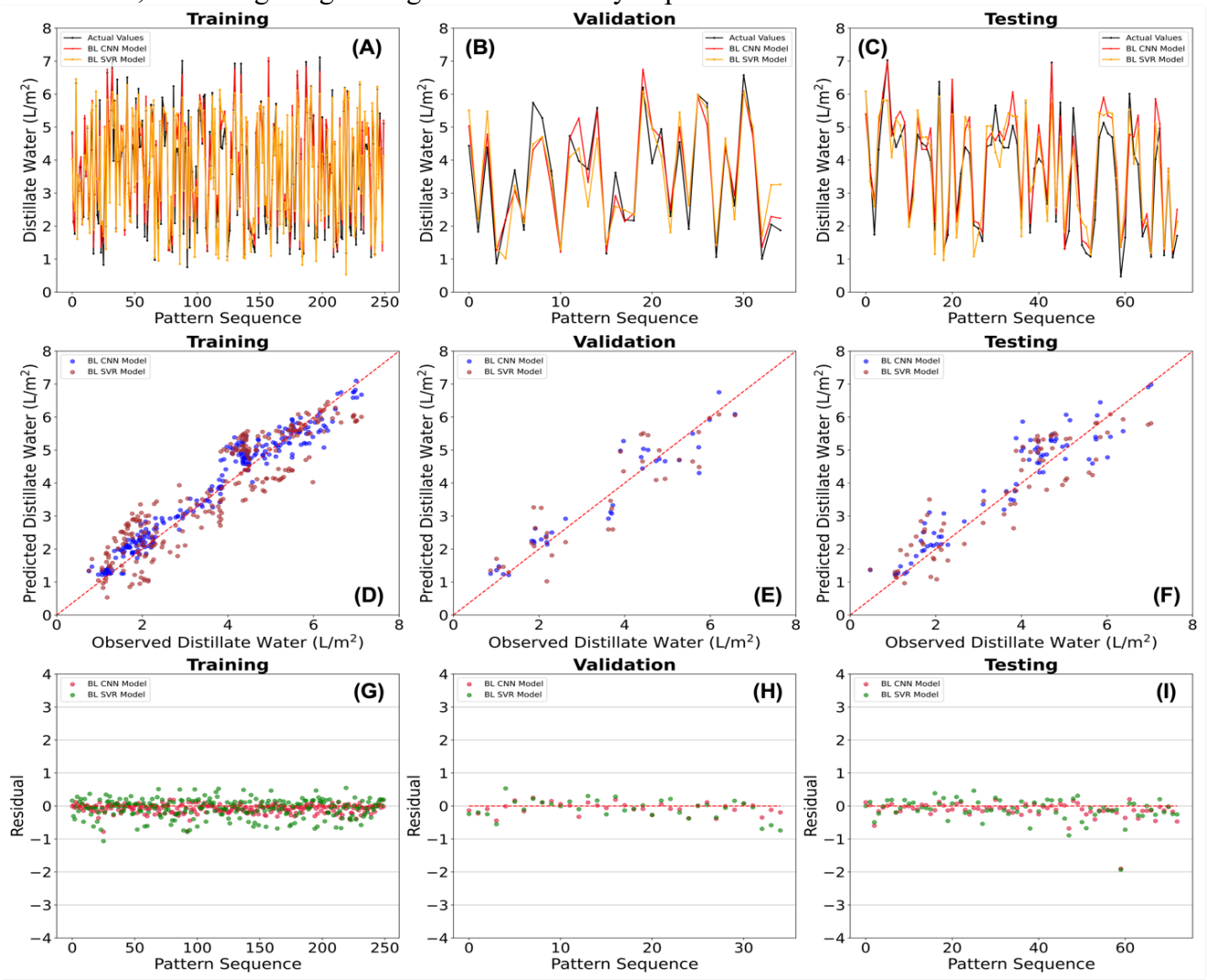


Figure 5: Baseline CNN-1D vs. SVR performance. Comparative temporal predictions, regression fits, and residuals across training, validation, and testing, with CNN showing closer alignment and tighter residuals.

Figure 6 presents an integrated performance analysis of the augmented CNN-1D and SVR models, elucidating the significant advancements attained by implementing the proposed time series data augmentation strategy. In the **training phase** (Figure 6A), the augmented CNN model exhibited near-perfect learning behavior, with predicted outputs overlapping seamlessly with the observed values. It achieved an outstanding **R^2 of 0.986**, an **EC of 0.994**, and an **OIMP of 0.987,** indicating that **99.4% of the output variance** was explained. By comparison, the augmented SVR model yielded lower values (**R^2 = 0.879**, **EC = 0.879,** and **OIMP of 0.893**), reflecting an **11.5% drop in predictive efficiency**. The **CV** for CNN was just **0.045**, compared to **0.202** for SVR—a **77.7% reduction in prediction variability**. Additionally, the CNN achieved significantly lower errors (**RMSE = 0.031**, **MAE = 0.023 L/day**) compared to SVR (**RMSE = 0.092**, **MAE = 0.084**), translating to **66.3% and 72.6% reductions**, respectively, and a **12.2%** enhancement in R^2 for the augmented CNN model. The regression plot in Figure 6D reveals that CNN predictions are tightly aligned along the 45° reference line, while SVR predictions are broadly scattered. Residual analysis in Figure 6G confirms that **over 95% of CNN residuals** fall within **±1%**, whereas SVR residuals are more dispersed and asymmetric.

During the **validation phase** (Figure 6B), the augmented CNN model sustained superior performance, with **R^2 = 0.971**, **EC = 0.981**, **OIMP of 0.971**, and **CV = 0.087**, indicating that **only 1.9% of variance remained unexplained**. Conversely, the augmented SVR model achieved an **R^2 = 0.813**, **EC = 0.810**, **OIMP of 0.841**, and a much higher **CV of 0.267**, reflecting a **67.6% increase in variability** over the CNN. The CNN also maintained lower error values (**RMSE = 0.045**, **MAE = 0.035**) compared to SVR (**RMSE = 0.114**, **MAE = 0.098**), with improvements of **60.5% in RMSE** and **64.3% in MAE,** accompanied by a 19.4% increase in R^2 in favor of the CNN model. Notably, the CNN's OIMP score demonstrated a 15.4% improvement over SVR, underscoring its more substantial overall predictive reliability and ability to capture complex nonlinear relationships within the augmented time-series data. Figure 6E clearly illustrates the alignment of CNN predictions with the 1:1 diagonal line, in contrast to the significant deviations noted in the SVR results. As shown in Figure 6H, **~90% of CNN residuals** remained within **±1%**, whereas SVR residuals display a more erratic and widely scattered residual pattern.

In the **testing phase** (Figure 6C), the augmented CNN demonstrated outstanding generalization capabilities, maintaining **R^2 = 0.971**, **EC = 0.979**, **CV = 0.079,** and **OIMP of 0.9696**. These values affirm the model's ability to sustain predictive accuracy across unseen data. By comparison, the augmented SVR recorded **R^2 = 0.853**, **EC = 0.852**, **CV = 0.226**, **and OIMP of 0.877**, which reflect a **13.8% drop in explained variance**, a **13.7% reduction in efficiency**, and a **65.1% increase in output variability** relative to CNN. Additionally, the CNN model achieved a 9.1% improvement in OIMP over SVR, highlighting its superior overall performance and robustness in capturing the nonlinear patterns in the data. The CNN model also produced much lower errors (**RMSE = 0.045**, **MAE = 0.035**) than SVR (**RMSE = 0.101**, **MAE = 0.087**), with corresponding error reductions of **55.4% and 59.8%**, respectively. Figure 6F illustrates a tight distribution of CNN predictions along the 45° slope line, while SVR predictions exhibit a more dispersed trend. Residual plots in Figure 6I further validate this difference: **approximately 92% of CNN residuals** fall within **±1%**, while only **~75% of SVR residuals** do, underscoring CNN's greater consistency and reduced prediction uncertainty.

In conclusion, the aggregate evidence derived from Figures 5 and 6, in conjunction with the detailed statistics outlined in Table 9, unequivocally demonstrates that the augmented CNN-1D model consistently outperforms all other configurations, exhibiting the highest accuracy, lowest error magnitudes, and strongest generalization. It achieved near-perfect agreement with the observed data, as indicated by its consistently high R^2 values (~0.97), (EC $\geq$ 0.97), and OIMP $\approx$ 0.97, while maintaining exceptionally low error dispersion with CV values ranging from 0.045 to 0.087. In contrast, the SVR models, particularly in their baseline form, exhibited weaker performance with lower R^2 and EC values ($\leq$ 0.875), lower OIMP scores (as low as 0.84), and higher CVs (up to 0.268), reflecting greater variability and reduced reliability. The OIMP analysis further emphasizes the superiority of the augmented CNN model, with an improvement of over 9% compared to the augmented SVR and more than 15% over the baseline SVR, thereby reinforcing its capability to effectively balance accuracy and error minimization. The regression plots further emphasized these differences, with CNN predictions tightly following the 45° reference line and SVR predictions diverging more noticeably. Residual plots revealed that more than 90% of the augmented CNN residuals consistently fell within ±1%, compared to only 75–80% for SVR, highlighting the augmented CNN's superior calibration and error control. Collectively, these findings validate the effectiveness of the proposed augmentation strategy and confirm the augmented CNN 1D model as the most robust and accurate framework for forecasting daily freshwater output in solar still systems.

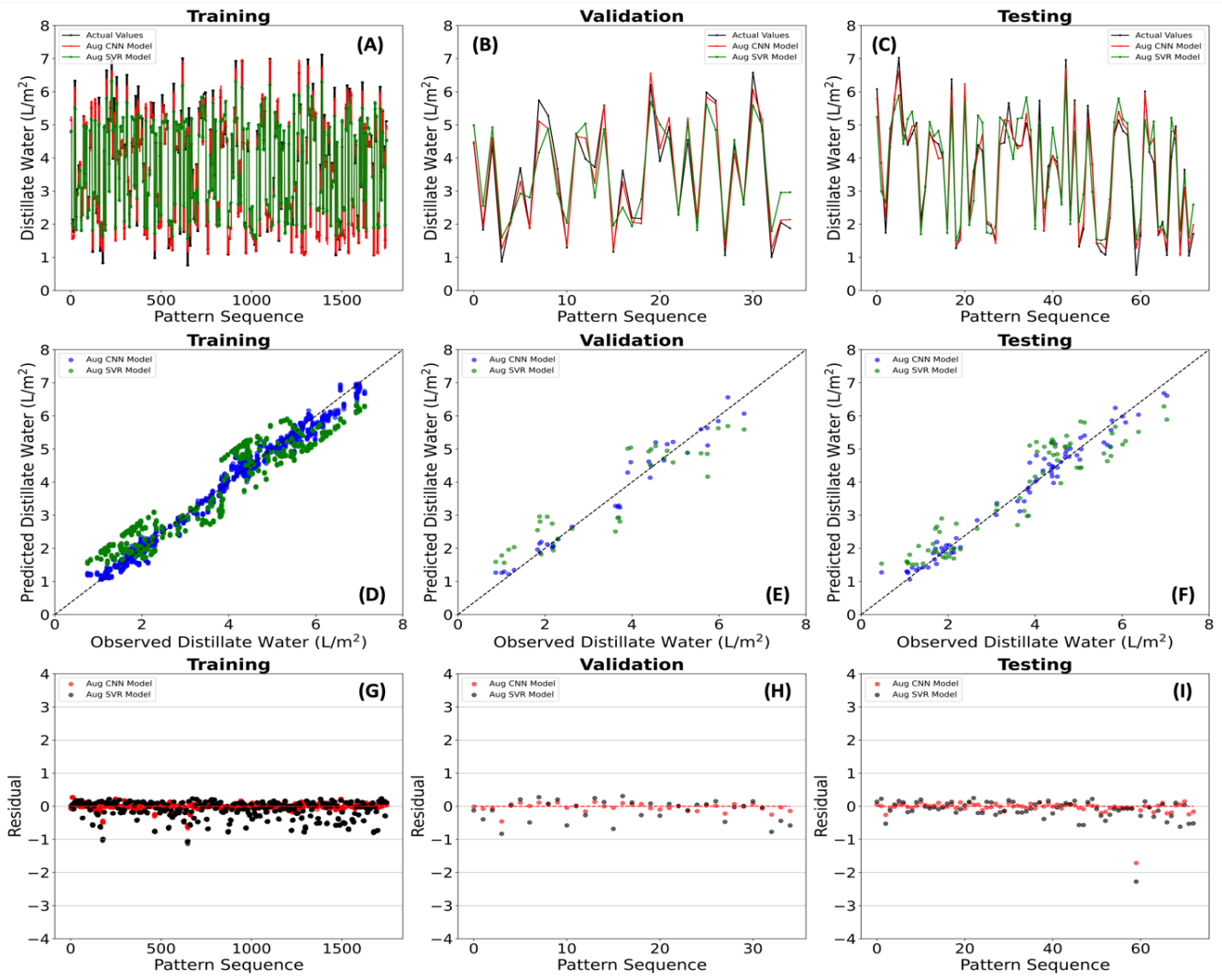


Figure 6: Augmented CNN-1D vs. SVR performance. Comparative performance of augmented CNN-1D and SVR, where CNN-1D demonstrates improved prediction accuracy and more compact residuals, especially during testing.

Figure 7 displays the residual heatmaps for the augmented CNN-1D and SVR models across the training, validation, and testing phases, offering a clear visualization of relative error magnitudes. The augmented CNN-1D model consistently shows lighter and more uniform color intensities, particularly during the testing phase, indicating its superior generalization ability and reduced residual dispersion. In the training phase (Fig. 7A), the relative error range for the CNN model spans approximately from 0.25 (light) to 1.75 (dark). In contrast, the validation (Fig. 7B) and testing (Fig. 7C) phases demonstrate narrower error ranges of 0.2 to 1.2 and 0.2 to 1.4, respectively. These lighter and more homogeneous patterns highlight the CNN model's ability to produce tightly bound residuals across all phases.

On the other hand, the SVR model exhibits darker and more scattered heatmaps, with pronounced color bands that suggest more prominent and more frequent prediction errors, especially during the validation and testing phases. The corresponding MAE values further validate these visual findings. The augmented CNN-1D achieved low MAE values of 0.023 for training, 0.035 for validation, and 0.035 for testing, which indicates consistent and minimal prediction errors.

Conversely, the SVR model yielded higher MAE values of 0.084 (training), 0.098 (validation), and 0.087 (testing), which aligns with the increased residual variance observed in the heatmaps. Additionally, the CNN model outperformed the SVR in overall performance metrics, achieving a lower test RMSE of 0.0814 and a higher R^2 of 0.903, compared to the SVR's test RMSE of 0.112 and R^2 of 0.841. These quantitative results and visual residual patterns demonstrate that the augmented CNN-1D model, enhanced through data augmentation and architectural optimization, provides more accurate predictions. Also, it maintains superior stability and generalization across all modeling phases compared to the SVR model.

The deviation histograms and kernel density estimation (KDE) plots presented in Figure 8 for the testing dataset provide a valuable analysis of prediction accuracy and consistency across the various models. The histogram for the Augmented CNN-1D model demonstrates a concentrated and symmetric distribution of deviations, with the highest frequency reaching 14 and deviation values closely clustered between -0.23 and 0.82. In contrast, the Baseline CNN-1D model exhibits a broader and more dispersed distribution, peaking at a frequency of only 7.9, while displaying a wider range of deviations from 0.25 to 1.22. This increased spread indicates a higher error variance and a greater frequency of significant deviations in the baseline model. Similarly, the KDE plots reinforce the superior generalization capabilities of the augmented CNN-1D model. Its distribution features a sharp, dominant peak with a maximum density of 1.8, suggesting that most predictions closely align with the actual values, with deviations ranging from -0.70 to 1.21. Conversely, the Baseline CNN-1D model reveals a flatter curve with a lower peak density of 1.25 and a broader range of deviations spanning from -0.75 to 1.5. These differences underscore the augmented model's ability to produce more consistent and less biased predictions, with fewer extreme errors, thereby highlighting its enhanced predictive stability and accuracy.

Table 9: Model performance comparisons. Training, validation, and testing results of CNN-1D and SVR (baseline vs. augmented). Augmented CNN-1D achieved the highest accuracy and best generalization.

Stage	**Training**					
Model / Metric	R^2	RMSE	MAE	CV	EC	OIMP
CNN-1D Baseline	0.91	0.08	0.06	0.17	0.91	0.91
CNN-1D Augmented	0.99	0.03	0.02	0.04	0.99	0.99
SVR Baseline	0.87	0.09	0.08	0.20	0.87	0.89
SVR Augmented	0.88	0.09	0.08	0.20	0.88	0.89
Stage	**Validation**					
Model / Metric	R^2	RMSE	MAE	CV	EC	OIMP
CNN-1D Baseline	0.87	0.09	0.07	0.22	0.87	0.91
CNN-1D Augmented	0.97	0.04	0.03	0.08	0.98	0.97
SVR Baseline	0.81	0.11	0.09	0.27	0.81	0.84
SVR Augmented	0.81	0.11	0.09	0.27	0.81	0.84
Stage	**Testing**					
Model / Metric	R^2	RMSE	MAE	CV	EC	OIMP
CNN-1D Baseline	0.86	0.09	0.07	0.22	0.86	0.88
CNN-1D Augmented	**0.971**	**0.04**	**0.03**	**0.08**	**0.98**	**0.97**
SVR Baseline	0.85	0.10	0.09	0.23	0.84	0.87
SVR Augmented	0.853	0.10	0.08	0.22	0.85	0.88

Note: R^2: coefficient of determination, RMSE: root mean square error, MAE: mean absolute error, CV: coefficient of variance, EC: efficiency coefficient, OIMP: overall index of model performance.

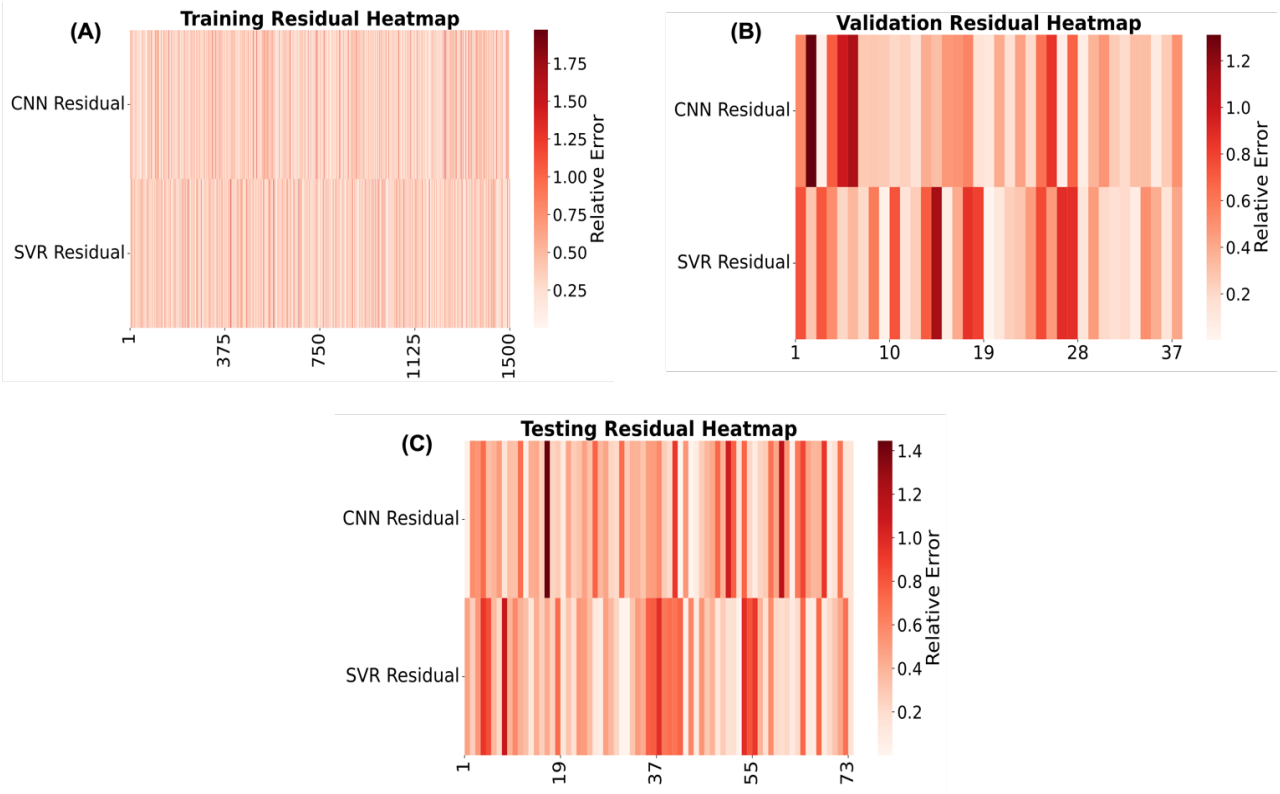


Figure 7: Residual heatmaps of augmented CNN-1D and SVR. Error distributions for augmented CNN-1D and SVR during training, validation, and testing, showing CNN-1D residuals clustered around zero and SVR residuals more scattered.

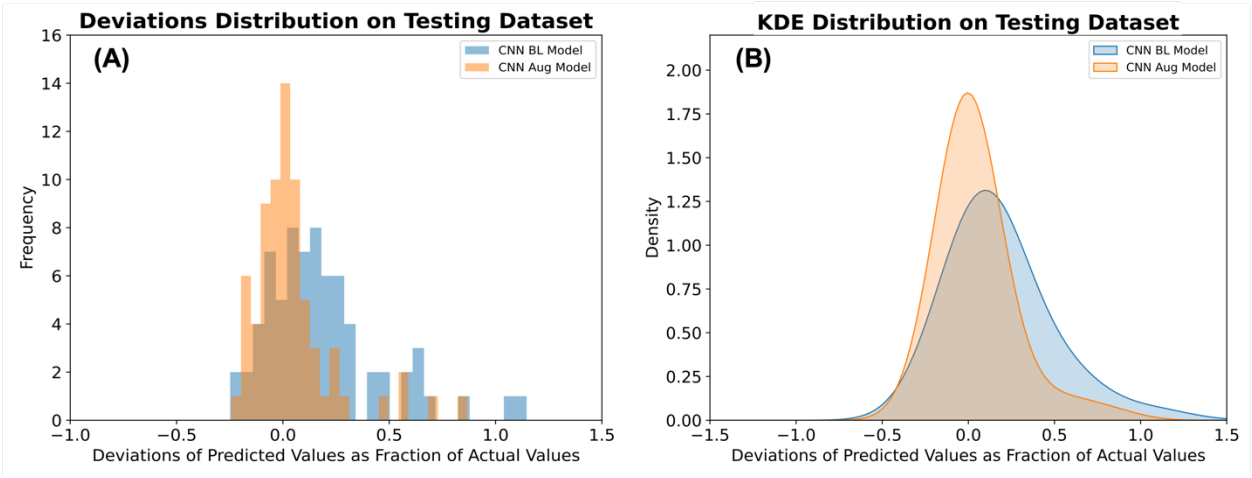


Figure 8: Error distributions in testing data. Histograms and kernel density plots of prediction deviations for baseline and augmented CNN-1D, with augmentation producing sharper error concentration around zero.

The integrated analysis of Figure 9 and Figure 10 presents a comprehensive comparative evaluation of four developed models: augmented CNN-1D, augmented SVR, baseline CNN-1D, and baseline SVR. This evaluation is predicated upon six critical statistical metrics assessed during the testing phase. The augmented CNN-1D exhibited the most consistent and superior performance across all identified metrics among the examined models. As illustrated by both figures, it attained the highest accuracy-related scores—$R^2 = 0.97$, EC = 0.98, and OIMP = 0.97—indicative of nearly

perfect concordance between predicted and observed values and a highly efficient model structure. In stark contrast, the baseline SVR model recorded the lowest values across these metrics (R^2 = 0.85, EC = 0.85, OIMP = 0.87), underscoring its limited capability to generalize and accurately capture the underlying patterns within the data. The radar plots vividly represent these distinctions, where the augmented CNN-1D model delineates the largest and most balanced polygon.

On the contrary, the SVR model demonstrates a considerably smaller and irregular shape, reflecting diminished predictive power and more significant variability. Further validation of these findings is provided through the error-based metrics. The augmented CNN-1D achieved the lowest RMSE (0.04) and MAE (0.04), indicating a significant reduction of approximately 60% in RMSE and 55% in MAE compared to both baseline models. Conversely, while exhibiting improvements relative to its baseline counterpart, the augmented SVR still manifested significantly higher error values (RMSE = 0.10, MAE = 0.09) and greater variability (CV = 0.23). Unlike the augmented CNN-1D, it maintained a substantially lower CV of 0.08, signifying more stable and less dispersed predictions. The radar plots also illustrated these discrepancies, where the augmented CNN-1D model exhibited tight clustering and minimal distortion along the error axes. Meanwhile, the SVR-based models displayed considerable asymmetry and extended error regions. Collectively, the histogram and radar plots provide robust evidence that the augmented CNN-1D model achieves the highest predictive accuracy and demonstrates the most robust and reliable performance during the testing phase. These results underscore the efficacy of the proposed time series data augmentation strategy and further substantiate the CNN-1D architecture as a highly suitable and generalizable framework for forecasting freshwater output in solar still desalination systems.

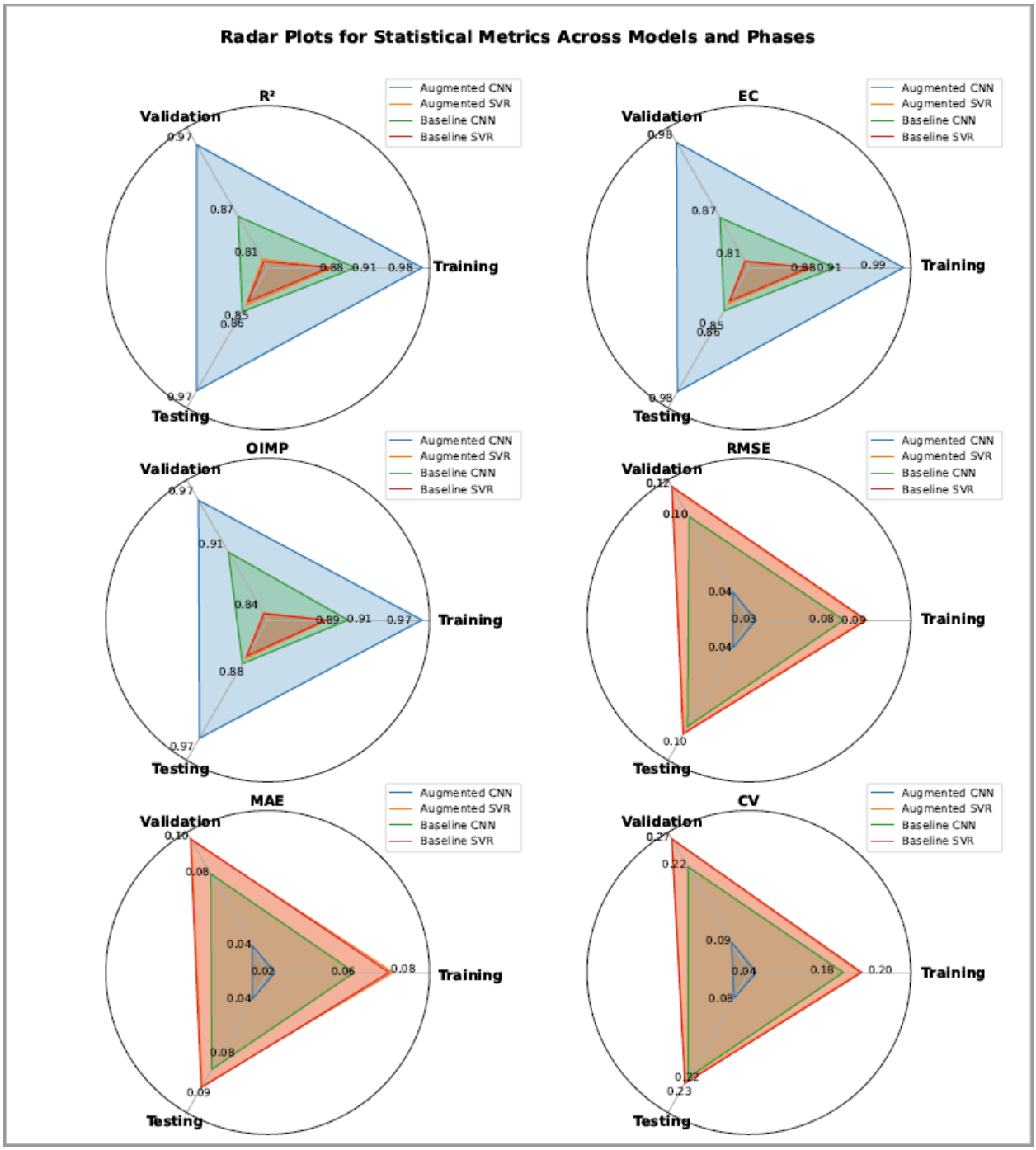


Figure 9: Radar plots of statistical metrics. Comparative radar charts of R^2, EC, OIMP, RMSE, MAE, and CV across phases for augmented CNN-1D, augmented SVR, baseline CNN-1D, and baseline SVR, showing augmented CNN-1D as the best performer.

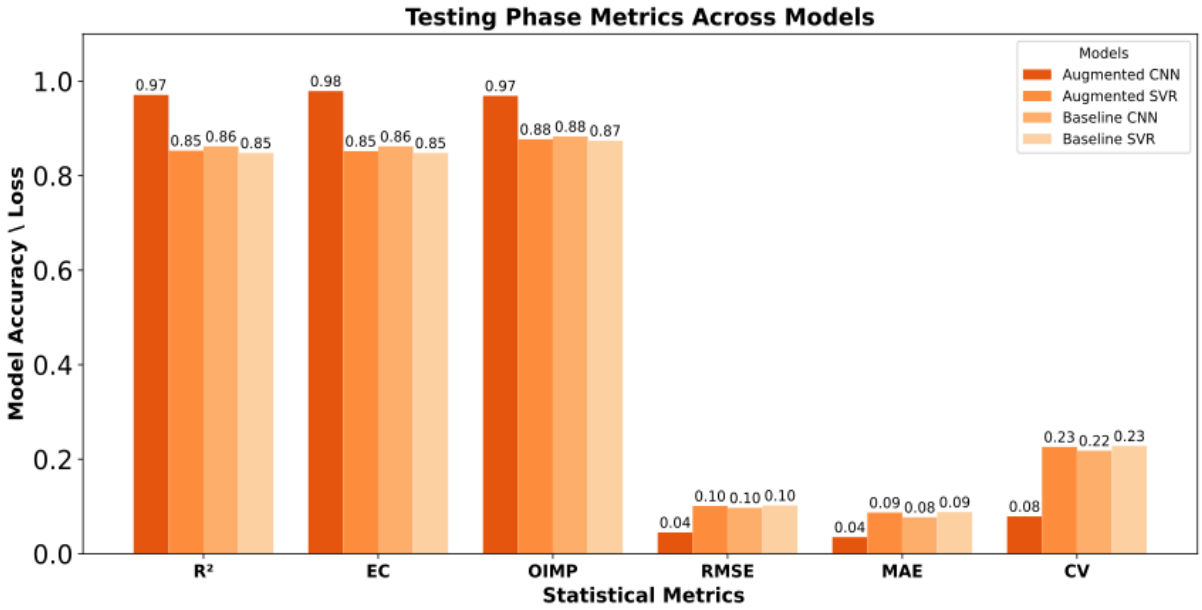


Figure 10: Testing phase metrics comparison. Bar chart comparing R^2, EC, OIMP, RMSE, MAE, and CV for all models in the testing phase, with augmented CNN-1D outperforming others in accuracy and error reduction.

Model performance on unseen climatic conditions was assessed using an independent test set, which contained meteorological states that are not present in the training set. The augmented CNN-1D achieved an $R^2 = 0.97$, an RMSE of 0.04, and an MAE of 0.03 on this set, confirming strong generalization. Residual analyses further showed that CNN-1D predictions remained stable across both frequent and less common conditions, while SVR exhibited greater error variance in extremes such as high irradiance or humidity. Beyond augmentation, transfer learning has been investigated as an additional strategy for improving robustness. In our previous study [32], the same solar still dataset was used to train a target ANN model (Python, randomly initialized) with transferred weights from a physics-based source ANN model developed in MATLAB/Simulink. This approach demonstrated the potential of combining data-driven and physics-based knowledge to improve generalization under variable conditions. Building on these findings, future work will focus on physics-informed neural networks (PINNs), which embed governing equations directly into the architecture, thereby enhancing extrapolation capacity under entirely new climatic regimes.

The comparative evaluation of the developed CNN-1D and SVR models was primarily conducted using the testing dataset, with performance assessed through a comprehensive statistical analysis, as shown in Figure 10. The results for freshwater prediction indicate that the CNN-1D model surpasses the SVR model in terms of accuracy, robustness, and generalization capability. This distinction is particularly significant when modeling complex, nonlinear interactions among meteorological inputs and distillate output, which are frequently encountered in solar desalination systems. This study highlights the importance of comparing these models for several reasons. First, SVR, while recognized for its non-linear regression capabilities, serves as a valuable benchmark for evaluating the enhanced ability of the deep learning model to capture temporal and spatial dependencies. Second, using CNN-1D enables hierarchical feature learning from time-series data, allowing it to capture intricate dynamics that are often oversimplified or overlooked by kernel-based methods, such as SVR.

Lastly, validating the performance of the CNN-1D model against SVR strengthens the robustness of the proposed framework, as the observed improvements are not only statistically significant but

also practically applicable to real-world solar still forecasting tasks. These findings are consistent with previous research studies [31] [32], demonstrating that deep learning architectures, including ANNs, outperform traditional machine learning approaches like MLR through the transfer learning approach in predicting distillate outputs.

Although CNN-1D achieved superior predictive accuracy in this study, several inherent weaknesses should be noted. First, CNN-1D models are highly data-dependent, and their stability decreases with small datasets; this limitation necessitates the introduction of data augmentation to mitigate the risks of overfitting. Second, CNN-1D primarily captures local temporal patterns through convolutional filters, making it less effective than recurrent architectures (e.g., LSTM, GRU) in scenarios requiring long-term dependency modeling. Finally, CNN-1D operates as a "black-box" model, offering limited interpretability of learned features relative to physics-based or statistical models. These limitations suggest that while CNN-1D is a strong choice under the conditions of this study, future research should explore hybrid approaches—such as physics-informed neural networks, or explainable AI frameworks—to mitigate these shortcomings and further enhance model robustness.

This study was conducted under several assumptions. The dataset of 365 daily samples was assumed sufficient to capture seasonal patterns in productivity. Feedwater flow rate, not explicitly measured in the Santos dataset, was estimated from efficiency correlations. Gaussian noise $\mathcal{N}(0, 0.01^2)$ was assumed to represent realistic meteorological variability, and augmentation was applied exclusively to the training set to avoid biasing validation or testing. Despite the strong results, several limitations should be acknowledged. The dataset was restricted to a single year of daily resolution, excluding inter-annual and hourly variations. Input features were limited to five non-redundant drivers, with clearness index and day length excluded to avoid collinearity with irradiance and seasonality. While CNN-1D effectively modeled local temporal dependencies, it may underperform in capturing longer-term memory compared to recurrent architectures. Finally, the model's ability to generalize to entirely new climatic regimes is limited and would require broader datasets or hybrid strategies such as transfer learning or physics-informed neural networks (PINNs). At the same time, this study addressed two critical research gaps through data augmentation. First, it systematically applied Gaussian noise with a look-back window to enrich the training space and reduce overfitting in solar still prediction. Second, it demonstrated that augmentation enables the practical training of deep learning models even with limited data, enhancing forecasting accuracy and supporting downstream analyses such as productivity classification. Together, these strategies mitigate the small-sample limitation and provide a scalable foundation for future research. The results confirm that advanced data-driven methods can serve as a powerful complement to experimental studies of solar stills. By enabling accurate daily forecasting of freshwater productivity, the augmented CNN-1D framework supports optimization of energy conversion efficiency in real-world applications.

4. Recent Advancements and Insights

Recent innovations in solar desalination have concentrated on improving thermal performance and predictive accuracy through experimental modifications and machine learning techniques. These advancements offer valuable insights into cutting-edge technologies, underscoring their importance in developing next-generation solar desalination solutions. Several experimental investigations have explored thermal enhancement strategies and their effects on freshwater yield, providing context for these breakthroughs. For instance, Issaq et al. [33] explored performance enhancement strategies for single-inclination solar stills through controlled trials focused on basin water depth, sensible heat storage, and inner glass surface treatment. Their findings revealed that

thermal efficiency exceeded 40% under favorable weather without modifications, providing a strong baseline for comparison. However, increasing water depth beyond 6 cm significantly reduced yield and efficiency. Similarly, adding black gravel beyond a 2% gravel-to-water mass ratio showed negligible thermal improvements, while treating the inner glass surface with waxy materials led to a marked decline in performance. These results emphasize the sensitivity of solar still output to subtle design and material adjustments.

Mahala et al. [34]thoroughly evaluated the performance enhancement of pyramid solar stills by incorporating rectangular fins, phase change material (PCM), and gravel. They developed and tested two configurations: a conventional solar still (CSS) and a modified solar still (MSS), along with variations that included gravels (CSS + G and MSS + G). This study, conducted under actual climatic conditions in Greater Noida, India, provided insights into improvements in productivity, energy, and exergy efficiency, as well as the impact of economic and environmental factors. The findings revealed that the MSS + G configuration delivered superior performance, achieving an impressive 84% increase in daily freshwater output, along with enhancements in energy and exergy efficiency of 81.1% and 273%, respectively, compared to the CSS. The MSS + G configuration also led to a 29.2% reduction in cost per liter and a 29% shorter payback period. Environmentally, it resulted in a 76% increase in carbon credit gains and mitigated 19.14 tons of CO_2 emissions. These significant findings underscore the effectiveness of integrating thermal and material enhancements in developing cost-efficient and sustainable solar desalination systems.

Complementing these efforts, Elsheikh et al. conducted a study to evaluate the thermal and predictive performance of a stepped solar still equipped with a copper corrugated absorber plate, comparing it to a traditional single-slope design. The research focused on key thermal parameters, such as convection, evaporation, radiation heat transfer coefficients, and energy and exergy efficiencies. The experimental findings highlighted that the stepped solar still significantly enhanced freshwater production, achieving approximately 128% more output than the conventional system. To forecast hourly freshwater yield, the authors implemented a Long Short-Term Memory (LSTM) neural network, trained on time-series data from field experiments. They assessed the predictive accuracy of the LSTM model against that of the traditional ARIMA model. The results revealed that the LSTM model exhibited remarkable forecasting precision, with coefficients of determination of 0.97 for the conventional still and 0.99 for the stepped design. These findings underscore the advantages of combining innovative absorber designs with deep learning models to optimize real-time prediction and performance in solar distillation systems [35].

The studies reviewed indicate that integrating advanced materials, design optimizations, and AI-based models significantly boosts the performance of solar stills. These findings emphasize the crucial role of data-driven approaches, such as the proposed augmented CNN-1D model, in shaping the future of sustainable desalination technologies.

4.1 Performance Analysis

The performance comparison in Table 10 elucidates the significant advancements accomplished by the proposed augmented CNN-1D solar still model when benchmarked against a diverse range of recent experimental and machine learning-based studies. This table consolidates various desalination configurations, including pyramid stills with wick materials, systems enhanced by external condensers or phase change materials (PCM), and alternative geometrical designs such as tubular and hemispherical stills. Furthermore, it reflects the growing adoption of predictive modeling, wherein artificial neural networks and hybrid optimization algorithms are employed to

forecast system performance. This transition towards data-driven methodologies complements traditional thermal enhancements and often yields superior results. A distinctive feature of the proposed CNN-based model is its utilization of time series data augmentation, which facilitates high prediction accuracy without requiring physical modifications to the system. This comparative analysis exemplifies the potential of intelligent modeling techniques in solar desalination and advocates for their implementation as scalable and cost-effective solutions for optimizing freshwater production. These findings offer crucial insights for researchers and engineers aiming to improve the efficiency, adaptability, and long-term sustainability of solar still technologies.

Table 10: Benchmarking augmented CNN-1D with prior solar still studies. Performance comparison of the proposed augmented CNN-1D model against selected recent studies, including experimental, theoretical, simulation, and machine-learning approaches.

Study Type / Design	**Achieved Efficiency / Productivity**	**Methodology**	**Ref.**
Pyramid Solar Still (Wick Type with Reflectors)	55.3% Efficiency, 1476 mL/m²·d Productivity; up to 136.6% improvement	Experimental	[36]
Wick Material Solar Still with External Condenser	27.13% Efficiency, Productivity improved by up to 75%	Experimental + CFD Simulation	[37]
Pyramid Solar Still (PSS)	45.7% Efficiency (implied), 2.585 kg/m² Productivity	Optimization (ANN+RSM) Models	[38]
Single-Slope Solar still (Modified with External Heater)	PCM: Prod. increase ~30.6%, Vaporizer: 415% yield increase	Experimental (HFU Vaporizer + PCM capsules)	[39]
Hemispherical Solar Distillers (Sponge-Enhanced)	Exergy and Energy Efficiencies were improved ~512.87 ,70.53%, respectively.	Experimental (Material Enhancement)	[40]
Passive Solar Thermal Desalination	32.9% Efficiency, 0.5 kg/m²·h Productivity; achieved 168 h (7 days).	Experimental (continuous salt-free operation)	[41]
Wick Solar Still (Conventional & developed)	~50% Productivity Increase; ANN-TSA model achieved up to R^2=0.99	ML (ANN + Tree–Seed Algorithm)	[42]
Active Solar Still with Condenser	53.21% productivity increased at a fan speed of 1350 rpm.	ML (ANN + Harris Hawks Optimizer)	[43]
Tubular Solar Still Distiller (Case III)	83.69% Energy Efficiency, 242.45% Exergy Efficiency, 79.88% Productivity increase	Experimental	[44]
Novel Solar Desalination	65.48% Energy Efficiency, 6.67% Exergy Efficiency, 20.95 L/m². day Productivity	Experimental (Heat Pump + Evacuated Tube)	[45]
Tubular Solar Still (Double-Trough)	58.5% Thermal Efficiency, 36.1% Productivity Improvement	Experimental (Multi-Evaporator Design)	[46]
Tubular Solar Still (TSS)	Fresh water production overall cost $0.0061-$0.2 per kg water	Theoretical Review: Economic Analysis	[47]

Hemispherical Solar Still (HSS-GB)	25.75% Thermal Efficiency increase, 26 % more distilled water with 6.30 L/m^2 yield	Experimental (Glass Bottle)	[48]
Hemispherical Solar Still (HSS-N)	Daily increase 6.8 L/m^2 with 76.6% Productivity Improvement	Optimization (Best Modification)	[49]
Solar Still with hybrid Nanofluid (MSS)	27.2% daily enhanced productivity in summer, 29.6 % CO_2 reduction in winter	Theoretical +Experimental	[50]
Conventional Single Slope Solar Still	R^2 = 97.1%, OIMP = ~97%, RMSE = 0.045, MAE = 0.035, (Aug. CNN 1D Testing Phase)	ML-Data Augmentation (CNN + SVR Models)	*Current Study*

5. Conclusion

This study developed and rigorously evaluated a data-efficient forecasting framework for predicting daily freshwater productivity (P_{std}) in a conventional single-slope solar still, directly addressing the persistent challenge of limited experimental data. By integrating a one-dimensional convolutional neural network (CNN-1D) with Gaussian noise–based time-series augmentation, the approach demonstrated how small-sample constraints can be effectively mitigated while ensuring model robustness. The methodological pipeline was designed to maximize predictive accuracy and generalization capacity. A 7-day look-back window was employed to restructure the raw dataset into a sequential time-series format, capturing the temporal dependencies between meteorological drivers and distillate yield. Data augmentation was performed using Gaussian noise $\mathcal{N}(0, 0.01^2)$, applied exclusively to the training set to enrich variability while preserving signal fidelity. Hyperparameter optimization (HPO) through a Tree-structured Parzen Estimator (TPE) search systematically tuned learning rate, kernel size, batch size, and regularization strength. The optimized CNN-1D architecture, comprising three convolutional layers (128 filters, ReLU activation), was trained using feed-forward backpropagation to extract complex spatial–temporal features autonomously.

Comparative analyses established a clear hierarchy of model performance: augmented CNN-1D > baseline CNN-1D > augmented SVR > baseline SVR**.** Under identical splits (70% training, 10% validation, 20% testing), the augmented CNN-1D achieved the strongest performance with an EC of 0.98, OIMP =0.97, RMSE = 0.04, MAE = 0.03, and CV = 0.08, confirming both high predictive accuracy and minimal residual dispersion. In contrast, the augmented SVR model, despite benefiting from Gaussian noise and kernel optimization (linear, polynomial, RBF, and sigmoid), achieved more moderate results (Test R^2 = 0.85, RMSE = 0.10, OIMP = 0.88). This performance gap underscores the structural limitations of margin-based learning and static support-vector mechanisms, which are less responsive to noise-induced variability than the hierarchical feature extraction achieved by CNN-1D. Nevertheless, SVR provided a valuable and computationally efficient benchmark, capturing key nonlinearities and demonstrating reasonable generalization. Diagnostic analyses reinforced these findings. For CNN-1D, observed vs. predicted plots showed tightly clustered predictions along the 1:1 reference line, while residual analyses indicated homoscedasticity and minimal bias. By contrast, the SVR model displayed wider residual dispersion, particularly under atypical climatic conditions, confirming its reduced sensitivity to fine-scale meteorological variability. From a practical perspective, these outcomes provide a scalable and computationally efficient forecasting tool for predictive monitoring and operational

optimization of solar stills. Importantly, they demonstrate that augmenting training data and leveraging deep learning architectures significantly improves predictive reliability compared to traditional machine learning approaches. The study is subject to several limitations, including reliance on a single year of daily data, exclusion of potentially relevant variables such as clearness index and day length, and restriction to CNN-1D and SVR benchmarks. These limitations create opportunities for future work. Planned directions include the use of multi-year and higher-resolution datasets, the application of advanced augmentation strategies such as generative adversarial networks (GANs), and the integration of physics-informed neural networks (PINNs) to enhance both generalization and physical interpretability.

In summary, this research demonstrates that combining data augmentation with hierarchical feature extraction in a CNN-1D framework offers a robust, scalable, and data-efficient strategy for solar still forecasting. The augmented CNN-1D consistently outperformed SVR and other benchmarks, but the comparative role of SVR highlights the importance of including traditional models for fair benchmarking. Together, these results underscore the transformative role of advanced data-driven modeling in enabling sustainable freshwater management and system optimization in resource-constrained arid regions.

Nomenclature

P_{std}	—	Freshwater productivity, L/day
Is	—	Solar Radiation, MJ/day
T_{amb}	—	Ambient temperature, ^{o}C
V_w	—	Wind speed, m/s
RH	—	Relative humidity, %
P_{st}	—	Saline water feed flowrate, L/day

List of Abbreviations

ML	—	Machine Learning
AI	—	Artificial Intelligence
CNN	—	Convolutional Neural Network
1D	—	One Dimensional
SVR	—	Support Vector Regression
MENA	—	Middle East and North Africa
MSF	—	Multi-Stage Flash
MED	—	Multi-Effect Distillation
PCM	—	Phase Change Materials
SWRO	—	Sea Water Reverse Osmosis
SEC	—	Specific Energy Consumption
GJO	—	Golden Jackal Optimization
ALSS	—	Aluminum-based Solar Still
PCSS	—	Polycarbonate-based Solar Still
NEPCMS	—	Nano-enhanced Phase Change Materials
SVM	—	Support Vector Machine
CPL	—	Cost per Liter

DA	—	Data Augmentation
BL	—	Baseline
KDE	—	Kernel Density Estimator
ReLU	—	Rectified Linear Unit
L2	—	Regularization Strength
R^2	—	Coefficient of determination
RMSE	—	Root mean square error
MAE	—	Mean absolute error
MSE	—	Mean squared error
EC	—	Efficiency coefficient
CV	—	Coefficient of variation
OIMP	—	Overall index of model Performance
RF	—	Random Forest
MAPE	—	Mean Absolute Percentage Error

Appendix

Mathematical Symbols Definition

Symbol Name		**Definition/ Meaning**
y_{min}	—	Minimum of input and output variables.
y_{max}	—	Maximum of input and output variables.
$y_{o,i}$	—	Observed input and output values.
$y_{p,i}$	—	Predicted value by the machine learning models.
$\bar{y}_o$	—	Averaged observed/ mathematically obtained values.
$\bar{y}_p$	—	Predicted values/ theoretically estimated from averaging.
n	—	Number of observations or dataset size.
X_{data}	—	Original input feature matrix (365 × 5)
y_{data}	—	Original output column (365 × 1)
$X_{ts}[i]$	—	Time-series input with look-back window = 7
$y_{ts}[i]$	—	Output value corresponding to each time-window sample
X_{train}, y_{train}	—	Training data split (70%)
X_{val}, y_{val}	—	Validation data split (10%)
X_{test}, y_{test}	—	Testing data split (20%)
$\tilde{X}_{train}[i,j]$	—	Augmented input sample using Gaussian noise
$\tilde{y}_{train}[i,j]$	—	Ground truth label (unchanged during augmentation)
$\mathcal{N}(0,\sigma^2)$	—	Gaussian noise with mean 0 and variance σ²
X_{scaled}, y_{scaled}	—	Normalized features and labels using Min-Max scaling
N	—	Total number of training instances
$Z^{(l)}[i,j]$	—	Output of convolution before activation
$A^{(l)}[i,j]$	—	Activation output at time step $[i,j]$ from the current layer
$A^{(l-1)}[i+k]$	—	Activation output at time step $[i,j]$ from the previous layer
$A^{(0)} = x$	—	Original input time series to the first convolutional layer

$w^{(l)}[k,j]$	—	Weight of kernel position k for filter j
$b^{(l)}[j]$	—	Bias term for filter j
P	—	Padding size
L_{in}	—	Input sequence length
L_{out}	—	Output sequence length
S	—	Stride (1)
K	—	Kernel size (3)
k	—	Kernel index from 0 to $K-1$ ($K=3$)
i	—	Time index (temporal position)
j	—	Filter index (1 to 128)
l	—	Layer index (l = 1,2,3)
$f \in \mathbb{R}^M$	—	Flattened feature vector, M = T × filters
h_m	—	Output from the m-th neuron in the dense hidden layer
$\max(0,.)$	—	ReLU activation function.
W_{mn}	—	Weight connecting n-th feature to m -th neuron in the dense hidden layer.
b_m	—	Bias term associated with the m-th neuron in the dense hidden layer.
f_n	—	n-th input value from the flattened layer $f \in \mathbb{R}^M$
M	—	Total number of neurons in the dense hidden layer (128).
$\hat{y}$	—	Predicted scalar output value from the network (water productivity).
$w_m^{(0)}$	—	Weight connecting m-th neuron in the dense hidden layer to output layer.
$b^{(0)}$	—	Bias term in the output layer.
η	—	Learning rate ∈ [10^{-4}, 10^{-2}]
λ	—	Regularization strength ∈ [10^{-4}, 10^{-2}]

Acknowledgment

The authors, H. H. Migaybil and B. Gopaluni, gratefully acknowledge the invaluable support and guidance provided by the Data Analytics and Intelligent Systems (DAIS) Laboratory team and the Chemical and Biological Engineering Department at the University of British Columbia. The authors also wish to convey their heartfelt appreciation to the esteemed Deanship of Scientific Research (DSR) at King Abdulaziz University for their generous funding and support, which played a pivotal role in facilitating this research endeavor.

Conflicts of Interest

There are no conflicts of interest to declare by the authors, considering the publication of this paper.

Data Availability

The data used in this study to support the findings are available upon request from the corresponding authors.

References

[1] UN-Water, "The United Nations World Water Development Report 2024: Water for Prosperity and Peace," Paris, France: UNESCO, 2024.

[2] B. OECD, "OECD Environmental Outlook to 2050: The Consequences of Inaction," Paris, France: OECD Publishing, Mar. 2012.

[3] UNICEF and WHO, "Progress on Household Drinking Water, Sanitation and Hygiene 2021 Update and SDG Baselines," World Health Organization, Geneva, 2021.

[4] UNESCO, "World Water Development Report 2023: Partnerships and Cooperation for Water," United Nations Educational, Scientific and Cultural Organization, Paris, 2023.

[5] NEOM, "NEOM's Vision for Water: Pioneering Sustainable Solutions," *NEOM*, 2024. [Online]. Available: https://www.neom.com/en-us/sectors/water. [Accessed: May. 5, 2025].

[6] M. A. Eltawil, Z. Zhengming, and L. Yuan, "A review of renewable energy solutions for desalination," *Desalination*, vol. 245, pp. 457-464, 2009.

[7] A. Giwa, S. W. Hasan, N. Yousuf, S. Chakraborty, D. J. Johnson, and N. Hilal, "Biomimetic and bioinspired membranes: A review of their status and potential for desalination and water treatment," *Desalination*, vol. 499, pp. 114753, Feb. 2021.

[8] M. Shokri and A. Sanavi Fard, "A comprehensive overview of environmental footprints of water desalination and alleviation strategies," *Int. J. Environ. Sci. Technol.*, vol. 20, no. 2, pp. 2347–2374, 2023.

[9] Y. Hong, S. Park, K. Kim, A. B. Alayande, and J. Kim, "Seawater Reverse Osmosis (SWRO) Desalination: Energy consumption in plants, advanced low-energy technologies, and future developments for improving energy efficiency," *Renew. Sustain. Energy Rev.*, vol. 180, p. 113212, 2023.

[10] Y. G. Kim, J. Byun, K. Park, and K. Park, "Comprehensive analysis of energy saving and high-quality permeate production strategies for a large-scale seawater reverse osmosis desalination plant with diverse process configurations and external resource utilization," *Desalination*, vol. 596, p. 118292, 2025.

[11] S. A. Kalogirou, "Solar thermal collectors and applications," *Prog. energy Combust. Sci.*, vol. 30(3), pp. 231–295, 2004.

[12] F. A. Omara, Z. M. Kabeel, A. E. Abdullah, A. S. Essa, "Experimental investigation of corrugated absorber solar still with wick and reflectors," *Desalination*, vol. 381, pp. 111–116, 2016.

[13] R. V. Arunkumar, K. R. Kaiwalya, D. D. Winfred Rufuss, D. Denkenberger, G. Tingting, and L. Xuan, "A review of efficient high productivity solar stills," *Renew. Sustain. Energy Rev.*, vol. 101, pp. 197–220, 2019.

[14] Z. M. Hammoodi, K. A. Dhahad, H. A. Alawee, and W. H. Omara, "Enhancement of pyramid solar still productivity through wick material and reflective applications in Iraqi conditions," *Math. Model. Eng. Probl.*, vol. 10, no. 5, pp. 1258–1264, 2023.

[15] A. N. Mohammed, A. H. Attalla, M. Shmroukh, "Performance enhancement of single-slope solar still using phase change materials," *Environ. Sci. Pollut. Res.*, vol. 28, no. 14, pp. 17098–17108, 2021.

[16] N. Wang, A. W. Kandeal, A. Swidan, S. W. Sharshir, G. B. Abdelaziz, M. A. Halim, and Y. Yang, "Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm," *Appl. Therm. Eng.*, vol. 184, p. 116233, 2021.

[17] S. S. Ghandourah, E., Prasanna, Y. S., Elsheikh, A. H., Moustafa, E. B., Fujii, M., & Deshmukh, "Performance prediction of aluminum and polycarbonate solar stills with air cavity using an optimized neural network model by golden jackal optimizer," *Case Stud.*

Therm. Eng., vol. 47, p. 103055, 2023.

[18] N. Shoeibi, S., Kargarsharifabad, H., & Rahbar, "Effects of nano-enhanced phase change material and nano-coated on the performance of solar stills," *J. Energy Storage*, vol. 42, p. 103061, 2021.

[19] N. I. Santos, A. M. Said, D. E. James, and N. H. Venkatesh, "Modeling solar still production using local weather data and artificial neural networks," *Renew. Energy*, vol. 40, no. 1, pp. 71–79, 2012, doi: 10.1016/j.renene.2011.09.018.

[20] N. H. Venkatesh, "Performance evaluation of single and double-basin solar stills in Las Vegas, Nevada," M.S. thesis, Univ. of Nevada, Las Vegas, NV, USA, 2012.

[21] NASA, "NASA Prediction Of Worldwide Energy Resource (POWER)," *Https://Power.Larc.Nasa.Gov/Data-Access-Viewer/*, 2020. https://power.larc.nasa.gov.

[22] O. Burn, M. Hoang, D. Zarzo, F. Olewniak, E. Campos, B. Bolto, and G. Barron, "A review of the opportunities for desalination in agriculture," *Desalination*, vol. 364, pp. 2–16, 2015.

[23] H. H. Migaybil and H. A. Maddah, "Design and simulation of a novel solar photovoltaic system assisted a single-slope solar still distillation unit," *Can. J. Chem. Eng.*, Accepted for publication, 2022.

[24] B. K. Rahman, M. M., & Bala, "Modelling of jute production using artificial neural networks," *Biosyst. Eng.*, vol. 105(3), pp. 350–356, 2010.

[25] A. Zangeneh, M., Omid, M., & Akram, "A comparative study between parametric and artificial neural networks approaches for economical assessment of potato production in Iran," *Spanish J. Agric. Res.*, vol. 9(3), pp. 661–671, 2011.

[26] P. O. Tang, J., Xia, H., Aljerf, L., Wang, D., & Ukaogo, "Prediction of dioxin emission from municipal solid waste incineration based on expansion, interpolation, and selection for small samples," *J. Environ. Chem. Eng.*, vol. 10(5), p. 108314, 2022.

[27] P. Xia, H., Tang, J., Aljerf, L., Wang, T., Gao, B., Xu, Q., ... & Ukaogo, "Assessment of PCDD/Fs formation and emission characteristics at a municipal solid waste incinerator for one year," *Sci. Total Environ.*, vol. 883, p. 163705, 2023.

[28] M. T. Alazba, A. A., Mattar, M. A., ElNesr, M. N., & Amin, "Field assessment of friction head loss and friction correction factor equations," *J. Irrig. Drain. Eng.*, vol. 138(2), pp. 166–176, 2012.

[29] J. Brownlee, "Use early stopping to halt the training of neural networks at the right time," Machine Learning Mastery, Dec. 10, 2018. [Online]. Available: https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/. [Accessed: Jun. 11, 2025].

[30] P. W. Lee, *Machine Learning Projects for .NET Developers*, Birmingham, UK: Packt Publishing, 2019.

[31] A. A. Mashaly, A. F., & Alazba, "Neural network approach for predicting solar still production using agricultural drainage as a feedwater source," *Desalin. Water Treat.*, vol. 57(59), pp. 28646–28660, 2016.

[32] H. H. Migaybil and B. Gopaluni, "A performance neural network model for conventional

solar stills via transfer learning," *Appl. Energy*, vol. 375, p. 124118, 2024.

[33] A. A. Issaq, S. Z., Talal, S. K., & Azooz, "Experimentation on enhancement of solar still performance," *Int. J. Renew. Energy Dev.*, vol. 12(4), 2023.

[34] N. Mahala, T., & Sharma, "Experimental investigations of a novel solar still with heat storage materials-energy, exergy, economic and environmental analyses," *Desalination*, vol. 578, p. 117467, 2024.

[35] S. M. Elsheikh, A. H., Katekar, V. P., Muskens, O. L., Deshmukh, S. S., Abd Elaziz, M., & Dabour, "Utilization of LSTM neural network for water production forecasting of a stepped solar still with a corrugated absorber plate," *Process Saf. Environ. Prot.*, vol. 148, pp. 273–282, 2021.

[36] A. S. Hammoodi, K. A., Dhahad, H. A., Alawee, W. H., Omara, Z. M., Essa, F. A., & Abdullah, "Improving the performance of a pyramid solar still using different wick materials and reflectors in Iraq," *Desalin. Water Treat.*, vol. 285, pp. 1–10, 2023.

[37] A. O. Hyal, L. S., Jalil, J. M., & Hanfesh, "Numerical and Experimental Study of a Single-Slope Solar Still Integrated with Wick Material and External Condenser," *Int. J. Heat Technol.*, vol. 42(4), 2024.

[38] R. Yuvaperiyasamy, M., Senthilkumar, N., Deepanraj, B., Gokilakrishnan, G., & Premkumar, "Application of response surface methodology and neural networks in pyramid solar still for seawater desalination: An optimization and prediction strategy," *Glob. NEST J.*, vol. 26(4), p. 05773, 2024.

[39] M. H. Abed, A. H., Hoshi, H. A., & Jabal, "Experimental investigation of modified solar still coupled with high-frequency ultrasonic vaporizer and phase change material capsules," *Case Stud. Therm. Eng.*, vol. 28, p. 101531, 2021.

[40] R. Sathyamurthy, A. E. Kabeel, and E. H. Attia, "Influence of high porous sponges for improving the interfacial evaporation from hemispherical solar distillers," *Sci. Rep.*, vol. 13, p. 17210, 2023.

[41] Y. Babb, P. I., Ahmadi, S. F., Brent, F., Gans, R., Lopez, M. A., Song, J., ... & Zhu, "Salt-rejecting continuous passive solar thermal desalination via convective flow and thin-film condensation," *Cell Reports Phys. Sci.*, vol. 4(12), 2023.

[42] A. Sharshir, S. S., Abd Elaziz, M., & Elsheikh, "Augmentation and prediction of wick solar still productivity using artificial neural network integrated with tree–seed algorithm," *Int. J. Environ. Sci. Technol.*, vol. 20(7), pp. 7237–7252, 2023.

[43] A. H. Essa, F. A., Abd Elaziz, M., & Elsheikh, "An enhanced productivity prediction model of active solar still using artificial neural network and Harris Hawks optimizer," *Appl. Therm. Eng.*, vol. 170, p. 115020, 2020.

[44] M. O. A. Kandeal, A. W., El-Naggar, A. A., Sharaby, M. R., Sharshir, S. W., Swidan, A., Abdelaziz, G. B., ... & El-Samadony, "Augmentation of the tubular distiller performance via hot air injection from a parabolic trough collector, nanocoating, and nanofluid," *Sol. Energy*, vol. 277, p. 112743, 2024.

[45] Z. Sharshir, S. W., Kandeal, A. W., Joseph, A., Elsayad, M. M., Abdullah, A. S., Jang, S. H., ... & Yuan, "Assessment of thermoeconomic and thermoenvironmental impacts of a

novel solar desalination system using a heat pump, evacuated tubes, cover cooling, and ultrasonic mist," *Appl. Therm. Eng.*, vol. 254, p. 123869, 2024.

[46] M. M. Elashmawy, M., Nafey, A. S., Sharshir, S. W., Abdelaziz, G. B., & Ahmed, "Experimental investigation of developed tubular solar still using multi-evaporator design," *J. Clean. Prod.*, vol. 443, p. 141040, 2024.

[47] A. Kabeel, A. E., Harby, K., Abdelgaied, M., & Eisa, "A comprehensive review of tubular solar still designs, performance, and economic analysis," *J. Clean. Prod.*, vol. 246, p. 119030, 2020.

[48] A. Abdel-Aziz, M. M., Attia, M. E. H., & Bouabidi, "Boosting solar distillation performance with recycled transparent glass bottles in hemispherical designs," *Sep. Purif. Technol.*, vol. 365, p. 132643, 2025.

[49] G. B. Attia, M. E. H., Kabeel, A. E., Abdelgaied, M., & Abdelaziz, "A comparative study of hemispherical solar stills with various modifications to obtain modified and inexpensive still models," *Environ. Sci. Pollut. Res.*, vol. 28(39), pp. 55667–55677, 2021.

[50] A. A. El-Gazar, E. F. Yousef, A. M. Elshaer, M. A. Khattab, T. A. Mouneer, and M. A. Hawwash, "Enhancing solar still performance with hybrid nanofluid: A comprehensive assessment of energy, exergy, economics, and environmental impact using a novel fractional model," *Environ. Dev. Sustain.*, pp. 1–28, 2024.