water

MDPI

Article

Application of a Multi-Teacher Distillation Regression Model Based on Clustering Integration and Adaptive Weighting in Dam Deformation Prediction

Fawang Guo [1], Jiafan Yuan [2,*], Danyang Li [2] and Xue Qin [2]

[1] POWERCHINA Guiyang Engineering Corporation Limited, Guiyang 550000, China; guofw_gyy@powerchina.cn
[2] College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China; danyangcl@163.com (D.L.)
* Correspondence: jiafanyuancs@163.com

Abstract: Deformation is a key physical quantity that reflects the safety status of dams. Dam deformation is influenced by multiple factors and has seasonal and periodic patterns. Due to the challenges in accurately predicting dam deformation with traditional linear models, deep learning methods have been increasingly applied in recent years. In response to the problems such as an excessively long training time, too-high model complexity, and the limited generalization ability of a large number of complex hybrid models in the current research field, we propose an improved multi-teacher distillation network for regression tasks to improve the performance of the model. The multi-teacher network is constructed using a Transformer that considers global dependencies, while the student network is constructed using Temporal Convolutional Network (TCN). To improve distillation efficiency, we draw on the concept of clustering integration to reduce the number of teacher networks and propose a loss function for regression tasks. We incorporate an adaptive weight module into the loss function and assign more weight to the teachers with more accurate prediction results. Finally, knowledge information is formed based on the differences between the teacher networks and the student network. The model is applied to a concrete-faced rockfill dam located in Guizhou province, China, and the results demonstrate that, compared to other knowledge distillation methods, this approach exhibits higher accuracy and practicality.

Keywords: dam deformation prediction; deep learning; knowledge distillation; cluster ensemble; information entropy

Academic Editors: Jie Yang, Lin Cheng and Chunhui Ma

Received: 25 February 2025
Revised: 18 March 2025
Accepted: 25 March 2025
Published: 27 March 2025

Citation: Guo, F.; Yuan, J.; Li, D.; Qin, X. Application of a Multi-Teacher Distillation Regression Model Based on Clustering Integration and Adaptive Weighting in Dam Deformation Prediction. *Water* **2025**, *17*, 988. https://doi.org/10.3390/w17070988

1. Introduction

As the core structure of hydraulic engineering, dams play a critical role not only in flood control, power generation, and water supply but also in effectively regulating the water environment and aquatic ecosystems. They have a significant impact on the safety of life and property, economic development, and the ecological environment of downstream populations [1]. Over time, the structural performance of dams inevitably degrades, and, in the event of a failure, the consequences are often catastrophic, leading to substantial loss of life and property for downstream residents. In fact, most dam failures are not sudden events but rather a gradual evolutionary process [2]. Therefore, it is essential to adopt reasonable and effective methods to monitor and evaluate the operational conditions of dams, with deformation prediction being a crucial component

of the monitoring process [3–5]. The comparison between predicted values and actual measurements can help identify abnormal conditions in a timely manner, enabling prompt response measures and ensuring the safe operation of dams [6,7].

Traditional dam deformation prediction methods are predominantly based on regression-based statistical models. These statistical approaches primarily establish functional relationships between effect quantities (such as deformation, seepage, and stress) and independent variables (such as water level, temperature, and time). Due to their simplicity and ease of implementation, they have been widely used in the field of dam safety monitoring [8–10]. However, during the actual operation of a dam, the relationship between the independent variables and the effect quantities is often extremely complex, exhibiting strong nonlinear characteristics. Consequently, traditional linear models are unable to accurately describe such nonlinear relationships, which further restricts the prediction accuracy. With the rapid development of machine learning and deep learning, techniques including but not limited to artificial neural networks, support vector machines, random forests, and genetic algorithms have been applied to dam deformation analysis and prediction. These algorithms exhibit significant advantages in addressing nonlinear problems and have demonstrated promising results [11–13]. Yang et al. employed long short-term memory (LSTM) networks to address the long-term dependencies present in dam deformation [14]. Building on this foundation, Yang et al. introduced Convolutional Neural Networks (CNNs) to identify and learn the spatial relationships within-time series data, developing a reliable dam prediction model [15]. Jiedeerbieke et al. proposed a dam deformation prediction model based on clustering partitioning and bidirectional long short-term memory (BiLSTM) networks. By considering the intrinsic correlations among monitoring points and integrating multiple feature information, the model achieves more comprehensive deformation monitoring [16]. Su et al. proposed a data processing framework that uses a long short-term memory (LSTM) model coupled with an attention mechanism to predict the deformation response of a dam structure in order to obtain more accurate nonlinear prediction results [17]. Zhou et al. established a dam displacement prediction model based on a multi-expert network. The model employs long short-term memory (LSTM) to extract temporal features and utilizes a multi-head attention mechanism to capture adjacency values, constructing a spatial graph to describe spatial relationships within the multi-expert network. Additionally, graph convolutional networks (GCNs) are applied to extract temporal relationships, enabling the prediction of multi-point dam displacements [18]. Wang et al. proposed a hybrid algorithm that combines backpropagation with a genetic algorithm (GA-BP) and a multiple population genetic algorithm (MPGA) to optimize the BP neural network. This hybrid algorithm has improved convergence speed and prediction accuracy [19]. The current research shows that hybrid neural networks often achieve better performance than single-structure models. However, this improvement in performance comes at a cost. As the model structure becomes more complex, the number of parameters in the neural network also increases significantly. This means that more computational resources are required, and longer training times are needed. The increase in computational overhead directly affects the application efficiency of the model, especially in real-world application scenarios with extremely high real-time requirements. Therefore, it is necessary to design a method that can significantly reduce the number of model parameters while improving the model's accuracy, thus shortening the training and inference times.

The principle of knowledge distillation (KD) [20] is effectively optimizin the original model by developing a lightweight, efficient, and small model. A lightweight student network is used to mimic the behavior and output of a powerful teacher network. However, currently, knowledge distillation has mostly been applied to classification problems [21–24],

where knowledge has been transferred through the soft labels generated by the teacher model. Nevertheless, this rule is not applicable in regression problems.

In response to the above-mentioned problems, this paper proposes a knowledge distillation method for regression problems, which is used for dam deformation prediction. This method consists of multiple modules, such as the pre-training of teacher networks, clustering integration based on information theory, the knowledge distillation process, and the prediction module. The main content is as follows: First, considering the robustness of teacher networks, we pre-trained teacher networks based on the original data to expand the teacher team. Subsequently, we adopted the clustering ensemble method to effectively reduce the number of redundant and low-precision teacher networks, thus avoiding the confusion in guidance caused by an excessive number of teacher networks with varying qualities during the subsequent distillation process. Then, we used the proposed weighted loss function for knowledge transfer to train a lightweight student network. Finally, the high-precision prediction of the student network was achieved. We used an actual concrete-faced rockfill dam as the experimental object, and the experimental results showed that the proposed model had higher accuracy and practicality in specific engineering cases.

Based on the above description, the contributions of this study are as follows:

(1) A multi-teacher distillation framework for regression problems is proposed. Traditional knowledge distillation methods have certain limitations when dealing with regression problems. We provide a new paradigm for this field. With the help of our framework, it is possible to provide the student network with richer and more comprehensive knowledge information.
(2) We adopted a clustering integration method based on information theory to reduce the number of redundant and low-precision teacher networks. This avoids knowledge redundancy and conflicts caused by too many teacher networks, further enhancing the performance and stability of the model.
(3) We propose of a novel weighted loss function for regression to guide the student network. Different weights are assigned according to the performance of different teacher networks, enabling the student network to pay more attention to the knowledge content of high-quality teachers during the learning process.
(4) Our method was applied to a concrete-faced rockfill dam in Guizhou province, China. The results show that, through our distillation algorithm, the prediction accuracy (MSE) of the student network can be effectively reduced by 40% to 73%. Compared with other knowledge distillation methods, our method has higher accuracy and practicality.

Next, an outline of the remaining parts of this paper is presented. Section 2 introduces the theoretical basis and traditional knowledge distillation (KD) methods. In Section 3, we present the proposed method. Section 4 reports the experimental results. In Section 5, the conclusions and future work are presented.

2. Related Works

2.1. HST Statistical Model

First, we briefly introduce the classic HST statistical model [25], According to the current knowledge system and relevant experience, the displacement of a dam is mainly related to three major influencing factors, namely, the upstream water level H, the temperature T, and the aging θ. We further express this as the water pressure component y_H, the temperature y_T, the aging component y_θ, and the constant term c as follows:

$$y = y_H + y_T + y_\theta + c \tag{1}$$

The water pressure component y_H of the horizontal displacement y of a gravity dam can be expressed by the upstream water levels H as follows:

$$y_H = \sum_{i=1}^{3} a_i H^i \tag{2}$$

where $a_i (i = 1, 2, 3)$ are the regression coefficients.

The actual situation of a dam is often complex. Since it is difficult to install a sufficient number of thermometers on the bedrock and the dam body, when the internal temperature field of the dam reaches a stable state, we usually use multi-cycle harmonic variations as the factors influencing the temperature component y_T as follows:

$$y_H = \sum_{i=1}^{3} a_i H^i y_T = \sum_{i=1}^{2} \left(b_{1i} \sin \frac{2\pi i t}{365} + b_{2i} \cos \frac{2\pi i t}{365} \right) \tag{3}$$

where $b_1 i$ and $b_2 i$ represent regression coefficients, and t represents the number of days from the detection date to the initial monitoring date.

The aging component y_θ of dam deformation is irreversible. For most dams, the displacement caused by aging develops rapidly during the initial operation period but tends to stabilize in the later stages. Therefore, y_θ is usually generalized as a combination of a logarithmic function and a linear function as follows:

$$y_\theta = c_1 \theta + c_2 \ln \theta \tag{4}$$

where $\theta = t/100$; t is calculated as the cumulative number of days from the monitoring day to the start of the monitoring period. Combining Equations (1)–(4), the complete mathematical expression of the HST model is as follows:

$$y = \sum_{i=1}^{3} a_i H^i + \sum_{i=1}^{2} (b_{1i} sin \frac{2\pi i t}{365} + b_{2i} cos \frac{2\pi i t}{365}) + c_1 \theta + c_2 \ln \theta + c \tag{5}$$

where the other symbols have the same meanings as in the previous equation.

2.2. Transformer Model

This section provides a comprehensive introduction to the Transformer. The Transformer [26] model was initially designed to address the challenges in natural language processing (NLP). Traditional models perform poorly in handling long-range dependencies. In contrast, the Transformer, with its unique architecture and innovative mechanisms, demonstrates outstanding performance and has extended its influence to various fields, including time-series prediction [27–31]. In time-series data, accurately capturing long-term dependencies and complex patterns is crucial for improving prediction accuracy. The Transformer can precisely meet this requirement. By modeling the relationships between different time steps in the time series, it can make more accurate predictions. The Transformer mainly consists of two core components: the encoder and the decoder. The encoder module is used to capture the complex relationships and contextual information among the elements in the input sequence, laying the foundation for subsequent processing. The decoder module focuses on feature reconstruction. Figure 1 shows the detailed framework of the Transformer.

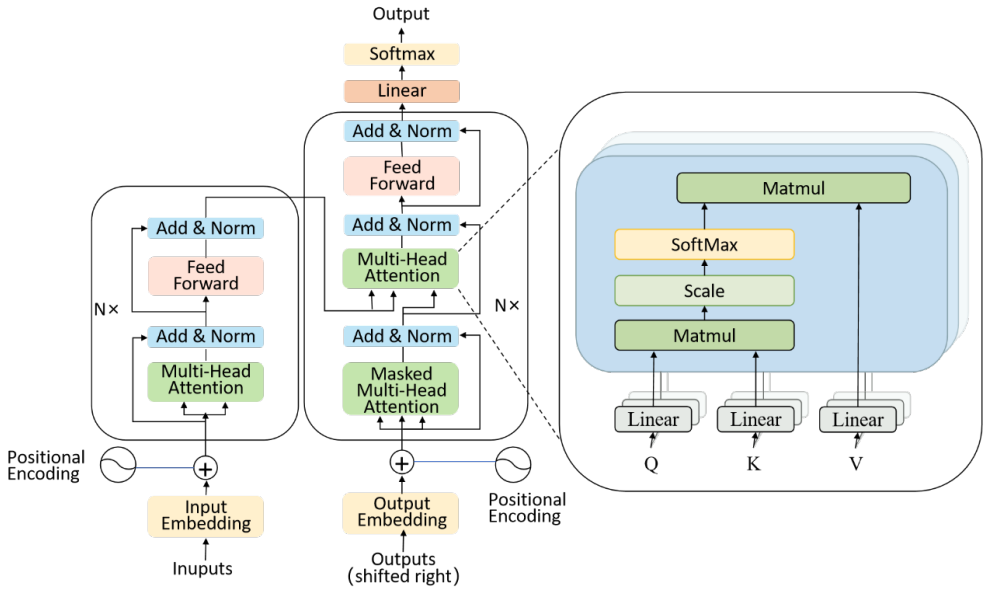

Figure 1. Architecture diagram of the Transformer [32].

The encoder is composed of multiple identical layers, and its primary function is to capture the dependencies within the input sequence. Each encoding layer takes the output of the previous layer as its input, and adjacent layers are connected through residual connections. This structure can effectively transmit information and mitigate the vanishing gradient problem during the training process. A normalization operation is performed after each residual module. Each layer of the encoder contains two core sub-layers: multi-head attention and a fully-connected feed-forward network. Different from a single self-attention mechanism, multi-head attention compares each query with a set of key vectors from multiple representation sub-spaces. Similar to the encoder, the decoder is also composed of multiple identical layers, with each layer taking the output of the previous layer as its input. Each decoding layer consists of three core sub-layers: masked multi-head attention, a fully-connected feed-forward network, and multi-head attention. The outputs of these sub-layers are passed through residual links and combined with layer normalization to ensure the stability of the input for each layer. The computation processes of the fully connected feed-forward network and multi-head attention are the same as those in the encoder. Masked multi-head attention is the same as multi-head attention, except for the self-attention computation process.

Compared with Recurrent Neural Networks (RNNs) that process data sequentially, the Transformer can process all data simultaneously to improve efficiency and performance. However, when applied to complex tasks such as dam deformation prediction, the Transformer faces a series of challenges, which can be mainly summarized as high computational resource requirements, large memory usage, and limited processing speed.

2.3. Multi-Teacher Knowledge Distillation

Knowledge distillation (KD), as an important method for model compression, mainly aims to effectively transfer the "dark knowledge" from complex and large-scale models to concise and small-scale models, enabling the smaller models to achieve performance similar to that of large models. In 2014, reference [20] proposed a method that allows the student network to learn the output of the teacher network, defining the results of the teacher model as "soft labels". Knowledge distillation works well in classification tasks because of its advantage in "dark knowledge", which refers to the softened hidden layer representations (logits) output of the teacher model. The specific process of knowledge distillation is shown in Figure 2a,b.

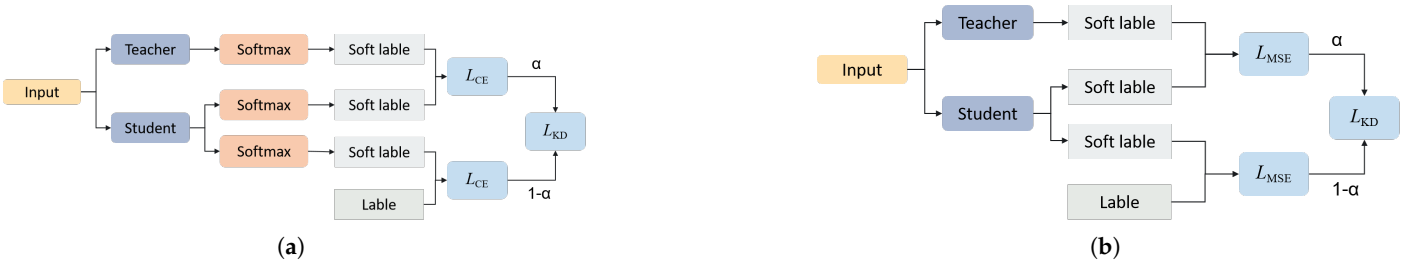

Figure 2. Knowledge distillation processes for classification tasks and regression tasks. (**a**) Knowledge distillation for classification. (**b**) Knowledge distillation for regression.

The objective function of knowledge distillation for classification is as follows:

$$L_{\mathrm{KD}} = \alpha \cdot L_{\mathrm{CE}}(O_s^{\mathrm{T}}, O_t^{\mathrm{T}}) + (1-\alpha) * L_{\mathrm{CE}}\left(O_s^1, y_{label}\right) \tag{6}$$

where α is the balance coefficient of the loss function, which is used to adjust the balance between the soft loss and the hard loss. Its value ranges from [0, 1]; L_{CE} is the cross-entropy loss function, T is the temperature coefficient, O_s^{T} is the probability distribution of the output of the student network after passing through the Softmax layer with the temperature coefficient T, O_t^{T} is the probability distribution of the output of the teacher network after passing through the Softmax layer with the temperature coefficient T; O_s^1 is the probability distribution of the output of the student network after passing through the Softmax layer; y_{label} is the label of the input data, which is defined as the hard label.

The objective function of knowledge distillation commonly used for regression is as follows:

$$L_{KD} = \alpha * L_{MSE}(O_s, O_t) + (1-\alpha) * L_{MSE}(O_s, y_{label}) \tag{7}$$

where O_s represents the direct output information of the student network; O_t represents the soft-label information directly output by the teacher network; L_{MSE} is the mean squared error loss function.

Multi-teacher distillation [33–36], as a variant of knowledge distillation, employs multiple teacher models to jointly guide the training of the student model. Compared with the traditional single-teacher knowledge distillation method, the multi-teacher distillation strategy provides more comprehensive and stable knowledge for the student model by aggregating the expertise from different teachers. Under the multi-teacher distillation framework, the learning process typically includes the following key steps:

(1) Teacher model training: Multiple teacher models are trained independently. Each teacher model may be based on a different model architecture or a different training subset, ensuring that they possess unique knowledge and capabilities.
(2) Teacher model integration: in multi-teacher distillation, weighted averaging of the prediction results is usually adopted to ensure more accurate knowledge transfer.
(3) Student model training: The student model is trained to mimic the outputs of multiple teacher models as closely as possible. In classification tasks, the cross-entropy loss function is commonly used to measure the consistency between the outputs of the student model and the teacher models so as to achieve efficient knowledge transfer.

Compared with the traditional knowledge distillation that relies on a single teacher model, multi-teacher distillation can provide more diverse and robust knowledge, thereby significantly improving the performance and generalization ability of the student model [37–39]. When dealing with complex tasks and large-scale model training, multi-teacher distillation shows excellent results. It not only optimizes the efficiency of knowledge transfer but also enhances the adaptability and robustness of the model in the

face of different data and tasks, demonstrating its wide application potential and important value in the field of deep learning.

2.4. The Basic Concept of Information Entropy

Here, we present some basic concepts of information entropy. Let v be a random variable with k possible values. $D(v) = \{v_1, v_2, \ldots, v_k\}$ is the set containing all possible values of the variable. If v is a discrete variable, its information entropy is defined as

$$H(v) = -\sum_{i=1}^{k} P(v_i)\log P(v_i) \tag{8}$$

where $P(v_i)$ is the probability of event $v = v_i$ occurring. If v is a continuous variable, then the information entropy is defined as

$$H(v) = -\int f(v)\log f(v)dv \tag{9}$$

where $f(v)$ is the probability density function of v; the smaller the value of $H(v)$, the more certain the value of this variable.

Conditional entropy quantifies the uncertainty of a random variable, assuming the value of another random variable is known. Let u be a random variable with k' possible values and $D(u) = \{u_1, u2, \ldots, u_{k'}\}$ be the set containing all possible values of this variable. At this time, the conditional entropy of v given u can be defined as

$$H(v|u) = \sum_{j=1}^{k'} P(u_j)H(v|u_j) \tag{10}$$

If v is discrete variable, it is defined as

$$H(v|u_j) = -\sum_{i=1}^{k} P(v_i|u_j)\log P(v_i|u_j) \tag{11}$$

If v is continuous, it is defined as

$$H(v|u_j) = -\int f(v|u_j)\log f(v|u_j)dv \tag{12}$$

Therefore, information entropy can be used to evaluate the consistency of an object set over a given feature set. The more objects have the same or similar feature values, the more certain these feature values of the objects.

3. Regression Method Based on Multi-Teacher Distillation

3.1. Method Overview

Traditional single-teacher networks may encounter bottlenecks in generalization ability and expressiveness when guiding student networks [40]. However, when adopting the multi-teacher distillation strategy, we may face challenges such as knowledge conflicts and high resource consumption [41]. To overcome these limitations, we propose an innovative method: using the idea of clustering ensemble to accurately cluster and efficiently integrate multi-teacher networks, ensuring that teacher networks with similar structures can be clustered into the same category. This strategy helps reduce model redundancy and improve the efficiency of knowledge transfer to form a more powerful guiding model. To further optimize the application of the multi-teacher distillation method in regression problems, we introduce a new loss function, aiming to solve the knowledge transfer problems faced

by multi-teacher network distillation in regression tasks and improve the accuracy and stability of the student model when dealing with complex regression problems.

In the entire knowledge distillation process, first, we pre-train N teacher networks. Then, we perform clustering ensemble on the pre-trained N teacher networks. The clustering method is explained in detail in Section 3.2. Through our method, the N teacher networks are clustered into k clusters. In each cluster, we select the top b teacher networks for weighted averaging, thereby integrating multiple teacher networks into a small number of high-quality teacher networks, which helps improve the subsequent distillation efficiency. Next, the predicted outputs of the aggregated k teacher networks are used to guide the learning process of the student network.

The detailed process is shown in Figure 3. The learning process of the student network consists of two parts. One part comes from the predicted outputs of the teacher networks, denoted as Limit, sand the other part comes from the real labels, Label, denoted as L_{MSE}. We use the balance factor α to adjust the weights of the student network for different knowledge sources. In addition, we provide Table 1 to summarize the main symbols used.

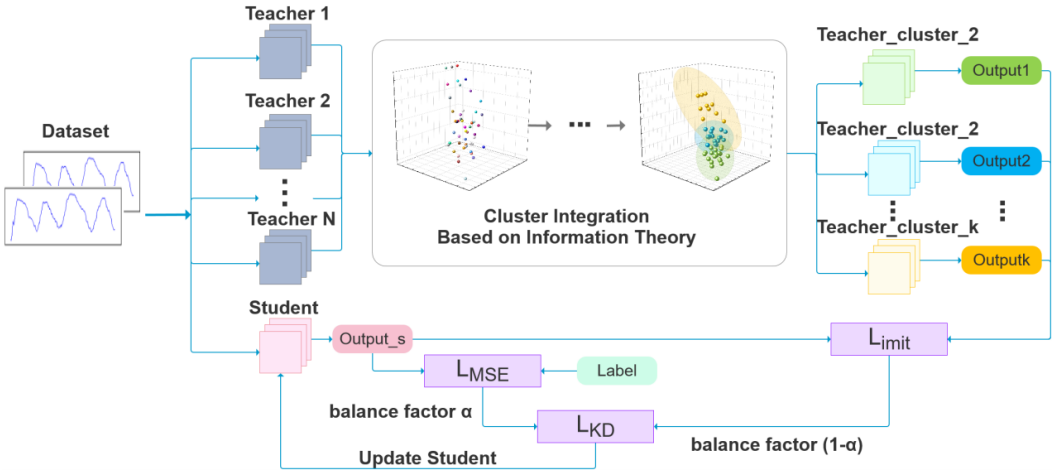

Figure 3. Multi-teacher knowledge distillation architecture for regression tasks.

Table 1. Explanations of some symbols used in this paper.

Symbol	Description
λ_h	The feature value of the h-th base clustering
D(v)	The domain of variable (feature) v
λ_{hl}	The clustering label of C_{hl}
$\mathbb{C}_h$	The h-th base clustering set
C_{hl}	The l-th cluster in $\mathbb{C}_h$
S $\subseteq$ T	A cluster composed of multiple objects in T
$\mathbb{S} = \{S_l\}_{l=1}^{p}$	A set composed of clusters S where S_l is the l-th cluster
$\mathbb{C}$	$\mathbb{S} = \mathbb{C}$ The set composed of the objects in T, which is defined as a partition
$\mathbb{C}_*$	The partition result formed by the final clusters
k	The number of clusters in partition $\mathbb{C}_*$
λ_*	The final clustering feature vector

3.2. Multi-Teacher Network Clustering Ensemble Based on Information Theory

Our goal is to integrate the wisdom of multiple teacher networks to form a small number of elite teacher networks, thereby enhancing the learning efficiency of the student network. To make the clustering results more reliable, we draw on the idea of clustering ensemble, aggregating multiple clustering results to generate more robust and stable clustering outcomes.

In our approach, we draw inspiration from the information–theoretic clustering ensemble methodology [42], as illustrated in Figure 4. We consider each teacher network as embedded within two feature spaces: the original feature space and the basic clustering feature space.

First, we generate a pool of teacher networks based on the original data. Then, through similarity calculation, we form the original data feature set $T(A)$. Subsequently, using multiple clustering methods on the original data feature set, we construct the base clustering feature set $T(\Delta)$. After that, we obtain the final clustering result through hierarchical clustering. Our requirement for the final clustering result is to have high consistency with the base clustering feature set. We use information entropy to evaluate the uncertainty between them to obtain high-quality clustering results. Finally, we obtain the final clustering result $T(\lambda_*)$ for the subsequent weighted integration of teacher networks. The detailed definition and process of this method are presented in Figure 4.

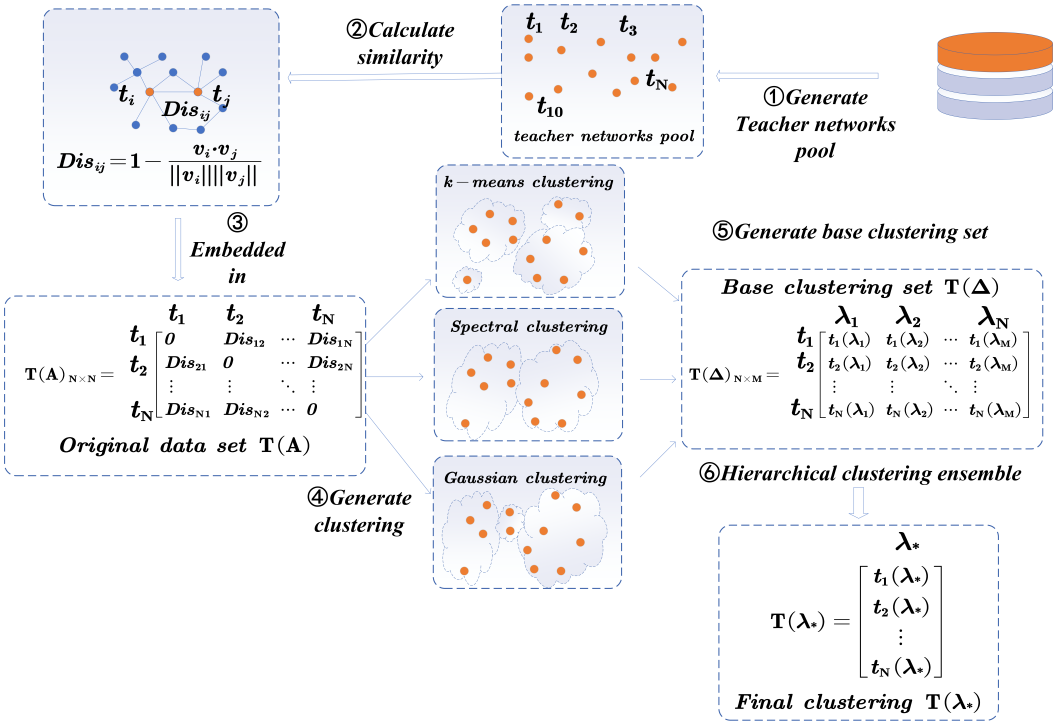

Figure 4. The basic process of the clustering ensemble method based on information theory.

$T(A)$ is an N × N matrix, where N is the number of teacher networks. We define it as the original data feature set. First, we present the construction processes of the original feature set and the basic clustering-feature set. For teacher networks t_i and t_j, we flatten the weights of the linear layers in the last layer of the teacher networks into one-dimensional vectors v_i and v_j. We use cosine similarity to evaluate the similarity between the two teacher networks:

$$\cos_{ij} = \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{13}$$

In addition, we define Dis_{ij} as representing the distance between the two teacher networks T_i and T_j:

$$Dis_{ij} = 1 - \cos_{ij} \tag{14}$$

We construct the original data feature set of teacher networks based on the distances between different teacher networks. The purpose of this is to fully consider the structural similarity among the different teacher networks when generating the basic clustering feature set. Subsequently, we select M clustering algorithms and generate M clustering

results according to the original data feature set $T(A)$. These results form the base clustering feature set $T(\Delta)$. $T(\Delta)$ is an N × M matrix, where M represents the number of clustering methods. It denotes the base clustering feature set composed of all initial clustering results.

Next, we introduce how to define information entropy to measure the consistency between the clustering results and the base clustering feature set Δ. We define each λ_h as a random variable, which represents the feature value formed by the h-th base clustering result. Let $D(\lambda_h) = \{\lambda_{hl}\}_{l=1}^{k_h}$, where λ_{hl} is the label of the clustering result of C_{hl}, and $P(\lambda_{hl})$ is the probability that λ_{hl}, where λ_h is the feature value of the h-th base clustering. Suppose $S \subseteq$ T is a cluster that contains multiple objects and S is a subset of the set T. We calculate $P(\lambda_{hl}|S)$ by means of the relative frequency of the label value λ_{hl} occurring in the cluster S:

$$P(\lambda_{hl} \mid S) = \frac{|\{x_i \mid x_i(\lambda_h) = \lambda_{hl}, x_i \in S\}|}{|S|} \tag{15}$$

The information entropy of λ_h given S is

$$H(\lambda_h \mid S) = -\sum_{i=1}^{k_j} P(\lambda_{hl} \mid S) \log P(\lambda_{hl} \mid S) \tag{16}$$

In addition, we assume that each λ_h is independent of the others. Then, Δ can be regarded as a set composed of M random variables. Given cluster S, the information entropy of Δ can be expressed as:

$$H(\Delta \mid S) = \sum_{h=1}^{M} H(\lambda_h \mid S) \tag{17}$$

Suppose $\mathbb{S} = \{S_l{}_{l=1}^{p}\}$ is the set of S, including p clusters, where S_l is the l-th cluster, $1 \leq l \leq p$. Therefore, according to the definition of conditional entropy, given S, the information entropy of Δ is

$$H(\Delta \mid \mathbb{S}) = \sum_{l=1}^{p} P(S_l \mid \mathbb{S}) H(\Delta \mid S_l) \tag{18}$$

$P(S_l|\mathbb{S})$ is the probability of S_l in $\mathbb{S}$, and the calculation process is as follows:

$$P(S_l \mid \mathbb{S}) = \frac{\mid S_l \mid}{\sum_{l=1}^{p} \mid S_l \mid} \tag{19}$$

$H(\Delta|\mathbb{S})$ can be regarded as the sum of the consistencies of all clusters $\mathbb{S}$, which we define as the consensus sum. If the consistency of each cluster S_l with respect to Δ is very high, then the value of $H(\Delta|\mathbb{S})$ is lower. Based on Formulas (17) and (18), we can define the consensus sum of $\mathbb{S}$ on Δ as follows:

$$I(\Delta|\mathbb{S}) = H(\Delta|S) - H(\Delta|\mathbb{S}) \tag{20}$$

We believe that if set $\mathbb{S}$ of clusters is a good set for cluster S, it can enhance the consensus of cluster S on Δ. The larger the $I(\Delta|\mathbb{S})$, the higher the clustering consistency on Δ.

Given T, $H(\Delta|T)$ is determined. Then, the more objects in each cluster of the set $\mathbb{S}$ share the same labels in the basic clustering feature set Δ, the lower the value of $H(\Delta|\mathbb{S})$. Therefore, clustering results with high consistency on Δ are considered to have good

performance. In addition, if a set of clusters $\mathbb{S} = \{S_i, S_j\}$ contains two clusters S_i and S_j, then, according to Formula (20), we have

$$I(\Delta|\{S_i, S_j\}) = H(\Delta|S_i \cup S_j) - H(\Delta|\{S_i, S_j\}) \tag{21}$$

Given S_i and S_j, $H(\Delta|S_i \cup S_j)$ is determined. If most of the objects in $S_i \cup S_j$ have the same labels on Δ, then $I(\Delta|\{S_i, S_j\})$ decreases significantly. At this time, we consider that these two clusters are similar on Δ.

The clustering-ensemble task can be regarded as a clustering operation performed on the basic clustering feature set. We use information entropy as the objective function to screen out the optimal final clustering results. Assuming $\mathbb{S} = \mathbb{C}$ is the set of clusters formed by T, where $\mathbb{C}$ is defined as a partition, we ultimately use $H(\Delta|\mathbb{C}_*)$ as the objective function of the clustering ensemble method and design the corresponding optimization model to perform the clustering ensemble task on the basic clustering feature set.

The formal description of the clustering ensemble optimization model is

$$\min_{\mathbb{C}_*} H(\Delta|\mathbb{C}_*) \tag{22}$$

According to the above-mentioned content, we can conclude that, given T,

$$\max_{\mathbb{C}_*} I(\Delta \mid \mathbb{C}_*) = H(\Delta \mid T) - \min_{\mathbb{C}_*} H(\Delta \mid \mathbb{C}_*) \tag{23}$$

According to Formula (23), we know that minimizing the objective function $H(\Delta|\mathbb{C}_*)$ is equivalent to maximizing the consensus sum between the final clustering results and the basic clustering feature set.

We adopt a hierarchical clustering strategy to generate the final clustering results. Define $\mathbb{C}^{(z)}$ as the partition generated by merging cluster $C_i^{(z-1)}$ and cluster $C_j^{(z-1)}$ at the z-th time. First, we construct an initial partition $\mathbb{C}^0$, where each teacher network in $T = \{t_1, t_2, \ldots, t_N\}$ serves as a separate cluster; that is, $\mathbb{C}^{(0)} = \{C_1, C2, \ldots, C_N\}$. According to Formula (18), at this time, $H(\Delta|\mathbb{C}_0) = 0$. Next, we continuously combine two partitions to form a new partition $\mathbb{C}^{(z+1)}$. Based on the properties of conditional entropy, we have

$$H(\Delta|\mathbb{C}^{(z+1)}) \geq H(\Delta|\mathbb{C}^{(z)}) \tag{24}$$

As the number of clusters decreases, the consensus sum of the clustering results also gradually decreases. Therefore, each time we merge partitions, we hope that the consensus sum decreases as little as possible. The optimization problem can be transformed into

$$\min_{\mathbb{C}^{(z+1)}} \left[H(\Delta \mid \mathbb{C}^{(z)}) - H(\Delta \mid \mathbb{C}^{(z+1)}) \right] = \min_{C_i^{(z)}, C_j^{(z)} \in \mathbb{C}^{(z)}} I(\Delta \mid \{C_i^{(z)}, C_j^{(z)}\}) \tag{25}$$

Therefore, each time we need to merge the two clusters $C_i^{(z)}$ and $C_j^{(z)}$ with the highest consistency in $\mathbb{C}^{(z)}$, the merging condition is

$$\min_{C_i^{(z)}, C_j^{(z)} \in \mathbb{C}^{(z)}} I(\Delta|\{C_i^{(z)}, C_j^{(z)}\}) \tag{26}$$

When the number of clusters in $\mathbb{C}^{(z)}$ is k, the merging process terminates, and the final clustering result is $\mathbb{C}_* = \mathbb{C}^{(z)}$.

The aforementioned steps effectively group multiple teacher networks, which are followed by our screening process to enhance the overall model's performance and efficiency. We obtain the final clustering result $T(\lambda_*)$ from $\mathbb{C}_*$, which ensures that the clustering result

maximally reflects the similarities and differences among the teacher networks. Here, λ_* denotes the final clustering result, where $1 \leq \lambda_* \leq k$. Our objective is to further refine the team of teacher networks. To achieve this, we rank the teacher networks within each cluster based on their mean squared error (MSE), allowing us to clearly understand the performance of each teacher network within its cluster. Subsequently, we select the top b teacher networks from each cluster (where b is a predefined parameter, typically adjusted according to specific requirements and model characteristics). This selection aims to retain the best-performing teacher networks in each cluster while eliminating those with lower prediction accuracy, thereby ensuring the quality of the refined teacher network team.

After completing the screening of teacher networks, we determine the weighted average of the predictions from the selected teacher networks. This weighted average takes into account the performance differences among the teacher networks, assigning higher weights to those with superior performance. Based on the final clustering result, we rank the teacher networks in each cluster according to the MSE evaluation metric, aiming to further streamline the teacher network team. We sequentially select the top b teacher networks from each cluster and compute a weighted average of their predictions, ultimately generating k strong teacher networks. These strong teacher networks not only retain the advantages of multiple teacher networks but also eliminate redundant and low-performance networks. Consequently, this step provides the student network with more accurate and effective knowledge guidance, enhancing the learning efficiency and predictive performance of the student network. Our proposed method is formally described in Algorithm 1.

Algorithm 1 Multi-teacher clustering ensemble based on information theory.

Input: $T(A)$, $T(\Delta)$, k
Output: $D(O_t)$
Set $z = 0$ and produce an initial partition $\mathbb{C}^{(z)} = \{C_i^{(z)}\}_{i=1}^{N}$ where $C_i^{(z)} = \{t_i\}, 1 \leq i \leq N$;
while $|\mathbb{C}^{(z)}| > k$ **do**
  Select set two clusters $C_i^{(z)}$ and $C_j^{(z)}$ which satisfy $\min_{C_i^{(z)}, C_j^{(z)} \in \mathbb{C}^{(z)}} I_w(\Delta | \{C_i^{(z)}, C_j^{(z)}\})$;
  Execute clustering merging $C_i^{(z)} = C_i^{(z)} \cup C_j^{(z)}$;
  Update partition results $\mathbb{C}^{(z+1)} = \mathbb{C}^{(z)} - C_j^{(z)}$;
  The number of merges increases automatically $z = z + 1$;
end while
Get $T(\lambda_*)$ corresponding to $\mathbb{C}^{(z)}$;
Cluster the teacher networks into k clusters according to $T(\lambda_*)$;
while $i < k$ **do**
  Select the top b teacher networks in the i-th cluster;
  Calculate the weighted average result O_{ti};
end while
Return $D(O_t) = \{O_{t1}, O_{t2}, \ldots, O_{tk}\}$;

3.3. Knowledge Distillation

Under the framework of muti-teacher distillation, the design of the loss function is of vital importance. It not only determines the learning efficiency of the student network of the knowledge of the teacher networks but also directly affects the performance of the final network. Next, the carefully designed multi-teacher distillation loss function for the regression task is elaborated in detail.

Inspired by [43], we chose to use Limit to replace the mean squared error (MSE) used in traditional regression tasks. When the performance of the student network is better than that of the teacher network, we did not add the loss of the teacher network at this time

as an additional loss of the student network. In this way, we could effectively guide the model learning.

At the same time, the loss function we propose fully considers the performance differences among multiple teacher models. We set parameter λ to balance their impacts on the student model. This weighting strategy ensures that teacher models with excellent performance can play a greater role in the training process, thereby accelerating the learning efficiency of the student model and improving its generalization ability. The objective function we propose is as follows:

$$L_{KD} = \frac{1}{n}\sum_{i}^{n}\sum_{j}^{m}\alpha(O_s - y_{label})_i^2 + (1-\alpha)\lambda_j L_{mt}(O_s, O_{tj})_i \tag{27}$$

$$L_{imit}(O_s, O_t) = \begin{cases} (O_s - O_t)^2, & \text{if}|O_s - y_{label}| > |O_t - y_{label}| \\ 0, & \text{otherwise} \end{cases} \tag{28}$$

$$\lambda_j = \left(1 - \frac{(mse_j)}{sum(M)}\right) \tag{29}$$

$$M = \{\frac{1}{n}\sum_{i}^{n}(O_{it} - y_{label})_j^2 : j = 1, ..., k\} \tag{30}$$

where λ_j is the set weight coefficient, α is the balance coefficient, n is the number of samples, k is the number of clustered teacher networks, and $M = \{mse_j\}_{j=1}^{k}$ is the set of mean squared errors of the clustered teacher networks.

The following presents its mathematical derivative-derivation process:

$$\begin{aligned} \nabla = \frac{\partial L_{KD}}{\partial W} &= \frac{\frac{1}{n}\sum_{i}^{N}\alpha(Wx + b - y_{label})^2 + \sum_{j}^{k}(1-\alpha)\lambda_j(wx + b - O_{tj})^2}{\partial W} \\ &= \frac{2}{n}\alpha x(Wx + b - y_{label}) + 2(1-\alpha)x\sum_{j}^{k}\lambda_j(wx + b - O_{tj}) \end{aligned}$$

$$W = W_0 - \eta \cdot \nabla L_{KD}(W_0)$$

where W_0 is the parameter of the independent variable, η is the learning factor, and W is the updated W_0.

4. Results

4.1. Project Overview

This study focused on a concrete-faced rockfill dam located in Guizhou province, China. The dam had a crest elevation of 154 m, a crest width of 10 m, a crest length of 428.93 m, a maximum dam width of 454.085 m, an installed capacity of 90,000 kilowatts, and a total reservoir capacity of 1.325 billion cubic meters. To monitor the dam's operational status, various sensors were installed internally to collect data and assess its performance. Horizontal displacement and settlement monitoring sensors were placed at the cross-section 0+005.000, with four monitoring lines. For a visual representation, refer to Figure 5. Monitoring data from four monitoring points (SG19, SG20, SG25, and SG26) from September 2021 to May 2024 were selected for modeling, with a data interval of 5 days. The dataset was divided into 80% for training and 20% for testing. During model training, the training set was further split into 0.8 for training and 0.2 for validation. The deformation data of the monitoring points are shown in Figure 6.

The dam deformation prediction model developed in this study used influencing factors such as water level, temperature, and time effects as inputs, with dam deformation as the output. A total of 10 influencing factors were selected, including water pressure factors H^1, H^2, H^3, and H^4; temperature factors $sin(2\pi t/365)$, $cos(2\pi t/365)$, $sin(2 \cdot 2\pi t/365)$, and $cos(2 \cdot 2\pi t/365)$; as well as time-effect factors θ and $ln(\theta)$.

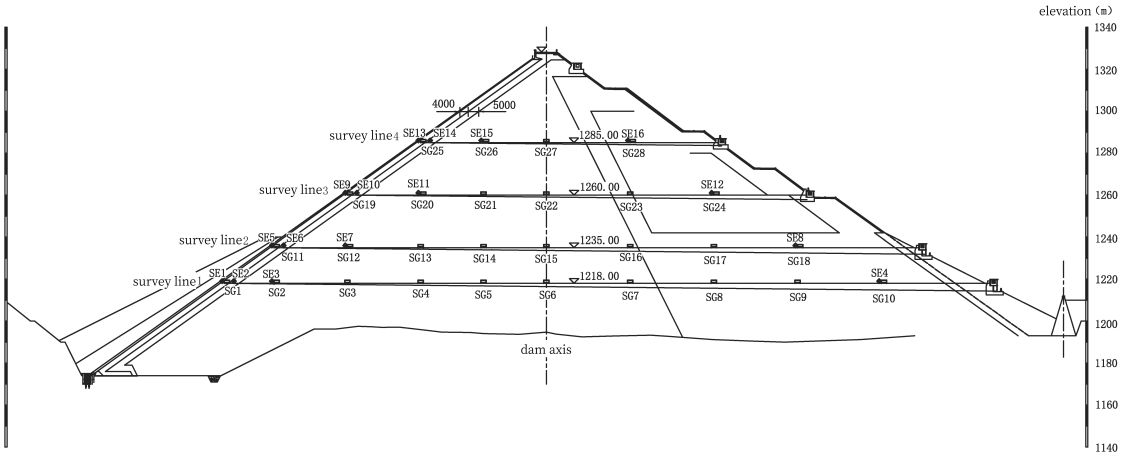

Figure 5. Monitoring profile diagram of the dam body on the left side of the cross-section 0+005.000.

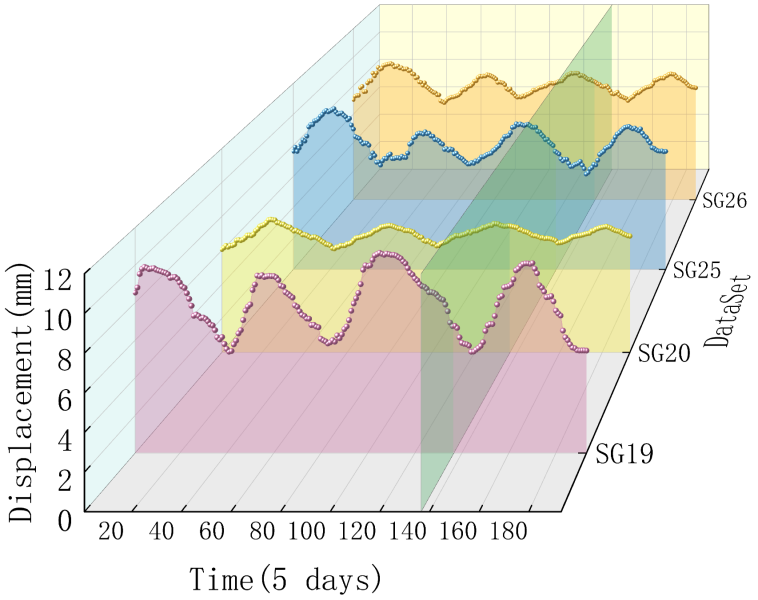

Figure 6. Deformation monitoring data of the monitoring points.

4.2. Evaluation Metrics

To comprehensively evaluate the prediction accuracy of the model, various indicators are usually used. We adopted RMSE (root mean squared error), MAPE (mean absolute percentage error), MAE (mean absolute Eerror), and the coefficient of determination (R^2) to evaluate the model. The specific formulas are given in Equations (31)–(34):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{31}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{32}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{33}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y}_i)^2} \tag{34}$$

where N is the total number of samples, y_i is the predicted value of the i-th sample, $\hat{y}_i$ is the observed value of the i-th sample, and $\overline{y}_i$ is the average value of the observed values of the i-th sample.

RMSE evaluates the prediction accuracy, MAPE evaluates the model's accuracy, and MAE reflects the actual situation of the prediction error. The smaller the values of these three parameters, the higher the model's accuracy. R^2 represents the fitting effect of the model, wehre the higher its value, the better the prediction performance.

4.3. Experimental Settings

We adopted the Transformer model as the teacher network, which included a position-encoding module, encoder layers, and a decoder composed of a simple linear layer. It was specifically designed for processing and predicting complex sequential data. We used a Temporal Convolutional Network (TCN) as the student network model. We selected the Temporal Convolutional Network (TCN) as the student network because it adopts a convolutional structure. Compared with traditional recurrent neural networks, convolutional operations can process data in parallel, thereby improving computational efficiency. Moreover, convolutional kernels can automatically extract local features from the data. Most importantly, the TCN has a relatively simple structure with a small number of parameters and thus has a relatively low demand for computing resources. Due to space limitations, we omit the description of the TCN and recommend readers refer to [44] for detailed information.In addition, we provide the algorithmic complexities of the Transformer and TCN in Table 2. On the one hand, it served as a reference for selecting the teacher network and the student network. On the other hand, it was used to verify the degree of reduction in model complexity after distillation. We set the initial number N of teacher networks to 50, the number k of clustering categories of teacher networks to 3, and the basic clustering set M composed of basic clustering results to 15. We used three methods, namely, the classic k-means, spectral clustering, and Gaussian mixture clustering, to cluster the 50 teacher networks according to the original feature set, with the number of clusters being 5. Each method selected 5 different random seeds to obtain 5 different clustering results, and b was set to 6 during the clustering selection process. Both the teacher network and the student network used the Adam optimizer. We initialized the learning rate at 0.0001, set the batch size to 8, and set α to 0.8 during the knowledge distillation process.

Table 2. Complexity of teacher networks and student network.

Network	#Params	FLOPs
Transformer	87007	8735
TCN	20.735 M	0.9928 M

4.4. Parameter Sensitivity Experiment

In the experiments presented in this paper, parameter α was incorporated into the objective function. Its purpose is to regulate the contribution of the knowledge extracted from the teacher network. The value of α has a significant and intuitive impact on the

knowledge distillation process.Theoretically, a larger value of α means that the teacher network has more influence on the distillation process. Conversely, when α is smaller, the student network places more emphasis on its own learning ability during the distillation. In the experiments in this study, to comprehensively and thoroughly investigate the impact of parameter α on the model's performance, we selected a set of representative values from the set {0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 0.99}. To more intuitively demonstrate the relationship between the model's performance and parameter α, we plotted a trend graph showing the model's R_2 (coefficient of determination; the closer the R^2 value to ond, the better the model fits the data) against α, as shown in Figure 7.

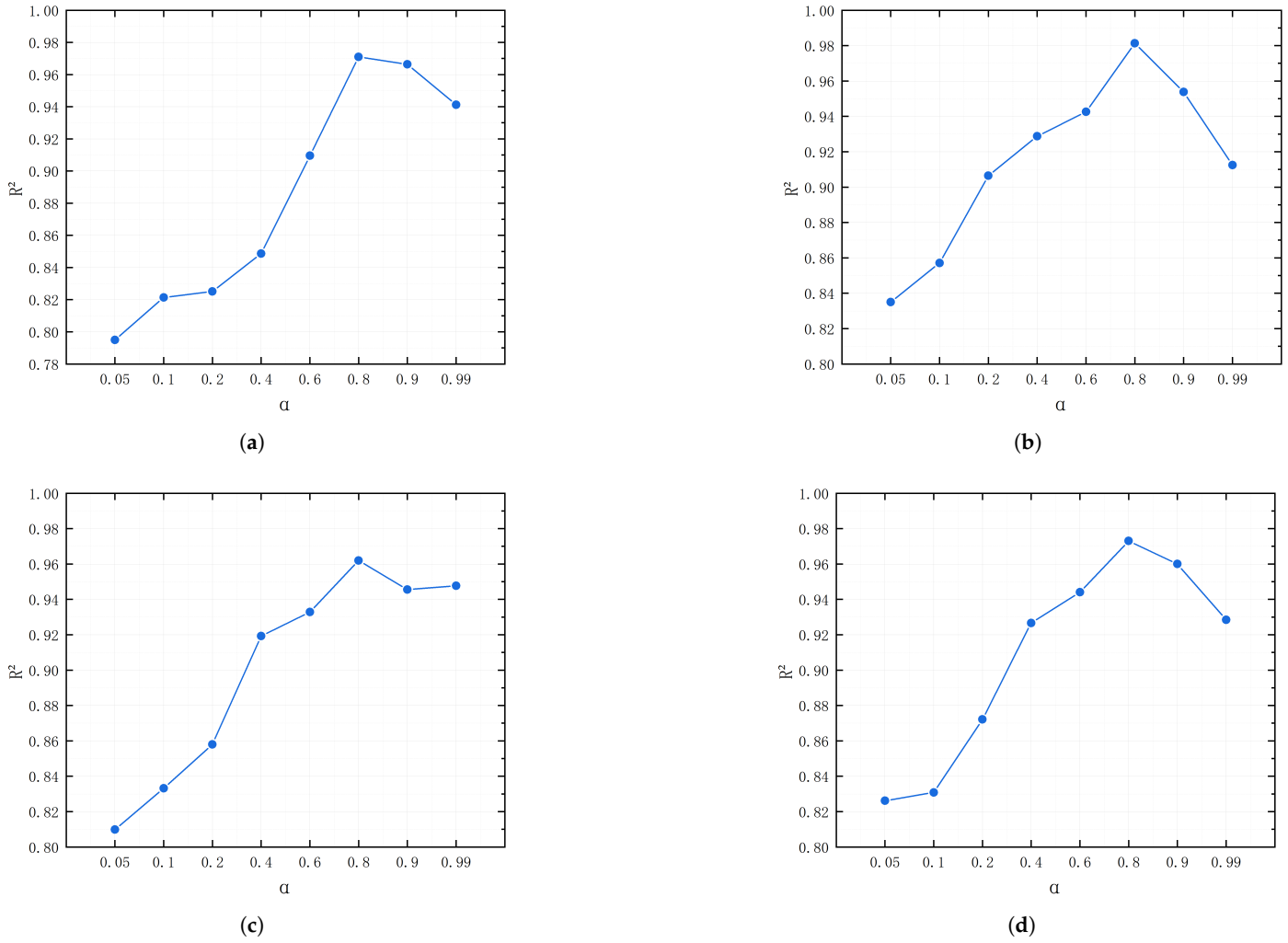

Figure 7. The influence of α on the model's R_2: (**a**) monitoring point SG19; (**b**) monitoring point SG20; (**c**) monitoring point SG25; (**d**) monitoring point SG26.

We found that when α was set to 0.8, the model achieved the optimal accuracy. This result indicates that when $\alpha = 0.8$, an ideal balance is reached between the teacher network and the student network. On the one hand, it can fully utilize the rich knowledge provided by the teacher network; on the other hand, it gives the student network enough room to exert its own learning ability, enabling the model to achieve the best performance during the knowledge distillation process. Based on this important experimental result, we made a reasonable and scientific decision. In the subsequent experiments, we fixed the value of α at 0.8.

4.5. Distillation Performance Evaluation

Table 3 shows the prediction results of the teacher networks after clustering, the student network, and CLMTKD on the test set. TeacherCL1,TeacherCL2 and TeacherCL3 are advanced teacher networks generated by weighted averaging the top six teacher networks after clustering. From SG19, SG20, SG25, and SG26, it can be seen that the proposed CLMTKD provides a significant improvement compared with the student network and is close to or even better than the clustering results of the advanced teacher networks at some measurement points.

Table 3. Performance comparison among teacher networks after clustering, student network, and student network after distillation.

Dataset	Model	MSE (mm)	RMSE (mm)	MAE (mm)	MAPE	R^2
SG19	TeacherCL1	**0.06049**	**0.24595**	**0.18483**	**2.36203**	**0.97679**
	TeacherCL2	0.09914	0.31487	0.25275	3.35216	0.96196
	TeacherCL3	0.10856	0.32949	0.25964	3.57259	0.95834
	Student	0.13955	0.37357	0.29399	3.88849	0.94644
	CLMTKD	0.07578	0.27527	0.21325	2.77232	0.97092
SG20	TeacherCL1	**0.00267**	**0.0513**	**0.04072**	**0.59478**	**0.96749**
	TeacherCL2	0.00317	0.05634	0.04432	0.64194	0.96088
	TeacherCL3	0.00507	0.07117	0.05386	0.78496	0.93758
	Student	0.00520	0.07210	0.05923	0.85757	0.93594
	CLMTKD	0.00308	0.05554	0.04265	0.62291	0.96199
SG25	TeacherCL1	0.01446	0.12027	0.09045	1.14168	0.97692
	TeacherCL2	0.03564	0.18879	0.15237	1.93899	0.94313
	TeacherCL3	0.03783	0.19449	0.15885	1.98612	0.93964
	Student	0.04463	0.21127	0.17393	2.18014	0.92878
	CLMTKD	**0.01171**	**0.10819**	**0.07837**	**0.98788**	**0.98132**
SG26	TeacherCL1	0.01018	0.10090	0.08128	1.01031	0.96411
	TeacherCL2	0.01446	0.12026	0.09729	1.21837	0.94901
	TeacherCL3	0.01482	0.12174	0.09554	1.17603	0.94775
	Student	0.02176	0.14753	0.12029	1.50459	0.92327
	CLMTKD	**0.00766**	**0.08754**	**0.06781**	**0.84433**	**0.97298**

Figure 8 presents a comparison between the student network and the student network after knowledge distillation. It is evident from the figure that the fitting performance of the student network significantly improved after knowledge distillation. Specifically, the distilled student network demonstrates superior fitting capability in the regression fitting curves, better capturing the distribution characteristics of the data. This indicates that the knowledge distillation process effectively enhances the performance of the student network, enabling it to learn and predict more accurately while maintaining lower complexity. This result further validates the effectiveness of the proposed CLMTKD algorithm in improving the learning efficiency and prediction accuracy of the student network.

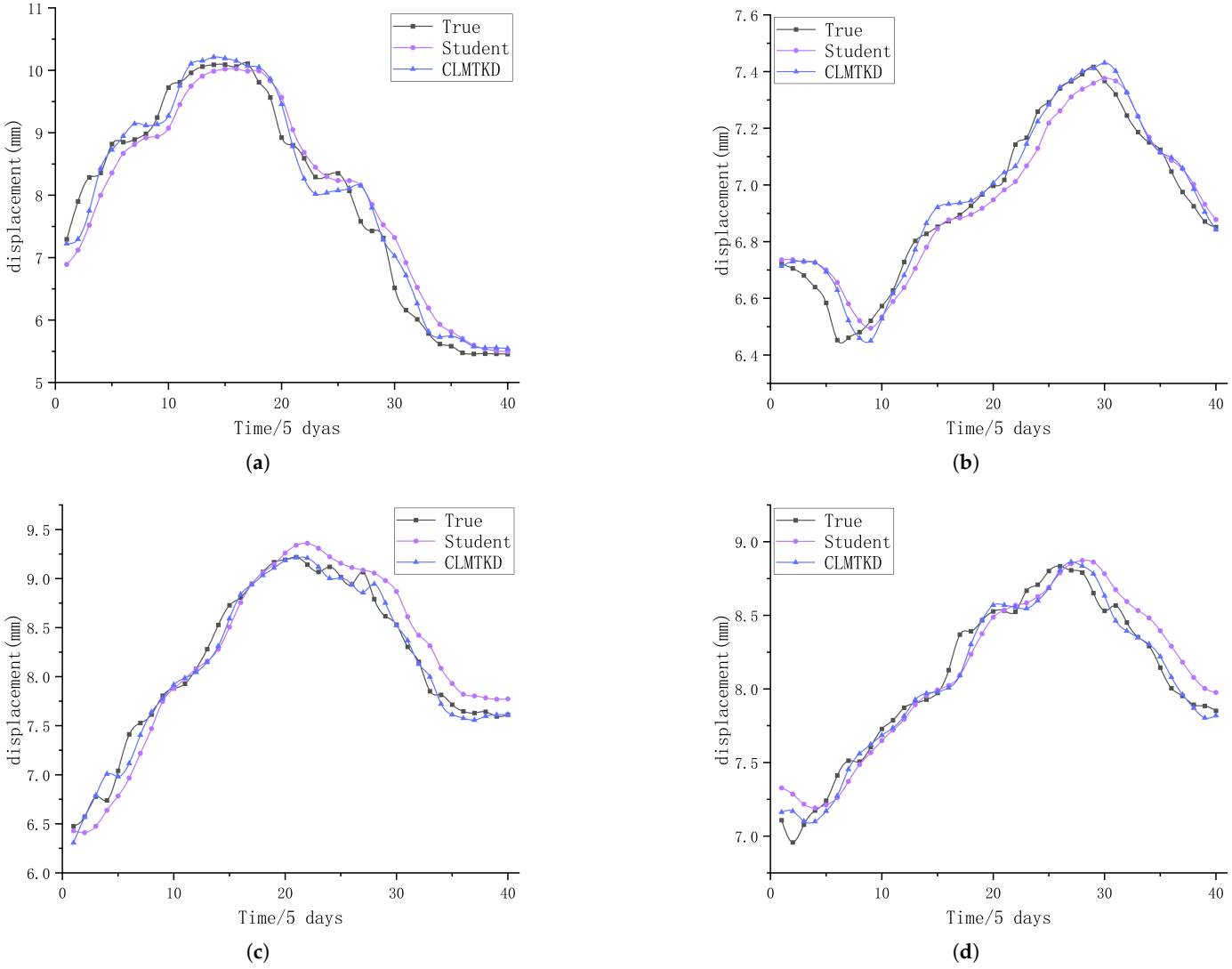

Figure 8. Prediction fitting images of the student network and the student network after distillation: (**a**) monitoring point SG19; (**b**) monitoring point SG20; (**c**) monitoring point SG25; (**d**) monitoring point SG26.

4.6. Comparative Experiment

To compare the effectiveness of our method with traditional knowledge distillation methods for regression, we selected a total of 5 comparison methods, including single-teacher network distillation and multi-teacher network distillation. Methods such as AIL [45], TBR [46], and TOR [47] use a single teacher network for knowledge distillation. Since the research on multi-teacher distillation models for regression is still in its infancy, we conducted two comparative experiments. The first method was forming a strong teacher network by weighted averaging all teacher networks and then selecting the regression loss of the AIL method to guide the learning of the student network (MTKD). The second method involved using all teacher networks simultaneously and selecting the regression loss in Formula (28) to guide the learning of the student network (ALLMTKD).

According to the comparison of the experimental results in Table 4, it can be found that the proposed CLMTKD achieves ideal results on the four datasets. Figure 9 is the radar chart of all algorithms, where the larger the area, the better the performance. It is not difficult to find that our method outperforms the other methods on all indicators.

Table 4. Comparison with other benchmark models.

Dataset	Model	MSE (mm)	RMSE (mm)	MAE (mm)	MAPE	R^2
SG19	AIL	0.11798	0.34348	0.26597	3.25020	0.954724
	TBR	0.09040	0.30066	0.23679	3.07238	0.965308
	TOR	0.11135	0.33638	0.27738	3.53747	0.95658
	MTKD	0.08714	0.29520	0.22549	2.94286	0.96656
	ALLMTKD	0.29588	0.54394	0.47393	6.32527	0.88645
	CLMTKD	**0.07578**	**0.27527**	**0.21325**	**2.77232**	**0.97092**
SG20	AIL	0.00534	0.07307	0.05780	0.83995	0.93421
	TBR	0.00423	0.06504	0.05466	0.79410	0.94788
	TOR	0.00432	0.05848	0.04533	0.66193	0.95786
	MTKD	0.00398	0.06305	0.04952	0.71876	0.95102
	ALLMTKD	0.01245	0.11157	0.10012	1.44319	0.84661
	CLMTKD	**0.00308**	**0.05554**	**0.04265**	**0.62291**	**0.96199**
SG25	AIL	0.04095	0.20237	0.16139	1.96994	0.93465
	TBR	0.02605	0.16139	0.12953	1.61069	0.95844
	TOR	0.02077	0.14413	0.11386	1.41403	0.96685
	MTKD	0.01336	0.11561	0.08560	1.07310	0.97867
	ALLMTKD	0.10177	0.31900	0.27808	3.44080	0.83762
	CLMTKD	**0.01171**	**0.10819**	**0.07837**	**0.98788**	**0.98132**
SG26	AIL	0.02030	0.14246	0.11936	1.47008	0.92845
	TBR	0.01038	0.10186	0.07854	0.97147	0.96342
	TOR	0.01212	0.11010	0.08304	1.02446	0.95727
	MTKD	0.00932	0.09648	0.07726	0.96672	0.96718
	ALLMTKD	0.05261	0.22937	0.19274	2.32723	0.81451
	CLMTKD	**0.00766**	**0.08754**	**0.06781**	**0.84433**	**0.97298**

Further analyzing the experimental results, compared with traditional single-teacher distillation networks (AIL, TBR, TOR), we find that relying solely on a single teacher network to transmit knowledge to students has a single source and limited features that it can cover. In contrast, our method aggregates the advantages of multiple teacher networks through clustering and integrated selection, ensuring performance while maintaining model diversity. This enables the student network to learn more comprehensive and rich knowledge, effectively demonstrating the necessity of introducing multi-teacher networks in knowledge distillation. At the same time, our method dynamically sets weights according to the performance of each teacher network, meaning that more capable teacher networks have more say in the knowledge transfer process, allowing the student network to learn more from the "high-quality information" of its outputs. However, single-teacher networks do not have such a mechanism for dynamically adjusting weights according to performance.

Compared with traditional multi-teacher networks, our method also achieves performance improvements. MTKD uses multiple teacher networks to form a single teacher network through weighted averaging. However, it is difficult to effectively distinguish and utilize the truly valuable parts in multi-teacher networks, and, instead, low-quality or repetitive information may interfere with the learning of the student network. Compared with ALLMTKD, which directly allows all teacher networks to guide students, due to the large number of teacher networks, the guidance weights are overly dispersed, meaning that the influence of each teacher network on the student network is very weak, and it is difficult to effectively transmit their own knowledge to the student network. The student network is likely to be lost in the large amount of weak and messy guiding knowledge, resulting in poor learning effects, effectively demonstrating the importance of directly introducing clustering and integrated selection.

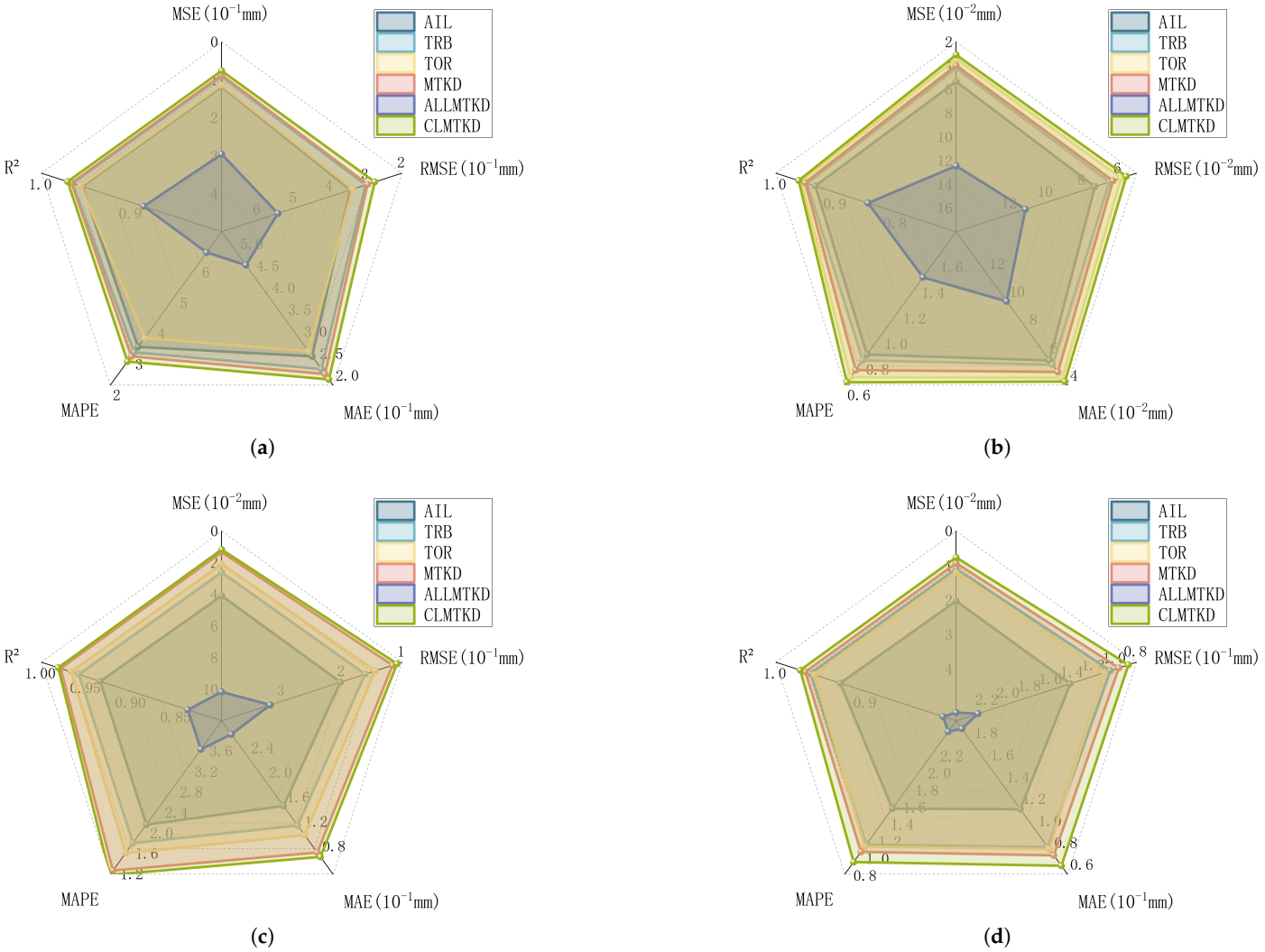

Figure 9. Radar chart showing comparison with other baseline models: (**a**) monitoring point SG19; (**b**) monitoring point SG20; (**c**) monitoring point SG25; (**d**) monitoring point SG26.

The results of the time and performance comparison of different models are shown in Figure 10. We can see that the training time of the single-teacher distillation method is shorter than that of multi-teacher networks. The main reason is that the multi-teacher distillation process needs to integrate knowledge from multiple teachers. Therefore, it usually requires a longer training time than the single-teacher distillation method. ALLMKD has the longest training time because multiple teacher networks participate in the guidance simultaneously, resulting in chaos in the knowledge transfer process, greatly increasing the training cost, and achieving the worst performance. This also demonstrates that the number of teacher networks in the distillation process is not the more, the better. The training time of the proposed CLMTKD is longer than that of MTKD. The reason is that we introduce a clustering and integration method, which consumes a part of the training time, but we achieve the best distillation results.

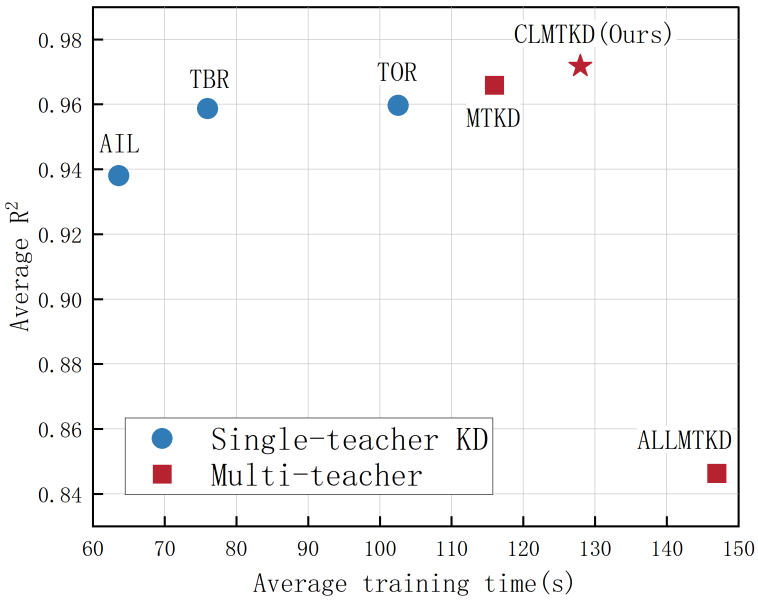

Figure 10. Deformation monitoring data of the monitoring points.

5. Conclusions

This paper proposed a multi-teacher distillation method based on clustering integration for dam displacement prediction. We discussed the advantages of multi-teacher networks over single-teacher networks in the knowledge distillation process. On this basis, a clustering selection and integration algorithm for multi-teacher networks was proposed to address issues such as the confusion and inefficiency caused by multi-teacher networks during the distillation process. Knowledge is transferred from the teacher network to the student network by minimizing the loss function. The experiments on all datasets showed that our model could effectively compress the model while improving the accuracy of the student model. Further analysis showed that we could significantly improve students' performance, and the prediction results were more consistent with the true values. In conclusion, the method we proposed can effectively improve the prediction accuracy, capture the subtle changes in dam deformation in advance, provide strong data support for dam safety management, reduce the risk of safety accidents, and ensure the safety of the lives and property of people downstream as well as the normal operation of water conservancy facilities. At present, this method was only applied to concrete face rockfill dams. In the future, we will consider exploring this framework for use with more general infrastructure such as concrete gravity dams and arch dams.

Author Contributions: Conceptualization, F.G. and J.Y.; methodology, J.Y.; software, F.G.; validation, F.G., J.Y. and D.L.; formal analysis, X.Q.; investigation, F.G.; resources, X.Q.; data curation, J.Y.; writing—original draft preparation, F.G.; writing—review and editing, J.Y.; visualization, J.Y.; supervision, J.Y.; project administration, D.L.; funding acquisition, X.Q. All authors have read and agreed to the published version of this manuscript.

Funding: This research was supported by the Guizhou Science and Technology Projects (QiankeheSupport [2023] General 251), the Guizhou Provincial Major Scientific and Technological Program (grant number QiankeheMajor[2024]008), and the Guizhou Provincial Basic Research Program (Natural Science) Youth Guidance Project (Qiankehe Basic-[2024] Youth 095).

Data Availability Statement: Data are contained within this article.

Conflicts of Interest: Author Fawang Guo was employed by the company POWERCHINA Guiyang Engineering Corporation Limited, Guiyang, China. The remaining authors declare that this research

was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflicts of interest.

References

1. Aydemir, A. Modified risk assessment tool for embankment dams: Case study of three dams in Turkey. *Civ. Eng. Environ. Syst.* **2017**, *34*, 53–67. [CrossRef]
2. Sivasuriyan, A. Health assessment of dams under various environmental conditions using structural health monitoring techniques: A state-of-art review. In *Environmental Science and Pollution Research*; Springer Nature: Berlin/Heidelberg, Germany, 2022; pp. 1–12.
3. Yang, J. An intelligent prediction model of settlement deformation for the earth-rock dam. In Proceedings of the 2021 International Conference on Artificial Intelligence, Big Data and Algorithms (CAIBDA), Xi'an, China, 28–30 May 2021; pp. 198–201.
4. Wang, Q. Research on dam deformation prediction based on variable weight combination model. In Proceedings of the 2024 4th International Conference on Neural Networks, Information and Communication (NNICE), Guangzhou, China, 19–21 January 2024; pp. 900–903.
5. Luo, C. Research on Dam Deformation Prediction Model Based on Wavelet Neural Network. In Proceedings of the 2021 3rd International Conference on Applied Machine Learning (ICAML), Changsha, China, 23–25 July 2021; pp. 362–365.
6. Zhu, Y. Automatic damage detection and diagnosis for hydraulic structures using drones and artificial intelligence techniques. *Remote Sens.* **2023**, *15*, 615. [CrossRef]
7. Zhu, M. Optimized multi-output LSSVR displacement monitoring model for super high arch dams based on dimensionality reduction of measured dam temperature field. *Eng. Struct.* **2022**, *268*, 114686. [CrossRef]
8. Mata, J. Constructing statistical models for arch dam deformation. *Struct. Control Health Monit.* **2014**, *21*, 423–437. [CrossRef]
9. Chen, R. Construction and selection of deformation monitoring model for high arch dam using separate modeling technique and composite decision criterion. *Struct. Health Monit.* **2024**, *23*, 2509–2530. [CrossRef]
10. Wu, B. A Multiple-Point Deformation Monitoring Model for Ultrahigh Arch Dams Using Temperature Lag and Optimized Gaussian Process Regression. *Struct. Control Health Monit.* **2024**, *2024*, 2308876. [CrossRef]
11. Xing, Y. Research on dam deformation prediction model based on optimized SVM. *Processes* **2022**, *10*, 1842. [CrossRef]
12. Li, X. An approach using random forest intelligent algorithm to construct a monitoring model for dam safety. *Eng. Comput.* **2021**, *37*, 39–56. [CrossRef]
13. Hipni, A. Daily forecasting of dam water levels: Comparing a support vector machine (SVM) model with adaptive neuro fuzzy inference system (ANFIS). *Water Resour. Manag.* **2013**, *27*, 3803–3823. [CrossRef]
14. Yang, D. A concrete dam deformation prediction method based on LSTM with attention mechanism. *IEEE Access* **2020**, *8*, 185177–185186. [CrossRef]
15. Yang, J. A CNN-LSTM model for tailings dam risk prediction. *IEEE Access* **2020**, *8*, 206491–206502.
16. Jiedeerbieke, M. Gravity Dam Deformation Prediction Model Based on I-KShape and ZOA-BiLSTM. *IEEE Access* **2024**, *12*, 50710–50722. [CrossRef]
17. Su, Y. Dam deformation interpretation and prediction based on a long short-term memory model coupled with an attention mechanism. *Appl. Sci.* **2021**, *11*, 6625. [CrossRef]
18. Zhou, Y. Multi-expert attention network for long-term dam displacement prediction. *Adv. Eng. Informatics* **2023**, *57*, 102060.
19. Wang, X. Study on MPGA-BP of gravity dam deformation prediction. *Math. Probl. Eng.* **2017**, *2017*, 2586107.
20. Hinton, G. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
21. Sun, T. Explainability-based knowledge distillation. *Pattern Recognit.* **2025**, *159*, 111095.
22. Chu, Y. Bi-directional contrastive distillation for multi-behavior recommendation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing: Cham, Switzerland, 2022; pp. 491–507.
23. Lopez-Paz, D. Unifying distillation and privileged information. *arXiv* **2015**, arXiv:1511.03643.
24. Yuan, Y. The Impact of Knowledge Distillation on the Energy Consumption and Runtime Efficiency of NLP Models. In Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI, Lisbon, Portugal, 14–15 April 2024; pp. 129–133.
25. Léger, P. Hydrostatic, temperature, time-displacement model for concrete dams. *J. Eng. Mech.* **2007**, *133*, 267–277.
26. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
27. Xu, C. A financial time-series prediction model based on multiplex attention and linear transformer structure. *Appl. Sci.* **2023**, *13*, 5175. [CrossRef]
28. Ahmed, S. Transformers in time-series analysis: A tutorial. *Circuits Syst. Signal Process.* **2023**, *42*, 7433–7466. [CrossRef]
29. Sasal, L. W-transformers: A wavelet-based transformer framework for univariate time series forecasting. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), Nassau, Bahamas, 12–14 December 2022; pp. 671–676.

30. Wen, Q. Transformers in time series: A survey. *arXiv* **2022**, arXiv:2202.07125.
31. Hittawe, M.M. Stacked Transformer Models for Enhanced Wind Speed Prediction in the Red Sea. In Proceedings of the 2024 IEEE 22nd International Conference on Industrial Informatics (INDIN), Beijing, China, 18–20 August 2024; pp. 1–7.
32. Jiang, W. Applicability analysis of transformer to wind speed forecasting by a novel deep learning framework with multiple atmospheric variables. *Appl. Energy* **2024**, *353*, 122155. [CrossRef]
33. Lin, Y. ATMKD: Adaptive temperature guided multi-teacher knowledge distillation. *Multimed. Syst.* **2024**, *30*, 292. [CrossRef]
34. Zhang, H. Adaptive multi-teacher knowledge distillation with meta-learning. In Proceedings of the 2023 IEEE International Conference on Multimedia and Expo (ICME), Brisbane, Australia, 10–14 July 2023; pp. 1943–1948.
35. Sarıyıldız, M.B. UNIC: Universal Classification Models via Multi-teacher Distillation. In *European Conference on Computer Vision*; Springer Nature: Cham, Switzerland, 2024; pp. 353–371.
36. Li, R. Cross-View Gait Recognition Method Based on Multi-Teacher Joint Knowledge Distillation. *Sensors* **2023**, *23*, 9289. [CrossRef]
37. Yang, Y. MKDAND: A network flow anomaly detection method based on multi-teacher knowledge distillation. In Proceedings of the 2022 16th IEEE International Conference on Signal Processing (ICSP), Beijing, China, 21–24 October 2022; Volume 1, pp. 314–319.
38. Pham, C. Collaborative multi-teacher knowledge distillation for learning low bit-width deep neural networks. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 6435–6443.
39. Ye, X. Knowledge distillation via multi-teacher feature ensemble. *IEEE Signal Process. Lett.* **2024**, *31*, 566–570. [CrossRef]
40. Wu, C. One teacher is enough? Pre-trained language model distillation from multiple teachers. *arXiv* **2021**, arXiv:2106.01023.
41. Zhang, J. Badcleaner: Defending backdoor attacks in federated learning via attention-based multi-teacher distillation. *IEEE Trans. Dependable Secur. Comput.* **2024**, *21*, 4559–4573. [CrossRef]
42. Bai, L. An information-theoretical framework for cluster ensemble. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 1464–1477. [CrossRef]
43. Lee, K. Fast and accurate facial expression image classification and regression method based on knowledge distillation. *Appl. Sci.* **2023**, *13*, 6409. [CrossRef]
44. Bai, S. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv* **2018**, arXiv:1803.01271.
45. Saputra, M.R.U. Distilling knowledge from a deep pose regressor network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 263–272.
46. Chen, G.; Choi, W.; Yu, X.; Han, T.; Chandraker, M. Learning efficient object detection models with knowledge distillation. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 742–751.
47. Takamoto, M. An efficient method of training small models for regression problems with knowledge distillation. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 6–8 August 2020; pp. 67–72.