Energy Conversion and Management: X 28 (2025) 101391

Contents lists available at ScienceDirect

Energy Conversion and Management: X

journal homepage: www.sciencedirect.com/journal/energy-conversion-and-management-x

# Veliformer: A periodicity-preserving model for short-term tidal energy forecasting in optimal power flow

Yangdi Huang [a], Lina Yang [b], Xinzhang Wu [a], Yunxuan Dong [a,*]

[a] School of Computer and Electronic Information, Guangxi University, Nanning, Guangxi 530000, China
[b] School of Electrical Engineering, Guangxi University, Nanning, Guangxi 530000, China

ARTICLE INFO

*Keywords:*
Tidal energy
Optimal power flow
Transformer
Masked modeling

ABSTRACT

With the increasing integration of renewable energy, tidal energy stands out for its high predictability, making it a valuable asset for stable power grid operation. However, accurate forecasting remains a critical challenge. Conventional deep learning models, despite their success in general time-series analysis, often struggle to preserve the inherent periodic features of tidal data, leading to reduced prediction accuracy and suboptimal grid scheduling. To address this gap, we propose Veliformer, a novel periodicity-preserving forecasting model. At its core, Veliformer introduces an innovative mask modeling technique. Unlike conventional methods that predict masked data points, our approach reconstructs the complete original sequence by learning to aggregate information from multiple, differently masked versions of the series. This unique reconstruction process is specifically designed to maintain the integrity of the underlying periodic structure of tidal energy, enabling the model to accurately capture both deterministic cycles and stochastic fluctuations. When applied to the optimal power flow (OPF) of tidal energy systems, Veliformer reduces power generation costs. Our theoretical analysis shows that the model preserves periodicity through masked sequence reconstruction. Numerical experiments demonstrate Veliformer's superior performance in optimizing power systems and reducing prediction errors compared to other popular models. The mask modeling mechanism enhances Veliformer's prediction accuracy by an average of 4.91%, further highlighting its effectiveness in handling tidal energy forecasting.

## 1. Introduction

In recent years, the development of renewable energy sources has received increasing attention due to the growing demand for energy conservation. Renewable energy is characterized as sustainable and non-polluting, which can solve the problems of energy supply and environmental pollution [1]. Among renewable energy sources, wind and solar power have been widely integrated into modern power grids [2,3]. With the maturity of tidal power generation technology, tidal energy has also garnered significant attention. Derived from the gravitational forces of the Earth, Moon, and Sun, the total global tidal energy potential is estimated at 2,700 GW, with approximately 2% (54 GW) being exploitable [4]. Compared to other renewable sources, tidal energy is more influenced by astronomical factors, resulting in clear and stable periodic patterns. This inherent predictability provides a reliable basis for grid scheduling and energy planning. As the potential for tidal energy exploitation grows, the expanding market makes accurate forecasting and the optimization of scheduling strategies increasingly important (see in Fig. 1, data source: Global Market Insights, 2025).

The integration of tidal energy into modern power systems presents a unique challenge, primarily centered around minimizing operational costs while ensuring security and economy, a task addressed by Optimal Power Flow (OPF). OPF is a critical tool for determining the optimal dispatch in a power network to minimize operational costs under given safety and performance constraints. Tidal energy is characterized by a highly predictable semidiurnal (approximately 12-hour) cycle, resulting in two pronounced power peaks and troughs each day. In OPF-based scheduling, tidal generation is thus modeled as a source with large, predictable but rapid fluctuations. The grid must accommodate these swings by coordinating other controllable generators to ramp up or down, ensuring real-time supply–demand balance [5]. The complexity in OPF arises not from unpredictability, but from the need to economically and securely manage these dramatic, periodic swings. Even small errors in short-term tidal power forecasting – such as phase or amplitude mismatches – can lead to significant deviations in dispatch plans, increasing reliance on costly reserves and increasing operational costs [6]. Therefore, enhancing prediction accuracy while preserving

* Corresponding author.
*E-mail address:* dyxiscool@outlook.com (Y. Dong).

https://doi.org/10.1016/j.ecmx.2025.101391
Received 6 June 2025; Received in revised form 26 September 2025; Accepted 7 November 2025
Available online 8 November 2025

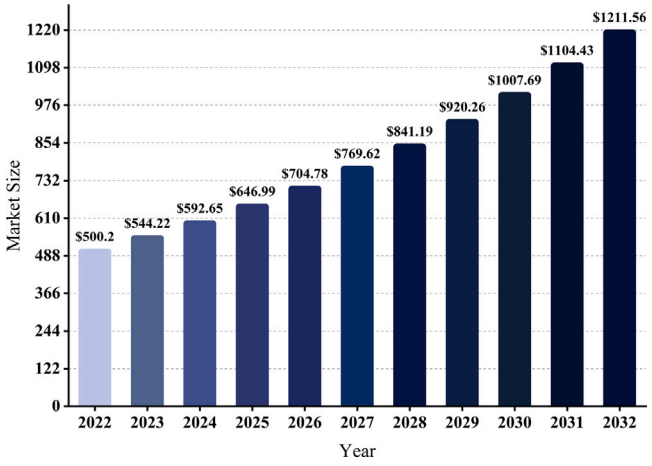


**Fig. 1.** Wave and tidal energy market size 2022 to 2032 (USD Million).

the inherent periodicity of tidal energy is critical for effective and economical grid operation.

While accurate forecasting is a recognized prerequisite for OPF, recent advancements in renewable energy forecasting have largely focused on challenges pertinent to wind and solar power, such as managing uncertainty and stochastic variability. For instance, Ref. [7] demonstrated the use of a binary prediction market to achieve probabilistic renewable energy forecasts, leveraging aggregated probabilities to enhance forecasting precision. To improve forecasting accuracy in wind power systems, Ref. [8] proposed a robust ensemble learning method that combined random forests and quantile arrays. Additionally, advanced machine learning techniques explored to enhance the adaptability of forecasting models. Ref. [9] presented an adaptive probabilistic wind power forecasting method, incorporating offline meta-learning for model training and online learning for real-time updates, showcasing flexibility across different lead times. Moreover, hybrid models integrating signal processing techniques like Singular Spectrum Analysis (SSA) with intelligent systems have been effectively applied to short-term wind speed forecasting [10]. For more robust interval predictions, Ref. [11] introduced a multi-objective optimization approach to construct wind power prediction intervals. Similarly, multi-objective optimization has been effectively utilized to enhance short-term power load forecasting models [12]. Despite these significant advancements, the core focus of these methods remains on handling unpredictability. Consequently, they do not adequately address the unique challenge of preserving the strong, deterministic periodic features inherent in tidal energy, as their mechanisms may inadvertently smooth out or misinterpret these crucial patterns.

Addressing this specific gap, a few recent studies have attempted to forecast tidal energy more explicitly. However, even these targeted efforts exhibit limitations in robustly preserving periodicity. For example, the hybrid point-interval forecasting system proposed in Ref. [13], despite incorporating sophisticated techniques like mode decomposition and attention mechanisms to capture complex patterns, does not guarantee the explicit preservation of fundamental tidal frequencies throughout its modeling pipeline. The interaction between decomposed modes and attention weights, or subtle shifts in data distribution, could inadvertently dampen or distort these vital periodic components. Likewise, while the use of deep neural networks with environmental variables [14] or Long Short-Term Memory (LSTM) networks for short-term tidal height forecasting [15] has demonstrated promise in terms of regression accuracy, their architectural designs do not inherently enforce the maintenance of the underlying temporal structure. Standard recurrent or deep feed-forward layers, if not specifically structured or regularized for periodicity, can struggle to distinguish between true periodic signals and aperiodic noise, particularly when dealing with long sequences or the non-stationary effects of environmental variables. Consequently, these models may capture average trends but falter in accurately predicting the timing and magnitude of periodic fluctuations.

To overcome the aforementioned limitations, we propose Veliformer, a novel deep learning model that incorporates a unique pre-training strategy designed to retain the natural periodicity of tidal energy data. Pre-training aims to learn the features of the data by training the model on large-scale data, and we summarize the differences in common approaches in Table 1. Based on this comparison, we selected masked modeling as the foundation for Veliformer. However, typical masking techniques, which reconstruct masked portions from visible data, can disrupt the temporal continuity crucial for time series [16]. To solve this, Veliformer introduces a novel objective: reconstructing the original from multiple, differently masked versions of it. This forces the model to learn the underlying periodic structure and temporal dependencies, rather than simply interpolating missing points. By applying this periodicity-preserving pre-training method to the powerful Transformer architecture, Veliformer is able to capture both long-range dependencies and the fundamental periodic patterns of tidal energy.

Based on the above insights, this paper proposes Veliformer to optimize tidal energy integration with the dual objectives of minimizing operational costs and maintaining system security. The specific contributions are as follows:

(1) To tackle the issue of periodicity preservation, Veliformer introduces a pre-training method that reconstructs the original series from multiple neighboring masked time series. This technique retains periodic features and enhances the accuracy of tidal energy power prediction when applied to the Transformer model.

(2) Veliformer is applied within the Optimal Power Flow (OPF) framework for power systems that integrate tidal energy. By leveraging deep learning techniques, Veliformer effectively minimizes operational costs while ensuring system security.

(3) A comprehensive comparative analysis is conducted between Veliformer and several popular deep learning models. The experimental results demonstrate Veliformer's superior performance in predicting tidal power generation, further underscoring its distinct advantages in both prediction accuracy and overall system optimization.

## 2. Method

### 2.1. Optimal power flow model with tidal energy

The Optimal Power Flow problem is a typical nonlinear programming challenge. Before describing the problem in detail, we introduce the vector $\mathbf{x}$ to denote the time series of active power. The input vector $\mathbf{x}^{\tau}(\mathbf{t})$ represents a mini-batch of time series samples. Each sample is a $\tau \times C$ matrix, where $\tau$ denotes the number of time points and $C$ denotes the number of observed variables. The input vector $\mathbf{x}(\mathbf{t})$ represents the $t$th row of the mini-batch matrix. We denote $\mathbf{x}(\mathbf{t})$ as $\mathbf{x}(\mathbf{t}) = \{x_1(t), \ldots, x_c(t), \ldots, x_C(t)\}$, where $t \in [1, T]$. And we express the scalar average of each vector $\mathbf{x}(\mathbf{t})$ as $\bar{x}(t) = \frac{1}{C}\sum_{c=1}^{C} x_c(t)$. The vector $\mathbf{x}_{\mathbf{c}}^{\tau}(\mathbf{t})$ represents the $c$th column, containing $\tau$ scalars, and is denoted as $\mathbf{x}_{\mathbf{c}}^{\tau}(\mathbf{t}) = \{x_c(t-\tau+1), x_c(t-\tau+2), \ldots, x_c(t)\}$. Then, We can slice this mini-batch data into $\mathbf{x}^{\tau}(\mathbf{t}) = \{\mathbf{x}_{\mathbf{1}}^{\tau}(\mathbf{t}), \ldots, \mathbf{x}_{\mathbf{c}}^{\tau}(\mathbf{t}), \ldots, \mathbf{x}_{\mathbf{C}}^{\tau}(\mathbf{t})\}$, where $c \in [1, C]$. The mathematical formulation of the AC optimal power flow problem adopted in this work is consistent with standard formulations in power systems literature [19], which can be described as follows:

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & h_a(\mathbf{x}) = 0, \quad a = 1, 2, \ldots, A, \\ & \underline{g} \le g_b(\mathbf{x}) \le \bar{g}, \quad b = 1, 2, \ldots, B, \end{aligned} \tag{1}$$

where $\mathbf{x}$ is the decision variable, $f(\mathbf{x})$ is the objective function with respect to the variable $\mathbf{x}$, $h_a(\mathbf{x})$ characterizes all equality constraints on the variable $\mathbf{x}$, and $g_b(\mathbf{x})$ characterizes all inequality constraints

**Table 1**
Comparison of different pre-training methods.

| Type | Description | Advantages | Disadvantages |
|---|---|---|---|
| Masked Modeling [16] | A pre-training method where parts of the input data are randomly masked, and the model learns to predict the masked parts. | Encourages the model to understand context and develop a deeper understanding of the data structure. | Can be computationally intensive and requires substantial data to avoid bias in predictions. |
| Contrastive Learning [17] | Learns by comparing pairs or sets of inputs to understand which features make two inputs similar or different. | Effective in learning robust features and useful for tasks requiring fine-grained distinction between inputs. | Requires careful design of the contrast sets and can be less effective if the negative examples are not well chosen. |
| Self-Supervised Learning [18] | The model learns by creating its own supervision signals from the data, such as predicting missing parts of data or solving puzzles generated from the data itself. | Reduces the need for labeled data and improves generalization by leveraging inherent data structures. | Can be challenging to design effective self-supervision tasks and may require significant computational resources. |

on the variable $\mathbf{x}$. $\underline{g}$ and $\bar{g}$ are the upper and lower bounds of the inequality constraints, respectively. $A$ and $B$ denote the number of equality constraints and inequality constraints respectively.

The objective function $f(\mathbf{x})$ can be determined according to the problem to be solved. In this paper, our goal is to evaluate the overall economic benefits of tidal energy projects when enhanced with high-precision forecasting models. Therefore, we adopt a cost function based on the Levelized Cost of Energy (LCOE) framework, which annualizes both capital and operational costs to assess the long-term impact of different scheduling strategies [20]. The function is defined as:

$$f(\mathbf{x}) = \frac{\text{CAPEX} \cdot FCR + \text{OPEX} + C_{\text{battery}}}{\text{AEP}}. \tag{2}$$

Here, CAPEX, OPEX, and FCR denote the capital expenditure, operating expenditure, and capital recovery factor of the power plant, respectively. AEP represents the annual energy production, which depends on the power output $\mathbf{x}$ over time. This formulation allows us to assess the impact on the average unit generation cost, rather than focusing solely on short-term fuel savings [21]. Furthermore, the term $C_{\text{battery}}$ represents the cost related to the energy storage system. Energy storage is an essential enabling technology for managing the high variability of tidal energy, allowing for peak shaving and valley filling. Including its cost is crucial to fully exploit the scheduling flexibility and economic potential brought by high-precision forecasting [22]. The detailed calculations for each component are provided as follows:

$$FCR = \frac{(1+r)^l \cdot r}{(1+r)^l - 1}. \tag{3}$$

Where $r$ refers to the discount rate, and $l$ denotes the economic lifespan of the power plant.

$$CAPEX = C_{\text{upfront}} + C_{\text{TCTs}} + C_{EI} + C_{\text{install}} + C_{\text{offshore}}. \tag{4}$$

In this equation, $C_{\text{upfront}}$ represents the upfront investment cost. $C_{TCTs}$, $C_{EI}$, $C_{\text{install}}$, and $C_{\text{offshore}}$ correspond to the costs of the tidal current system, power units, installation, and offshore construction, respectively.

$$OPEX = C_0 + C_{FR} + C_{FM}. \tag{5}$$

Where $C_0$, $C_R$, and $C_M$ denote the initial operation cost, maintenance cost, and repair cost, respectively.

$$C_{\text{battery}} = \kappa \cdot \left(P^{\text{ch}}_{\text{total}} + P^{\text{dis}}_{\text{total}}\right) + \gamma \cdot \text{SoC}^2. \tag{6}$$

Where $\kappa$ is a constant related to the cost per unit of charge and discharge power, $P^{\text{ch}}_{\text{total}}$ and $P^{\text{dis}}_{\text{total}}$ are the total charge and discharge powers of the storage system, respectively. $\gamma$ is a constant related to the degradation cost due to cycling. And SoC is the state of charge of the battery, which is related to the degradation cost.

The nodal power balance equations $h(\mathbf{x})$ is the tidal current equation of the following form:

$$\begin{aligned} P_{Gk} - P_{Dk} - P_k\left(V_e, V_f\right) = 0, \\ Q_{Gk} - Q_{Dk} - Q_k\left(V_e, V_f\right) = 0, \end{aligned} \tag{7}$$

where $k = 1, 2, \ldots, n$; $P_{Gk}$, $Q_{Gk}$ are the active and reactive generator outputs connected at node $k$; $P_{Dk}$, $Q_{Dk}$ are the active and reactive loads connected at node $k$; $P_k$, $Q_k$ are the active and reactive power injections at node $k$; $V_e$, $V_f$ are the real and imaginary part of the voltages at nodes.

The inequality constraint $g(\mathbf{x})$ mainly includes the generator's active and reactive power output constraints:

$$\begin{aligned} P^{\min}_{Gk} \le P_{Gk} \le P^{\max}_{Gk}, \\ Q^{\min}_{Gk} \le Q_{Gk} \le Q^{\max}_{Gk}. \end{aligned} \tag{8}$$

Nodal voltage amplitude constraint:

$$V^{\min}_k \le V_k \le V^{\max}_k, \tag{9}$$

and the line transmission power constraint:

$$\begin{aligned} P^{\min}_{kj} \le P_{kj} \le P^{\max}_{kj}, \\ P^{\min}_{jk} \le P_{jk} \le P^{\max}_{jk}. \end{aligned} \tag{10}$$

Where $P^{\min}_{Gk}$, $P^{\max}_{Gk}$ $\left(Q^{\min}_{Gk}, Q^{\max}_{Gk}\right)$ represent the lower and upper limits of generator $k$'s active (reactive) output; $V^{\min}_k$ is the lower limit of the voltage amplitude at node $k$. Similarly, $V^{\max}_k$ is the upper limit of the voltage amplitude at node $k$; $P_{kj}$, $P_{jk}$ denote the active power flowing from node $k$ to node $j$ and the active power flowing from node $j$ to node $k$, respectively; $P_{kj}{}^{\min}$, $P_{kj}{}^{\max}$ $\left(P^{\min}_{jk}, P^{\max}_{jk}\right)$ are the upper and lower limits of the active power of line $kj$ $(jk)$ respectively.

Energy storage capacity constraints:

$$E_{\min} \le E_t \le E_{\max}. \tag{11}$$

Charging and discharging Power Constraints:

$$\begin{aligned} 0 \le P^{\text{ch}}_t \le P^{\text{ch}}_{\max}, \\ 0 \le P^{\text{dis}}_t \le P^{\text{dis}}_{\max}. \end{aligned} \tag{12}$$

Where $E_t$ is the energy state of charge (SoC) at time $t$. $E_{\min}$ and $E_{\max}$ are the minimum and maximum energy storage limits, respectively. $P^{\text{ch}}_t$ and $P^{\text{dis}}_t$ represent the charging and discharging powers of the energy storage at time $t$. $P^{\text{ch}}_{\max}$ and $P^{\text{dis}}_{\max}$ are the maximum charging and discharging power limits, respectively.

### 2.2. OPF solution methodology

To solve the OPF problem (1) formulated in Section 2.1, we employ a standard Primal–Dual Interior-Point Method, a well-established and robust algorithm for nonlinear constrained optimization problems in power systems [23]. The core idea of this method is to convert the original problem with inequality constraints into a sequence of equality-constrained problems by introducing slack variables and a logarithmic barrier function. At each iteration, the Karush–Kuhn–Tucker (KKT) conditions are solved using a Newton–Raphson method to find the search direction for all variables. For a detailed mathematical derivation of this standard method as implemented in our study, please refer to Appendix A.

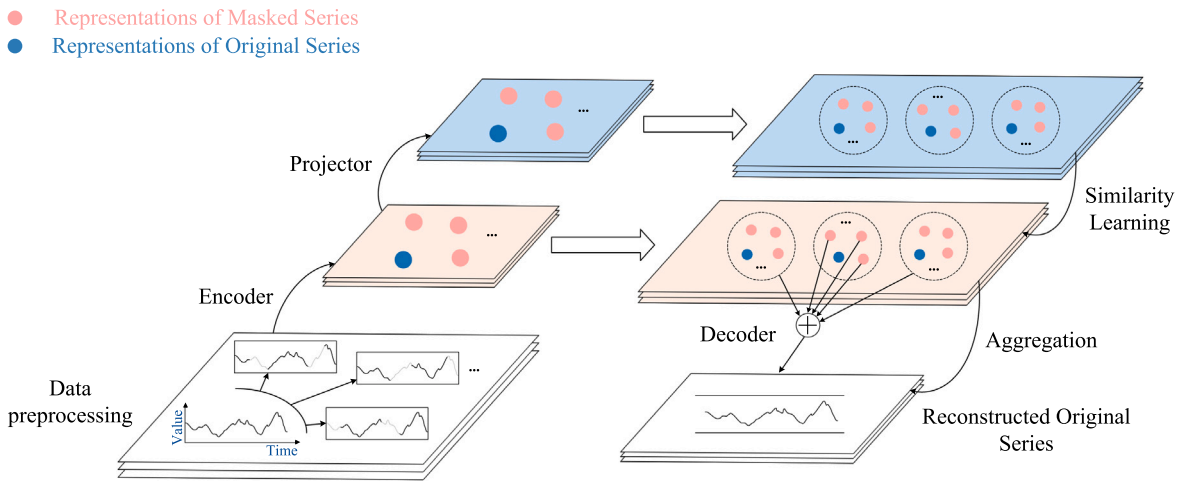


**Fig. 2.** Architecture of Veliformer, which reconstructs the original time series by adaptively aggregating multiple masked versions of the series. The aggregation process is driven by series-wise similarities.

### 2.3. Modeling based on masked mechanism

The richness of time series features is crucial for improving prediction accuracy. In this paper, we enhance the diversity of training data by adopting a masking mechanism, which significantly improves the accuracy of the prediction model. Specifically, we reconstruct the original time series from multiple adjacent masked time series. The overall framework of Veliformer is illustrated in Fig. 2.

Next, we describe the detailed process of modeling based on the masking mechanism. First, we need to generate the masked series. $\mathbf{x_i}$ represents a mini-batch of $N$ time series samples. We can generate a set of mask series for each sample $\mathbf{x_i}$ by randomly masking a portion of time points in the time dimension. The detailed formulation is presented as follows:

$$\begin{aligned}\left\{\bar{\mathbf{x}}_{\mathbf{i}}^{\mathbf{j}}\right\}_{j=1}^{M} &= \left\{\mathrm{Mask}_r(\mathbf{x_i})\right\}_{j=1}^{M} \\ &= \left\{\left(m_c(t)^j \cdot x_c(t)\right)_{c=1,t=1}^{C,\tau}\right\}_j^M .\end{aligned} \tag{13}$$

Eq. (13) denotes the generation of $M$ random masked series for each $\mathbf{x_i}$. $M$ is the hyperparameter of the number of masked time series, which indicates how many different random masked series are generated for each $\mathbf{x_i}$. And $\bar{\mathbf{x}}_{\mathbf{i}}^{\mathbf{j}}$ denotes the $j$th masked time series of $\mathbf{x_i}$. $m_c(t)^j$ is a random binary variable generated using a geometric distribution satisfying the following conditions:

$$m_c(t)^j = \begin{cases} 0, & \text{with probability } r, \\ 1, & \text{with probability } 1-r, \end{cases} \tag{14}$$

where $r$ denotes the masking part, $r$ is a decimal number between $0 \sim 1$, which indicates the proportion of the masked part in the total data length. The detailed procedure is summarized in Algorithm 1. Finally, a total of $N * (M+1)$ series can be obtained by randomly masking the $N$ time series and adding the original series. That is,

$$\mathcal{X} = \bigcup_{i=1}^{N}\left(\{\mathbf{x_i}\} \cup \left\{\bar{\mathbf{x}}_{\mathbf{i}}^{\mathbf{j}}\right\}_{j=1}^{M}\right). \tag{15}$$

Then $\mathcal{X}$ is passed through an encoder to get $\mathcal{Z}$, and $\mathcal{Z}$ is passed through a projection layer to get $\mathcal{S}$. $\mathcal{Z}$ is a feature vector of $\mathcal{X}$, and the role of $\mathcal{S}$ is to learn the similarity between features. The following equations formally describe the transformations from $\mathcal{X}$ to $\mathcal{Z}$ and from $\mathcal{Z}$ to $\mathcal{S}$.

$$\begin{aligned}\mathcal{Z} &= \textstyle\bigcup_{i=1}^{N}\left(\{\mathbf{z_i}\} \cup \left\{\bar{\mathbf{z}}_{\mathbf{i}}^{\mathbf{j}}\right\}_{j=1}^{M}\right) = \mathrm{Encoder}(\mathcal{X}), \\ \mathcal{S} &= \textstyle\bigcup_{i=1}^{N}\left(\{\mathbf{s_i}\} \cup \left\{\bar{\mathbf{s}}_{\mathbf{i}}^{\mathbf{j}}\right\}_{j=1}^{M}\right) = \mathrm{Projector}\ (\mathcal{Z}).\end{aligned} \tag{16}$$

**Algorithm 1** Geometric Masking

**Input:** $\mathbf{x_i}$ (The original time series to be masked), $lm$ (The average length of the masking subsequence), $r$ (The ratio of the series to be masked)
**Output:** *masked_sequence* (The series with random masking applied)
**Mask Generation:**
$L \leftarrow$ length of $\mathbf{x_i}$
*keep_mask* $\leftarrow$ array of True values with length $L$
$p_m \leftarrow 1/lm$
$p_u \leftarrow p_m \times r/(1-r)$
$p \leftarrow [p_m, p_u]$
Generate a random number between 0 and 1
**if** the random number is greater than $r$ **then**
    Set *state* $\leftarrow 1$
**else**
    Set *state* $\leftarrow 0$
**end if**
**for** each $i$ from 0 to $L-1$ **do**
    *keep_mask*[$i$] $\leftarrow$ *state*
    **if** random value $<$ $p$[*state*] **then**
        *state* $\leftarrow 1 -$ *state*
    **end if**
**end for**
*masked_sequence* $\leftarrow \mathbf{x_i} \times$ *keep_mask*
**return** *masked_sequence*

The encoder used in this paper is transformer and the projection layer is a simple MLP. Fig. 3 further details the Veliformer's reconstruction pipeline, showcasing how encoded and projected series representations are processed through similarity learning, aggregation, and a final decoder to yield the reconstructed original time series.

Using the series-level representation of the similarity between weighted aggregation, we get

$$\begin{aligned}\mathbf{R} &= \mathrm{Sim}(\mathcal{S}), \\ \mathbf{R}_{\mathbf{u},\mathbf{v}} &= \frac{\mathbf{u}\mathbf{v}^{\top}}{\cos\|\mathbf{v}\|\|\mathbf{u}\|},\end{aligned} \tag{17}$$

where $\mathbf{R}$ is the pairwise similarity matrix of $(N \times (M+1))$ input samples in the series representation space, with matrix size $(N \times (M+1)) \times (N \times (M+1))$. $\mathbf{u}$ and $\mathbf{v}$ are feature vectors from $\mathcal{S}$. And similarity is measured by the cosine similarity. Based on the learned series similarity, the

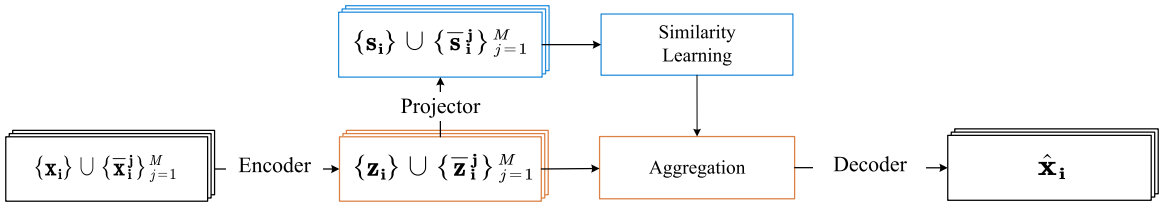


**Fig. 3.** Detailed schematic of Veliformer's reconstruction pipeline, which emphasizes the process of learning similarities between series representations, adaptively aggregating information from masked sequences, and subsequently decoding the aggregated representation to reconstruct the original time series.

aggregation process for the $i$th original time series is as follows:

$$\hat{\mathbf{z}}_i = \sum_{s' \in S \setminus \{s_i\}} \frac{\exp(\mathbf{R}_{s_i,s'}/\tau)}{\sum_{s'' \in S \setminus \{s_i\}} \exp(\mathbf{R}_{s_i,s''}/\tau)} \mathbf{z}', \tag{18}$$

where $\mathbf{z}'$ represents the corresponding point-wise representation of $s'$, $\hat{\mathbf{z}}_i$ is the reconstructed point-wise representation, and $\tau$ denotes the temperature hyperparameter for softmax normalization of series-wise similarities. Finally, after the decoder, the reconstructed original time series is obtained.

$$\{\hat{\mathbf{x}}_\mathbf{i}\}_{i=1}^{N} = \text{Decoder}\left(\{\hat{\mathbf{z}}_i\}_{i=1}^{N}\right), \tag{19}$$

where $\hat{\mathbf{x}}_\mathbf{i}$ is the reconstruction to $\mathbf{x}_\mathbf{i}$. The decoder is instantiated as a simple MLP layer along the channel dimension.

### 2.4. Periodic holding in Veliformer mask reconstruction

To establish the periodicity preservation of the Veliformer model during masked reconstruction, we first define periodicity formally. Let $\mathbf{x}(t)$ be a discrete time series with a fundamental period $T_p$, satisfying:

$$\mathbf{x}(t) = \mathbf{x}\left(t + T_p\right), \quad \forall t \tag{20}$$

Where $T_p$ represents the fundamental period. A masking mechanism is applied to generate $M$ masked versions of the series. These masked versions are denoted as $\left\{\overline{\mathbf{x}}^{(j)}(t)\right\}_{j=1}^{M}$, and each version is defined as:

$$\overline{\mathbf{x}}^{(j)}(t) = m^{(j)}(t) \cdot \mathbf{x}(t). \tag{21}$$

The variables $m^{(j)}(t)$ are binary random variables that are independently and identically distributed (i.i.d.), satisfying:

$$\mathbb{E}\left[m^{(j)}(t)\right] = p, \quad \text{Var}\left[m^{(j)}(t)\right] = p(1-p). \tag{22}$$

The reconstructed series $\hat{\mathbf{x}}(t)$ is defined as:

$$\hat{\mathbf{x}}(t) = \frac{1}{M} \sum_{j=1}^{M} \overline{\mathbf{x}}^{(j)}(t). \tag{23}$$

The autocorrelation function of the reconstructed series is defined as:

$$R_{\hat{\mathbf{x}}}(\tau) = \mathbb{E}[\hat{\mathbf{x}}(t) \cdot \hat{\mathbf{x}}(t+\tau)]. \tag{24}$$

Expanding Eq. (24), we have:

$$R_{\hat{\mathbf{x}}}(\tau) = \frac{1}{M^2} \sum_{j=1}^{M} \sum_{k=1}^{M} \mathbb{E}\left[\overline{\mathbf{x}}^{(j)}(t) \cdot \overline{\mathbf{x}}^{(k)}(t+\tau)\right]. \tag{25}$$

Under the assumptions of independence between different mask realizations and between the masks and the signal, the expectation of every term in the double summation is identical and evaluates to $p^2 R_{\mathbf{x}}(\tau)$. Therefore, the summation over all $M^2$ terms simplifies as follows:

$$R_{\hat{\mathbf{x}}}(\tau) = \frac{1}{M^2} \sum_{j=1}^{M} \sum_{k=1}^{M} \left(p^2 R_{\mathbf{x}}(\tau)\right) = \frac{1}{M^2} \cdot M^2 \cdot p^2 R_{\mathbf{x}}(\tau) = p^2 R_{\mathbf{x}}(\tau). \tag{26}$$

Thus, the autocorrelation function of the reconstructed series is proportional to that of the original series, preserving its periodicity.

When the mask intensity factor $p$ is nonzero, the autocorrelation function of the reconstructed series retains the periodicity of the original series, scaled by $p^2$. This proves that the masking and reconstruction mechanism in Veliformer effectively preserves the periodic structure of time series.

## 3. Experiment

### 3.1. Forecasting task formulation

To clarify the predictive capabilities of Veliformer and address potential confusion regarding its input–output structure, we formally define the forecasting task and distinguish between the model's pre-training and fine-tuning phases.

**Task Definition:** The forecasting task involves predicting future multivariate tidal velocity sequences based on historical observations. Specifically, given a historical sequence of multivariate tidal velocity data $\mathbf{X}_{input} = \{\mathbf{x}(t-T_{in}+1), \mathbf{x}(t-T_{in}+2), \ldots, \mathbf{x}(t)\}$ spanning $T_{in}$ time steps, where each $\mathbf{x}(t) = [x_1(t), x_2(t), x_3(t)]^T$ represents the three-dimensional velocity components (X1, Y1, Z1) at time $t$, the model aims to predict the future sequence $\mathbf{X}_{output} = \{\mathbf{x}(t+1), \mathbf{x}(t+2), \ldots, \mathbf{x}(t+T_{out})\}$ over the next $T_{out}$ time steps.

**Input–Output Specification:** In this study, the input sequence length $T_{in}$ was adjusted based on the forecasting horizon $T_{out}$ to provide sufficient historical information for the model. The configurations were as follows:

- *For short-term horizons:* To predict $T_{out}$ of 10 time steps (10 s), 300 time steps (5 min), and 600 time steps (10 min), we set the input length to $T_{in} = 720$ time steps (corresponding to 12 min of historical data).
- *For longer-term horizons:* To predict $T_{out}$ of 1200 time steps (20 min), 1800 time steps (30 min), and 3600 time steps (60 min), we used an increased input length of $T_{in} = 7200$ time steps (corresponding to 2 h of historical data).

In all cases, the model simultaneously processes all three velocity components (X1, Y1, Z1) as a unified multivariate input and generates predictions for all three components.

**Two-Stage Training Process:** Veliformer employs a two-stage training approach:

1. *Self-supervised Pre-training:* In this stage, the masked reconstruction mechanism described in Section 2.3 is employed. The model learns to reconstruct original tidal velocity sequences from multiple masked versions, thereby capturing the inherent periodic patterns and temporal dependencies without requiring future ground truth labels [24].
2. *Supervised Fine-tuning:* Following pre-training, the model parameters are fine-tuned using the standard supervised forecasting objective, where the model learns to map historical sequences to future sequences using the mean squared error loss function.

This two-stage approach enables Veliformer to leverage both the unsupervised periodic pattern learning from the masking mechanism and the supervised sequence-to-sequence mapping required for accurate forecasting.[25]

### 3.2. Dataset description

The tidal data used in this study come from the ReDAPT project, which collects data at a tidal energy test site near the Fall of Warness, one of the Orkney Islands in Scotland [26]. The Fall of Warness is a globally significant site for tidal energy research and development, known for its strong, consistent, and fast-flowing tidal currents, making it an ideal location for assessing the performance of tidal energy converters and related technologies. The availability of high-quality, high-resolution data from such a well-characterized energetic site was a key reason for its selection, providing a robust foundation for developing and validating our forecasting model.

Tidal current velocity data were collected using a four-beam acoustic Doppler current profiler (ADCP) deployed on a gravity-anchored frame on the seabed [27]. Tidal current velocity data along the X, Y, and Z directions at depths of 22, 23, and 24 m below sea level were selected for multilevel prediction. These measurements, taken at three different depths, were chosen to capture variations in tidal currents at multiple layers of the water column, providing a more comprehensive dataset for predictive modeling [28]. The multi-depth approach is crucial as tidal currents exhibit vertical shear, meaning their speed and direction can vary significantly from the sea surface to the seabed due to factors like bed friction and velocity gradients. Analyzing data from different depths thus allows for a more accurate representation of the overall energy potential and the complex three-dimensional flow structure.

While the horizontal velocity components (X1, Y1) are the primary contributors to tidal power generation, the inclusion of the vertical velocity component (Z1) serves as a crucial auxiliary information source in our multivariate forecasting framework. Although Z1 does not directly contribute to power calculations, it captures important three-dimensional flow dynamics, turbulence intensity, and vorticity patterns that significantly influence the stability and short-term variations of the horizontal flow components. In multivariate time series prediction, deep learning models can automatically learn complex correlations between Z1 variations and future changes in X1 and Y1, thereby leveraging this additional hydrodynamic information to improve prediction accuracy for the power-generating components. The vertical flow patterns often serve as early indicators of flow regime changes and environmental perturbations that subsequently affect horizontal currents, making Z1 a valuable predictor variable despite its indirect relationship to power output [29].

The data were recorded continuously over a seven-day period with a high-resolution sampling interval of 1 s, allowing for detailed temporal analysis. The 1-second sampling interval is particularly important for capturing the fine-grained dynamics and turbulent fluctuations inherent in tidal flows, which might be missed by coarser sampling rates.

The raw data were stored in .mat file format, and the dataset included missing values and outliers due to the possible malfunction of the recording instruments or interference caused by fish movement in the vicinity of the sensors [30]. We thoroughly pre-processed the data to address these issues. For non-continuous missing values, the average of neighboring time points was used to fill the gaps. This method was chosen as it provides a reasonable local estimate while preserving the underlying temporal structure without introducing significant bias. Continuous missing values were directly replaced with the default value of 0.01 to ensure consistency. This small constant value was used to maintain data integrity for numerical processing and to clearly distinguish these imputed points from actual zero readings, while minimizing their impact on overall statistical properties.

To understand the underlying statistical properties of the tidal velocity components, we first examined their probability density distributions. Fig. 4 illustrates the distributions for the key components X1, Y1, and Z1, derived using histograms and Kernel Density Estimation. The distributions for X1 and Y1 exhibit distinct multi-modal characteristics, suggesting the presence of several dominant operational states within the tidal flow, possibly corresponding to different phases and strengths of the ebb and flood tides. In contrast, the Z1 component shows a sharp, unimodal distribution highly concentrated around zero, indicating that vertical velocities are predominantly minimal but can experience occasional fluctuations. These varied and complex distributions underscore the non-Gaussian nature of the tidal data and highlight the necessity for sophisticated modeling approaches capable of capturing such diverse data patterns.

To further validate and characterize the inherent periodicity crucial for tidal energy forecasting, we performed frequency and time-domain analyses. A frequency domain analysis using the Fast Fourier Transform (FFT) was conducted on the velocity components. As illustrated in Fig. 5, both the X1 and Y1 components exhibit a distinct dominant frequency corresponding to a period of approximately 12 h, consistent with the known semidiurnal tidal cycles driven by lunar gravitation. Although the Z1 component exhibits more scattered spectral energy, a weak periodicity is still observable. These findings from the spectral analysis confirm that the dataset contains clear periodic patterns.

Complementing the frequency-domain insights, an analysis of the temporal dependence structure was conducted using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots, shown in Fig. 6 for a representative key tidal velocity component (X1). To clarify this analysis, the ACF measures the correlation between the time series and its own lagged values, revealing the overall strength of persistence and long-term memory in the data. The PACF, conversely, measures the direct correlation between an observation and a specific lag after removing the influence of the intermediate time steps, which is useful for identifying the order of autoregressive processes [31]. These tools are fundamental for diagnosing the underlying structure of time series data. The ACF plot for X1 demonstrates a very slow decay, indicating strong persistence and significant autocorrelation across many lags. Each lag represents a one-second time step, consistent with the data's sampling frequency. This pattern is characteristic of time series with strong underlying periodicities or trends. The PACF, on the other hand, cuts off sharply after a few lags, suggesting an autoregressive nature in the data once the influence of intermediate observations is removed. Together, these ACF and PACF characteristics strongly reinforce the presence of exploitable temporal structures and periodicities within the tidal velocity data, justifying the design choice in Veliformer to explicitly preserve such periodic features for improved forecasting accuracy.

To facilitate the subsequent computation of OPF, we converted the tidal flow rate into power using the Flux method [32]. The theoretical power ($P_{tidal}$) that can be extracted by a tidal stream turbine is generally calculated as:

$$P_{tidal} = \frac{1}{2}\rho A_{swept} u_{tidal}^3 C_P \eta_{overall} \tag{27}$$

where $\rho$ is the density of seawater, $A_{swept}$ is the cross-sectional area swept by the turbine rotors ($m^2$), $u_{tidal}$ is the velocity of the tidal current (m/s), $C_P$ is the power coefficient, and $\eta_{overall}$ is the overall conversion efficiency of the power train. The tidal current velocity data $u_{tidal}$, collected as described previously, serves as a primary input for this power calculation. For the purpose of OPF calculations in this work, appropriate and consistent values for $A_{swept}$, $C_P$, and $\eta_{overall}$ were utilized.

The dataset is categorized into a training set, a test set and a validation set with proportions of 70%, 20% and 10%, respectively. The validation set is used to tune the model's hyperparameters. The predictive accuracy of the forecasting models developed and evaluated using this dataset will be primarily assessed through Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics are formally defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{28}$$

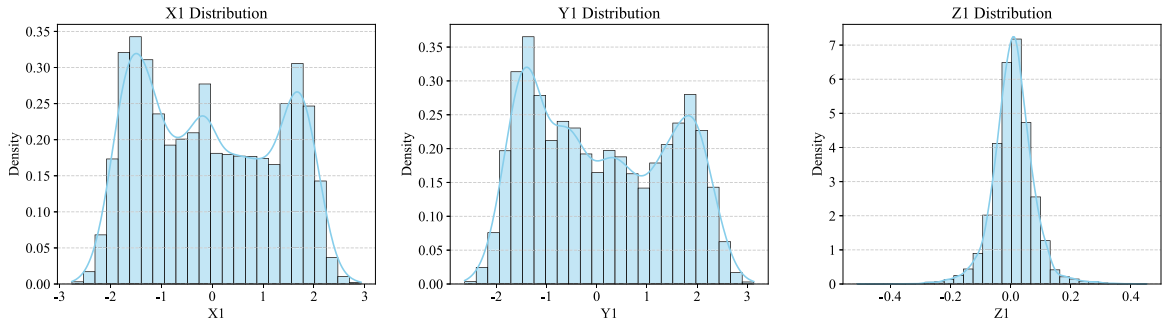


**Fig. 4.** Distribution analysis of key tidal velocity components (X1, Y1, Z1) via histograms and kernel density estimation, revealing multi-modal distributions for X1 and Y1, and a sharp unimodal distribution for Z1.

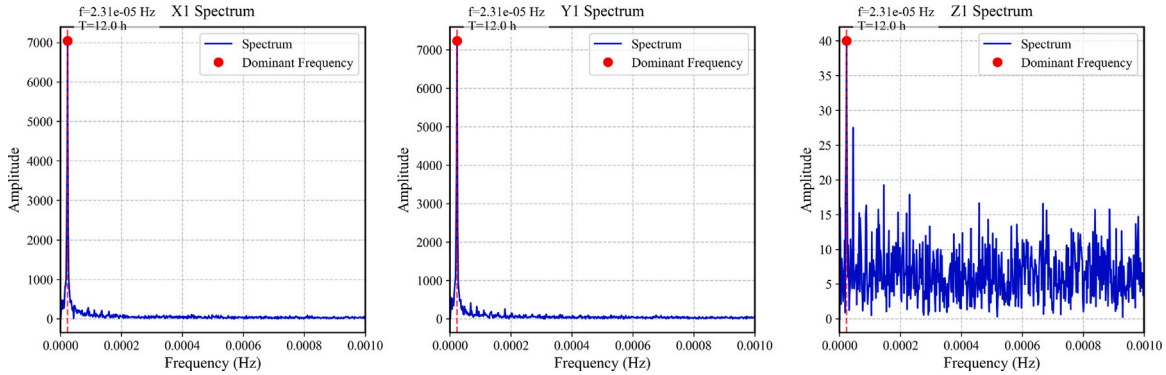


**Fig. 5.** Dominant 12-hour periodicity revealed by spectral analysis of tidal velocities in X, Y, and Z directions.

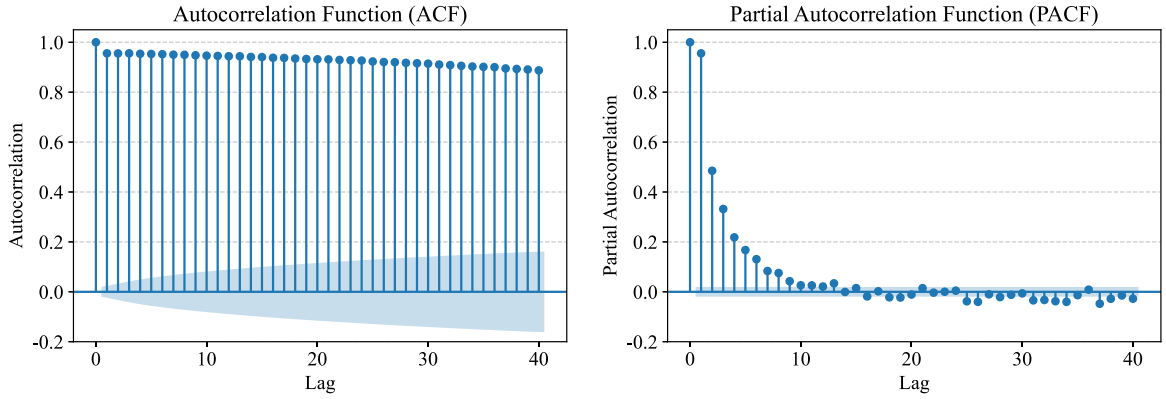


**Fig. 6.** ACF and PACF analysis of a key tidal velocity component (X1). The slow decay in ACF highlights strong persistence/periodicity, while the PACF suggests an underlying autoregressive structure. (Note: The figure displays data for component X1, chosen for its representative characteristics of the primary flow dynamics. The lag units are in time steps, where each time step corresponds to 1 s given our data sampling interval.)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{29}$$

where $n$ represents the total number of data points in the evaluation set, $y_i$ is the actual observed value for the $i$th data point, and $\hat{y}_i$ is the corresponding value predicted by the forecasting model. These metrics will be instrumental in comparing the performance of different models discussed in subsequent sections.

### 3.3. Optimal power flow cost comparison case study

The IEEE 118-bus power system is a complex and widely used standardized test system for power system research. It contains 118 buses, 186 lines, and 54 transformers, which together represent a complex urban power grid network [33]. The system simulates a power network in a region of the United States and is widely used to study power system dynamic behavior. Similarly, the IEEE 39-bus system is a well-known test case that represents the power grid in the northeastern region of the United States. It consists of 39 buses, 46 lines, and 10 generators, making it suitable for studying dynamic stability and power flow analysis.

In our case study, we set the reference power for both systems to 100 MVA, with reference voltages set to 380 kV and 110 kV, respectively. The nominal frequency for both systems is maintained at 60 Hz. To ensure the accuracy of the study, both systems utilize high-precision power flow calculation and optimization algorithms during the optimized tidal energy generation operation. The tidal energy generator is connected to the 110 kV voltage level in both the IEEE 118-bus and 39-bus systems. Specifically, the generator is connected to bus 25 in the IEEE 39-bus system and bus 61 in the IEEE 118-bus

**Table 2**
Cost in IEEE 118 bus and IEEE 39 bus systems.

| | | TCN [34] | | TCN-LSTM [35] | | LSTM-GRU [36] | | GRU-FCN [37] | | Veliformer | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | IEEE-39 | IEEE-118 | IEEE-39 | IEEE-118 | IEEE-39 | IEEE-118 | IEEE-39 | IEEE-118 | IEEE-39 | IEEE-118 |
| Cost | 15min | 29207 | 18737 | 31445 | 45 668 | 39 141 | 18782 | 37 089 | 19 466 | 28981 | 16767 |
| | 6h | 768806 | 839 535 | 796 994 | 751166 | 772078 | 791885 | 823 226 | 1 066 771 | 745341 | 678485 |
| | 20h | 2 772 456 | 1546014 | 2 811 036 | 1312254 | 2736331 | 1 795 600 | 2727652 | 1 860 472 | 2644167 | 1266977 |

1. All costs are in U.S. dollars (USD).
2. Gold represents the best performance, Silver represents the second best, Copper represents the third best.

**Table 3**
Model performance comparison across multiple time intervals (10 s, 5 min, 10 min, 20 min, 30 min, 60 min) for multivariate time-series forecasting (see Ref. [38]).

| Models | 10 s | | 5 min | | 10 min | | 20 min | | 30 min | | 60 min | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| Crossformer [39] | 0.256 | 0.391 | 0.298 | 0.386 | 0.302 | 0.427 | 0.348 | 0.461 | 0.276 | 0.409 | 0.303 | 0.429 |
| TCN [34] | 0.466 | 0.475 | 0.313 | 0.390 | 0.388 | 0.433 | 0.287 | 0.350 | 0.277 | 0.350 | 0.335 | 0.389 |
| LSTM-GRU [36] | 1.310 | 0.899 | 0.467 | 0.404 | 0.451 | 0.401 | 0.462 | 0.417 | 0.467 | 0.419 | 0.458 | 0.420 |
| TCN-LSTM [35] | 0.468 | 0.533 | 0.346 | 0.412 | 0.397 | 0.439 | 0.281 | 0.352 | 0.306 | 0.352 | 0.334 | 0.369 |
| GRU-FCN [37] | 1.843 | 0.985 | 0.373 | 0.364 | 0.457 | 0.402 | 0.461 | 0.414 | 0.475 | 0.424 | 0.502 | 0.439 |
| Informer [40] | 0.545 | 0.458 | 0.434 | 0.542 | 0.403 | 0.507 | 0.543 | 0.573 | 0.605 | 0.599 | 0.732 | 0.673 |
| DLinear [41] | 0.291 | 0.352 | 0.382 | 0.361 | 0.386 | 0.369 | 0.391 | 0.373 | 0.401 | 0.384 | 0.406 | 0.393 |
| FNet [42] | 0.297 | 0.385 | 0.384 | 0.354 | 0.381 | 0.360 | 0.378 | 0.367 | 0.385 | 0.377 | 0.413 | 0.415 |
| UTide [38] | 0.752 | 0.613 | 0.518 | 0.488 | 0.425 | 0.415 | 0.389 | 0.390 | 0.357 | 0.365 | 0.325 | 0.461 |
| Veliformer | 0.236 | 0.336 | 0.287 | 0.347 | 0.284 | 0.344 | 0.271 | 0.348 | 0.263 | 0.345 | 0.289 | 0.355 |

1. Gold represents the best performance, Silver represents the second best, Copper represents the third best.

system. The impact of the tidal energy generator on both power systems is analyzed through precise power flow optimization.

The effectiveness of the proposed model was validated by integrating tidal energy generation power into both the IEEE 118-bus and IEEE 39-bus systems, which allowed us to calculate the system generation costs under different time scenarios. This case study aims to assess the performance of various models in optimizing power flow costs within these systems, including the proposed Veliformer model, designed to enhance accuracy in such tasks.

The models compared in this study include TCN [34], TCN-LSTM [35], LSTM-GRU [36], GRU-FCN [37], and the proposed Veliformer model. We calculated generation costs across three scenarios: 15 min, 6 h, and 20 h, which denoted the total forecast horizons. The prediction time steps were set to 1 min for the 15-minute scenario and 10 min for the 6-hour and 20-hour scenarios, as shown in Table 2. These time intervals were selected to simulate both short-term and long-term operational scenarios within the IEEE 118-bus system, allowing for a comprehensive evaluation of each model's performance.

Table 2 shows that the proposed Veliformer model achieved the lowest generation costs across all three-time intervals. In different time scenarios, the Veliformer model demonstrated significant cost advantages across both the IEEE 118-bus and 39-bus systems. In the 15-minute scenario, Veliformer reduced costs by approximately 10.5% to 11.1% compared to the TCN and LSTM-GRU models in the IEEE 118-bus system, and by 0.8% to 7.7% compared to the TCN and TCN-LSTM models in the IEEE 39-bus system. In the 6-hour scenario, Veliformer achieved cost reductions of 9.6% and 19.2% in the IEEE 118-bus system (compared to TCN-LSTM and TCN models, respectively), while reducing costs by 3.0% and 6.5% in the IEEE 39-bus system (compared to the TCN and LSTM-GRU models). In the 20-hour scenario, Veliformer achieved savings of 3.4% to 16.1% in the IEEE 118-bus system (compared to TCN-LSTM and TCN models), and 4.6% in the IEEE 39-bus system (compared to both TCN-LSTM and TCN models).

These cost reductions highlight the efficiency of Veliformer in minimizing power flow costs, primarily attributed to its masking mechanism, which enhances the model's ability to focus on the most relevant data. The consistent performance of the Veliformer model across different time frames underscores its robustness and potential for application in integrated energy systems, where efficient power flow optimization is crucial.

### 3.4. Model prediction accuracy comparison experiment

We evaluate Veliformer against nine baseline models—Crossformer, TCN, LSTM-GRU, TCN-LSTM, GRU-FCN, Informer, DLinear, FNet, and UTide—across six different forecast horizons: *10 seconds*, *5 minutes*, *10 minutes*, *20 minutes*, *30 minutes*, and *60 minutes*. Table 3 summarizes the results in terms of Mean Squared Error (MSE) and Mean Absolute Error (MAE). We acknowledge that pointwise metrics such as MSE and MAE can overweight phase misalignment in periodic signals. However, our comparisons emphasize relative model performance and downstream OPF cost benefits, so our conclusions do not rely on a single pointwise metric. Therefore, we complement MSE/MAE with the Q value (a Sobolev-norm-based metric for surface similarity) and $R^2$ to assess both phase and amplitude consistency in a more robust manner. Table 4 summarizes the key advantages and limitations of these baselines, which provides insights into their design characteristics and forecasting capabilities.

**Short-term horizons (10 s, 5 min, 10 min).** For short-term forecasts, Veliformer consistently delivered superior accuracy over competing models. At the 10 s interval, it reduced MSE by roughly 8% and MAE by about 14% compared to the next-best model, Crossformer. Moving to 5 min predictions, Veliformer's MSE and MAE showed improvements of around 4% and 10%, respectively, relative to Crossformer, and outperformed TCN, TCN-LSTM, DLinear, and FNet by even wider margins. By 10 min, Veliformer retained a consistent edge, with its MSE and MAE about 6% and 19% lower, respectively, than Crossformer, while the gap against other baselines grew larger. Across the short-term horizons, Veliformer achieved an average improvement of approximately 7.64% in MSE and 9.84% in MAE.

**Longer horizons (20 min, 30 min, 60 min).** As the forecast window extended, Veliformer's relative lead remained pronounced. For instance, at 20 min, it demonstrated an approximate 4% drop in MSE and a 1% drop in MAE over the second-best TCN-LSTM. At the 30 min horizon, Veliformer improved MSE by around 5% relative to Crossformer. Even at the longest 60 min forecast, Veliformer still surpassed the runner-up by around 5% in MSE and 17% in MAE. Overall, these results confirm that Veliformer delivers robust gains across all time intervals, providing anywhere from a few percentage points to double-digit percentage error reductions compared to baseline models. Across

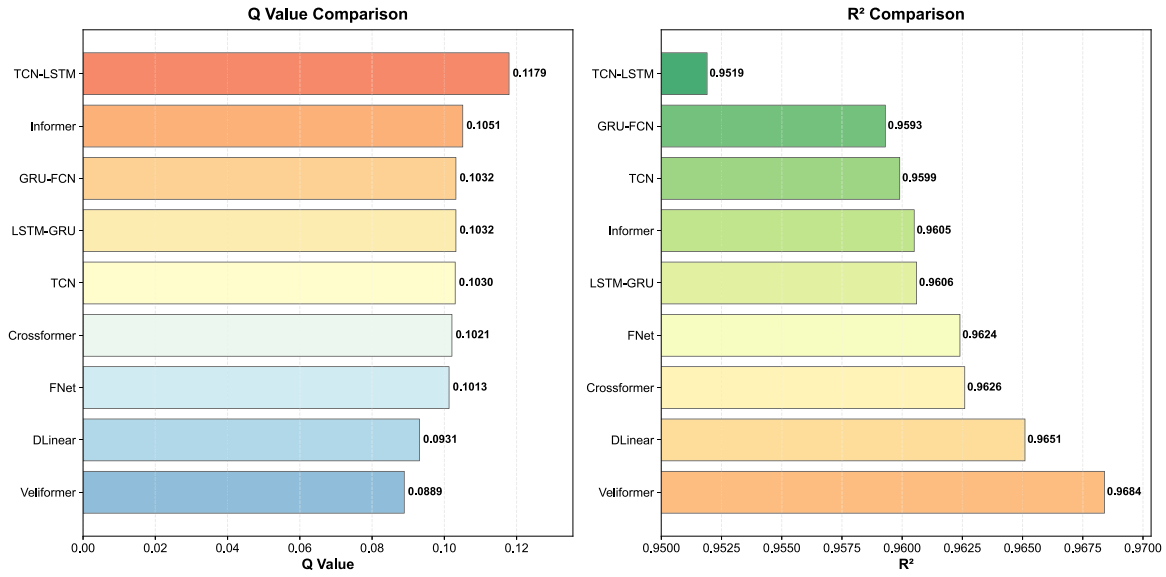


**Fig. 7.** Performance comparison of Veliformer against eight baseline models on Q value and $R^2$ metrics. The results are based on the 20-minute forecasting scenario. Both metrics indicate that Veliformer achieves the best performance.

**Table 4**
Summary of baseline models.

| Model | Advantages | Limitations | Ref. |
|---|---|---|---|
| Crossformer | Captures long-range dependencies using cross-dimension attention for inter-channel interactions and frequency mixing for spectral pattern recognition. | Coarse temporal granularity, optimized for long-term focus, can lead to sensitivity to short-term fluctuations and overlooking of localized details. | [39] |
| TCN | Employs causal and dilated convolutions ensuring valid temporal flow, enabling fast training, stable gradients, and large receptive fields for contextual understanding. | Fixed convolutional structure offers limited dynamic temporal adaptivity, potentially struggling with highly irregular or non-stationary time series patterns. | [34] |
| LSTM-GRU | Leverages gated recurrence mechanisms from LSTM and GRU units to effectively model complex sequential dynamics and manage information flow over long sequences. | May still encounter vanishing gradient challenges in very long sequences and can incur high inference latency due to its inherently sequential computation. | [36] |
| TCN-LSTM | Combines TCN's ability to capture broad temporal contexts via long receptive fields with LSTM's proficiency in memory retention for robust sequential data modeling. | The hybrid architecture can be computationally more intensive than its standalone components and typically lacks inherent mechanisms for frequency-aware processing. | [35] |
| GRU-FCN | Provides a lightweight and efficient architecture by integrating GRUs for temporal modeling with FCNs for convolutional feature encoding, suitable for faster processing. | May offer limited interpretability of learned features and possesses weaker inherent capabilities for explicit frequency domain analysis or decomposition. | [37] |
| Informer | Utilizes a ProbSparse attention mechanism to efficiently process very long sequences, significantly reducing the computational overhead associated with standard attention mechanisms. | Its strong focus on dominant long-range patterns via sparse attention may result in overlooking finer-grained, localized temporal details crucial for some predictions. | [40] |
| DLinear | Offers a simple yet robust baseline by decomposing time series into distinct trend and seasonal components, which are then modeled linearly for interpretability. | The inherent linearity restricts its ability to capture non-linear patterns and complex interactions, making it less adaptive to abrupt changes or localized disruptions. | [41] |
| FNet | Replaces computationally intensive self-attention with unparameterized Fourier Transforms for global token mixing, significantly accelerating inference and reducing model complexity. | Reliance on Fourier analysis, which assumes periodicity and stationarity, limits its adaptivity to non-stationary or non-periodic data and complex aperiodic events. | [42] |

the longer horizons, Veliformer achieved an average improvement of approximately 4.30% in MSE and 2.67% in MAE.

To further assess performance, Fig. 7 presents a comparison based on Q value and the coefficient of determination ($R^2$) for the 20-minute forecasting scenario. The Q value, a metric that quantifies the magnitude of error between predicted and observed values, shows Veliformer achieving the top score of 0.0889 [43]. Similarly, for the $R^2$, which represents the proportion of variance in the observed data that is predictable from the model, Veliformer again leads with a score of 0.9684. These results provide additional evidence of Veliformer's superior predictive accuracy. Fig. 8 illustrates the predicted time-series curves from Veliformer and the Crossformer baseline against the ground truth. The plots for both 20-minute and 60-minute horizons visually confirm Veliformer's ability to more closely track the actual tidal velocity, accurately capturing the critical peaks and troughs that are essential for reliable operational planning.

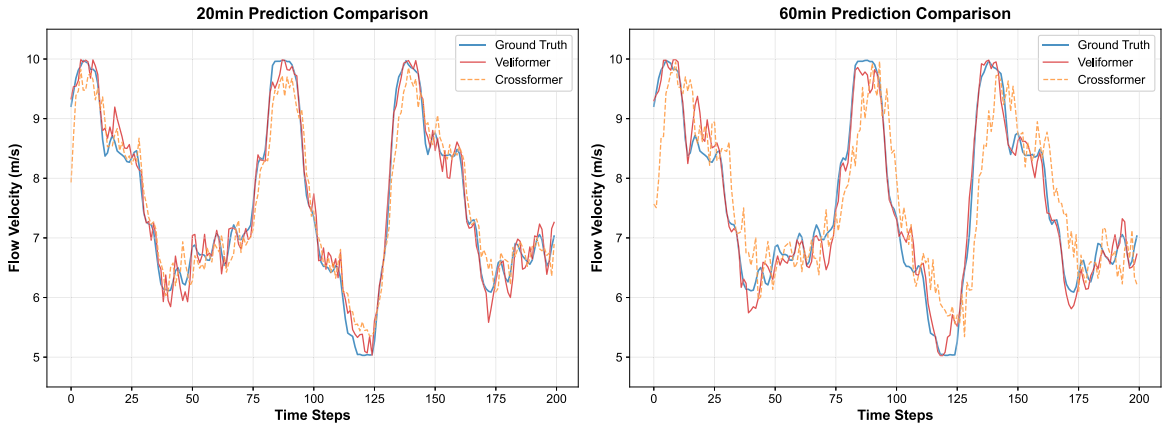


**Fig. 8.** Case visualization of prediction performance for 20-minute and 60-minute forecast horizons. The plot compares the ground truth with predictions from Veliformer and the baseline Crossformer.

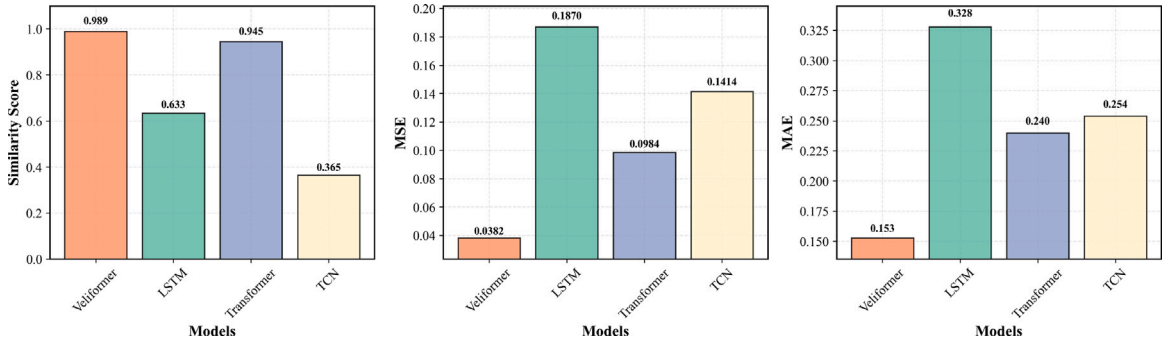


**Fig. 9.** Comparative analysis of Veliformer and baseline models on high-frequency forecasting metrics for a 20-minute prediction horizon. (Left) Spectral Similarity, where a higher score is better. (Center) High-Frequency MSE and (Right) High-Frequency MAE, where lower values indicate better performance. The results highlight Veliformer's superior capability in accurately predicting high-frequency components.

Overall, Veliformer provides consistently lower MSE and MAE values than the other methods across all six time intervals, with an average improvement of 4.91% in prediction accuracy, supplemented by its leading performance on Q and $R^2$ metrics, which indicates its robustness for both short-term and long-term multivariate time-series forecasting tasks.

### 3.5. Validation of high-frequency component forecasting

To address concerns regarding the importance of high-frequency components in tidal energy forecasting – where the cubic relationship between velocity and power ($P \propto u^3$) can amplify the impact of small, rapid fluctuations – a dedicated experiment was conducted. This experiment was designed to quantitatively assess Veliformer's ability to accurately predict these crucial high-frequency dynamics compared to baseline models (LSTM, Transformer, and TCN). The evaluation was performed on a 20-minute forecasting task, focusing on three specialized frequency-domain metrics: Spectral Similarity, High-Frequency MSE, and High-Frequency MAE.

Spectral Similarity measures the cosine similarity between the frequency spectra of the predicted and true signals, indicating how well the overall periodic structure is preserved. High-Frequency MSE and MAE are calculated after applying a high-pass filter to isolate components with periods shorter than 10 min, directly quantifying the model's accuracy on the most rapid variations.

The results, presented in Fig. 9, unequivocally demonstrate Veliformer's superior performance in capturing high-frequency components. Veliformer achieved a Spectral Similarity score of 0.989, significantly outperforming the next-best model, Transformer (0.945), and indicating a much higher fidelity in reconstructing the complete frequency spectrum. Most critically, in the direct evaluation of high-frequency errors, Veliformer obtained an MSE of 0.0382 and an MAE of 0.153. These error values are substantially lower – by a factor of approximately 2.6 for MSE and 1.6 for MAE compared to the Transformer – than those of all baseline models. This marked reduction in high-frequency error confirms that Veliformer's masked reconstruction mechanism is highly effective at preserving the fine-grained temporal details essential for accurate tidal power estimation, directly validating its advantage in handling the very components that are most critical to energy conversion calculations.

### 3.6. Sensitivity analysis of hyperparameters

To comprehensively evaluate the robustness of Veliformer and identify optimal or influential hyperparameter settings, an extensive sensitivity analysis was conducted. In these experiments, one hyperparameter was varied at a time, while all other parameters were maintained at their established baseline values. For each specific configuration, the Veliformer model underwent both its self-supervised pre-training and supervised fine-tuning stages before Veliformer's predictive performance, in terms of Test MSE and Test MAE, was evaluated on the designated test set. The summarized results of these experiments, illustrating performance trends for each tested hyperparameter, are presented in Fig. 10.

Parameters central to Veliformer's masking strategy were examined. For the Temporal Unit (M), defining the number of augmented masked sequences utilized, model performance, measured by both MSE and MAE, generally exhibited consistent improvement with an increasing number of units within the tested range of 1 to 7. This upward trend in performance suggests that incorporating a richer set of augmented

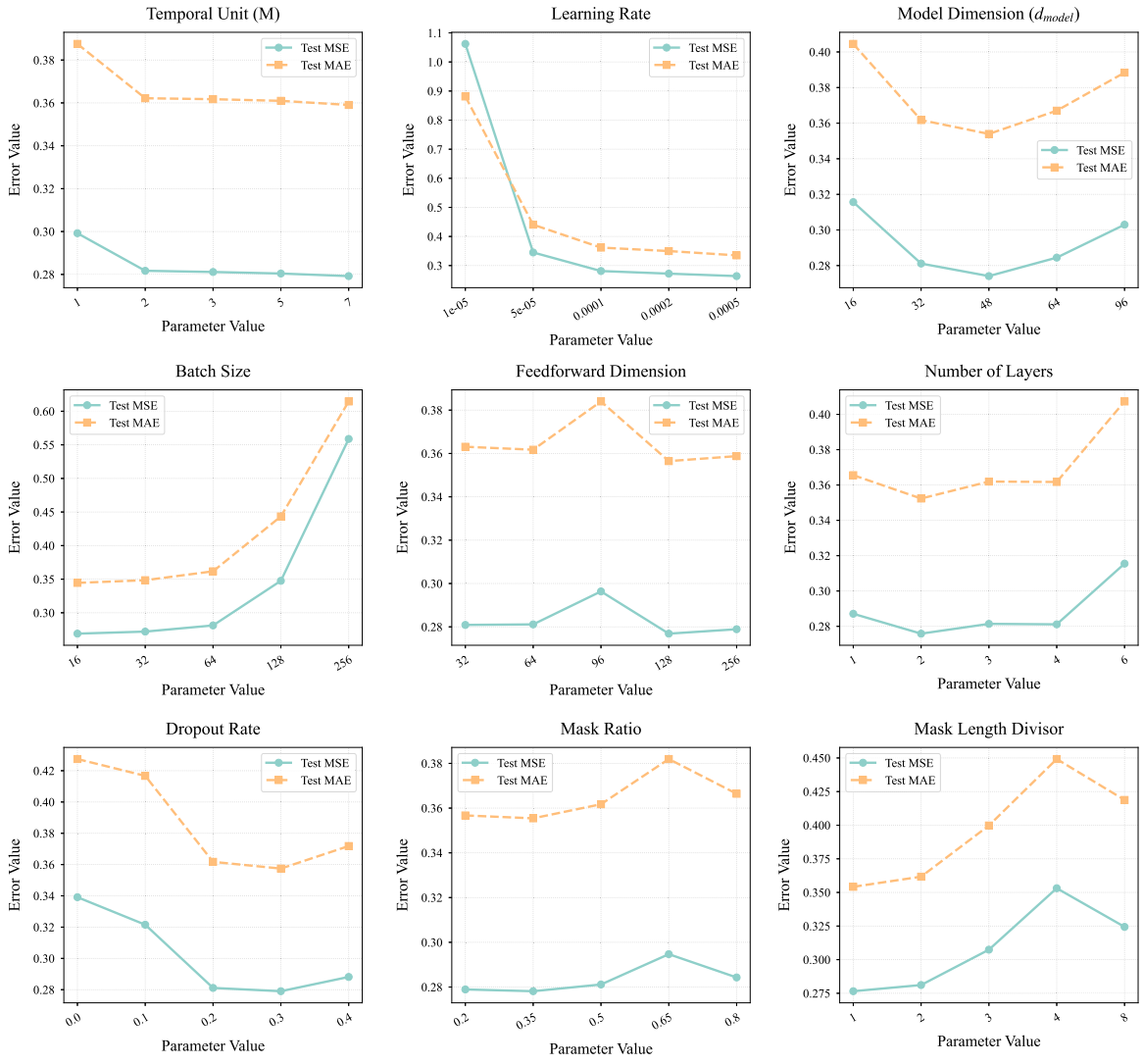


**Fig. 10.** Sensitivity of Veliformer's prediction performance (Test MSE and Test MAE) to variations in individual hyperparameters. Each subplot illustrates the change in error metrics as a single hyperparameter is varied, while all other parameters are held constant at their baseline values.

perspectives on the input series allows Veliformer to build more robust and comprehensive temporal feature representations. Such representations are crucial for accurate tidal energy forecasting where diverse and subtle short-term patterns exist. The mask ratio, representing the proportion of the input sequence that is masked, was evaluated from 0.2 to 0.8. The analysis showed that lower ratios within this evaluated range, specifically around 0.2, tended to yield significantly better results than higher ratios. As the mask ratio increased beyond 0.2, a clear upward trend in both MSE and MAE was observed, indicating performance degradation. This finding suggests that while a certain degree of masking is essential for the model's learning mechanism, an overly aggressive masking approach can be detrimental, likely by removing too much critical information about tidal patterns for the model to effectively learn and reconstruct. Regarding the mask length divisor, which inversely controls the length of contiguous masked segments, the analysis revealed a clear trend. Model performance, in terms of both MSE and MAE, consistently improved as the divisor increased when tested from 1 to 7. An increasing divisor corresponds to shorter contiguous masked segments. This outcome indicates that masking shorter, more distributed segments throughout the time series is more advantageous for Veliformer's performance than masking fewer, longer contiguous blocks. The latter approach might excessively disrupt local temporal dependencies or obscure entire short-term periodic events, whereas shorter, distributed masked segments may encourage the model to learn finer-grained contextual relationships and improve the model's ability to capture nuances of the tidal data. However, this is a preliminary observation under conditions where other parameters are held constant. The potential interaction effects between parameters warrant a more detailed investigation, which will be analyzed in the next section.

Key architectural parameters of the Transformer were analyzed. The model dimension ($d_{model}$) exhibited a distinct U-shaped sensitivity curve, indicating an optimal capacity. As $d_{model}$ increased from smaller values such as 16, predictive performance, reflected by decreasing MSE and MAE, improved significantly. An optimal capacity was observed around a dimension of 48 to 64. Beyond this optimal region, further increases in $d_{model}$ led to a slight degradation in performance, possibly due to the onset of overfitting with excessive parameters for the given dataset size or an increased difficulty in optimizing a larger, more complex network. A similar U-shaped trend was observed for the number of Transformer layers (num_layers). Both a low number of layers, such as a single layer, which may lack the hierarchical capacity to model complex temporal dependencies, and a high number of layers, such as 5 or 6 layers, which can be harder to train effectively and become prone to overfitting or issues like vanishing gradients, resulted in higher prediction errors. An intermediate depth, typically

around 2 to 3 layers, was found most effective, striking an optimal balance between model expressiveness and the model's ability to generalize from training data. The dimension of the feedforward network ('dim_feedforward') within the Transformer layers also demonstrated an optimal range. Performance generally improved (errors decreased) as this feedforward dimension increased from smaller values like 32 or 64. The lowest error metrics were typically observed for moderately larger dimensions, such as 128 or 256. Further increasing this feedforward dimension beyond this range did not yield substantial additional performance gains and, in some instances, led to a slight increase in error, indicating a point of diminishing returns regarding model capacity for this specific network component.

Common training and regularization hyperparameters were assessed. The learning rate (lr) demonstrated a critical impact on model training and final predictive accuracy, exhibiting a pronounced U-shaped curve. The analysis revealed a distinct optimal range, typically around $1 \times 10^{-4}$ to $5 \times 10^{-5}$, where the model achieved the lowest MSE and MAE. Learning rates of $1 \times 10^{-5}$ or less significantly hindered convergence speed and resulted in suboptimal performance, likely due to the optimizer struggling to escape shallow local minima. Conversely, learning rates of $2 \times 10^{-4}$, $5 \times 10^{-4}$, and above led to training instability and divergence, causing a sharp increase in prediction errors. The choice of batch size also proved influential, with smaller batch sizes generally yielding superior results. The experiments, testing batch sizes from 16 to 256, indicated that smaller values like 16 or 32 resulted in lower MSE and MAE compared to larger batch sizes such as 128 or 256. Larger batches tended to increase prediction errors, a phenomenon sometimes attributed to larger batches converging to sharper minima in the loss landscape, which may generalize less effectively than the flatter minima often found by smaller batches. The dropout rate analysis clearly confirmed the benefits of regularization for the Veliformer model. Performance was notably worse, with higher MSE and MAE, when dropout was not applied (a dropout rate of 0.0), indicating a tendency of the model to overfit the training data. An intermediate dropout rate, typically found most effective in the range of 0.1 to 0.2, minimized both error metrics. Dropout rates higher than this optimal range, such as 0.3 or 0.4, began to degrade performance again, likely due to excessive information loss during training, leading to underfitting.

In summary, the detailed analyses underscore that Veliformer's predictive accuracy is highly sensitive to the interplay of architectural design, masking strategy configuration, and training procedure. The identified trends and more precisely characterized optimal regions for these hyperparameters offer valuable and actionable insights. This enhanced understanding facilitates the effective deployment of Veliformer in tidal energy forecasting, ensuring robust performance and maximizing Veliformer's predictive capabilities when tackling complex time series data.

### 3.7. Effect of different masking strategies

Building on the preliminary analysis in Section 3.6, which suggested that shorter mask segments were optimal under a fixed set of hyperparameters, we recognize that the effect of a single hyperparameter may not fully capture its role in a complex model. Specifically, the optimal mask segment length might be significantly influenced by the chosen mask ratio. To investigate this interaction effect, we designed a more comprehensive experiment by systematically varying the combination of mask ratio and mask segment length. The mask segment length sets the average size of contiguous segments within the geometric masking process. These experiments utilized model hyperparameters established from prior sensitivity analyses. The Veliformer model incorporated both self-supervised pre-training and supervised fine-tuning stages. Test MSE and MAE served as the performance evaluation metrics. Fig. 11 summarizes the main effects of these masking parameters. Fig. 12 details their interaction effects through heatmaps, while Table 5 lists the top-performing parameter combinations.

An examination of the main effects, referencing Fig. 11, reveals the average trends for each masking parameter when the influence of the other is averaged out. For the masking ratio, a clear trend emerges: lower to moderate values, particularly around 0.2, generally yield superior average performance, with a noticeable degradation as the proportion of masking increases towards 0.5. Regarding the length of masked segments, the analysis presents a more nuanced picture that contrasts with the findings from the single-variable sensitivity analysis in Section 3.6. While the earlier analysis pointed to shorter segments being optimal for a specific fixed configuration, the main effects analysis here indicates that longer segments (specifically 90 and 180) are more beneficial on average when evaluated across all mask ratios. This apparent discrepancy strongly suggests the presence of a significant interaction effect, meaning the ideal segment length is highly dependent on the chosen mask ratio.

The heatmaps in Fig. 12 provide a more nuanced understanding by illustrating significant interaction effects between the masking ratio and mask segment length. These visualizations clearly show that the optimal setting for one parameter often depends on the value of the other, rather than a universally optimal value existing for each in isolation. For instance, a masking ratio of 0.4 paired with a relatively short mask segment length of 45 achieved one of the best overall performances. Conversely, a lower masking ratio, such as 0.2, also produced excellent results when combined with a very long mask segment length of 180. Furthermore, even a high masking ratio of 0.5 demonstrated competitive performance when matched with a moderate mask segment length of 60, while performing poorly with shorter mask segment lengths. Table 5 highlights these distinct optimal pairings by listing the top three configurations. This interplay suggests the model can adapt to different masking intensities if the segment length is appropriately chosen to balance context and reconstruction difficulty.

In conclusion, the investigation into different masking strategies underscores that simple monotonic rules do not govern the effectiveness of Veliformer's masking mechanism for the masking ratio and mask segment length individually. Instead, optimal predictive accuracy arises from specific combinations of these two parameters. The findings indicate a preference for moderate masking ratios, but the ideal segment length is interactive and context-dependent, with several distinct combinations yielding top-tier performance. This detailed understanding is crucial for configuring the masking strategy to maximize Veliformer's capabilities.

### 3.8. Ablation test

To assess the contribution of different components within the Veliformer model, an ablation study was conducted, as shown in Fig. 13. The base Veliformer model includes three key components: the masking mechanism, the Transformer architecture, and a fine-tuning process. To evaluate the impact of the masking mechanism and fine-tuning, two modified versions of the model were tested: one without the masking mechanism (denoted as "Veliformer w/o Masking") and another without the fine-tuning process ("Veliformer w/o Fine-tuning"). Additionally, the standard Transformer [44] model was included for comparison.

The performance of each model variant was evaluated using both MSE and MAE as key metrics. The base Veliformer model demonstrated the lowest error rates, achieving an MSE of 0.236 and an MAE of 0.336. When the masking mechanism was removed from the model, the MSE increased to 0.278 and the MAE rose to 0.386. Similarly, when the fine-tuning process was excluded, the error values further increased, with the MSE reaching 0.305 and the MAE rising to 0.392. The standard Transformer model exhibited the highest error rates among all tested configurations, with an MSE of 0.330 and an MAE of 0.421. The results indicate that both the masking mechanism and fine-tuning enhance model performance. The masking mechanism improves feature capture, while fine-tuning refines parameters for greater accuracy.

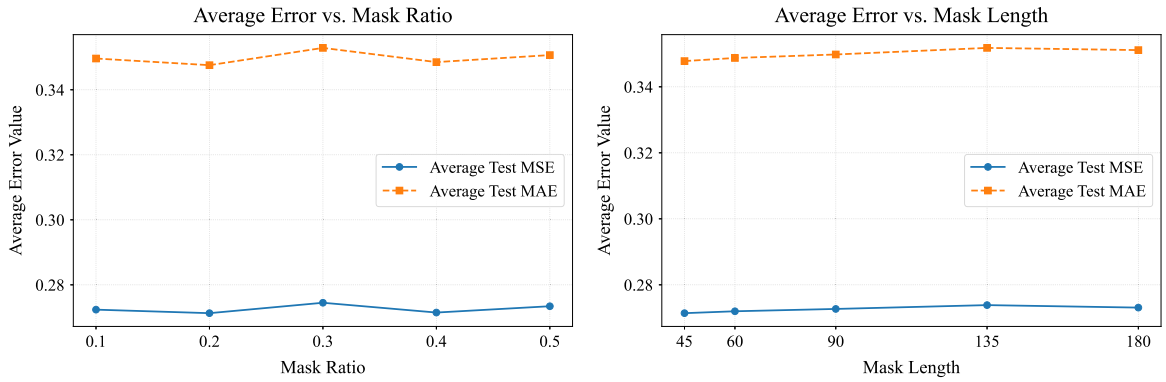


**Fig. 11.** Main effects of mask ratio and mask length on model performance.

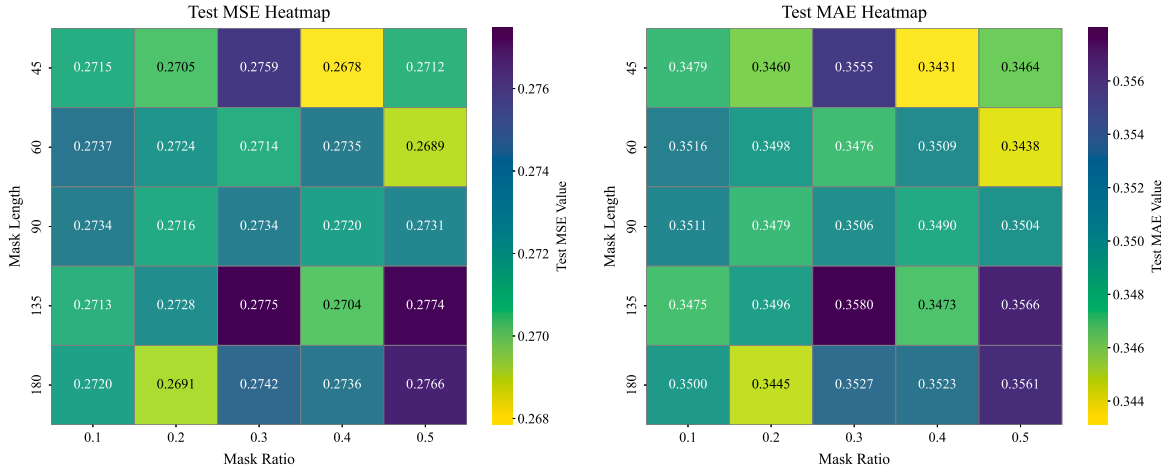


**Fig. 12.** Heatmaps showing the impact of combining different Mask Ratios and Mask Lengths on model prediction error. The left heatmap displays Test MSE, and the right heatmap displays Test MAE.

**Table 5**
Top 3 masking strategy combinations for optimal performance.

| Metric | Rank | Mask ratio | Mask length | Test MSE | Test MAE |
|---|---|---|---|---|---|
| Lowest MSE | 1 | 0.4 | 45 | 0.26784854 | 0.34311256 |
| | 2 | 0.5 | 60 | 0.26885887 | 0.34379373 |
| | 3 | 0.2 | 180 | 0.26907027 | 0.34449386 |
| Lowest MAE | 1 | 0.4 | 45 | 0.26784854 | 0.34311256 |
| | 2 | 0.5 | 60 | 0.26885887 | 0.34379373 |
| | 3 | 0.2 | 180 | 0.26907027 | 0.34449386 |

Note: The values for Test MSE and Test MAE are rounded to 8 decimal places for presentation. The ranking is based on the full precision values. The top configurations for MSE and MAE are identical in this dataset.

## 4. Conclusion

This paper addressed the prevalent challenge of periodicity disruption in deep learning-based short-term tidal energy forecasting by proposing Veliformer, a novel Transformer model. Veliformer utilizes a unique masking and reconstruction technique, rebuilding the original time series from multiple adjacent masked sequences. This approach is designed to effectively preserve the inherent periodic structure of tidal energy data, a capability supported by our theoretical analysis.

Comprehensive experimental evaluations demonstrated Veliformer's significant advantages. The model achieved an average prediction accuracy improvement of 4.91% over several state-of-the-art baseline models across multiple forecasting horizons. Furthermore, when applied to Optimal Power Flow (OPF) simulations on standard IEEE test systems, Veliformer delivered substantial reductions in power generation costs, highlighting its practical utility. The efficacy of the core masking mechanism was further validated through ablation studies, while extensive sensitivity and masking strategy analyses not only confirmed the model's robustness but also provided valuable guidelines for optimal hyperparameter configuration.

In summary, Veliformer offers a robust and efficient solution for enhancing short-term tidal energy forecasting and contributing to more effective power system optimization. By successfully preserving critical periodic information, the model facilitates more reliable grid integration and cost-effective management of tidal energy resources. The fundamental principles of its periodicity-preserving masking strategy may also hold promise for other time series forecasting domains where similar structural patterns are crucial. Future research will focus on exploring Veliformer's adaptability to other periodic renewable energy sources, assessing its performance and scalability in larger and more complex power systems, and pursuing further advancements in its adaptive masking techniques.

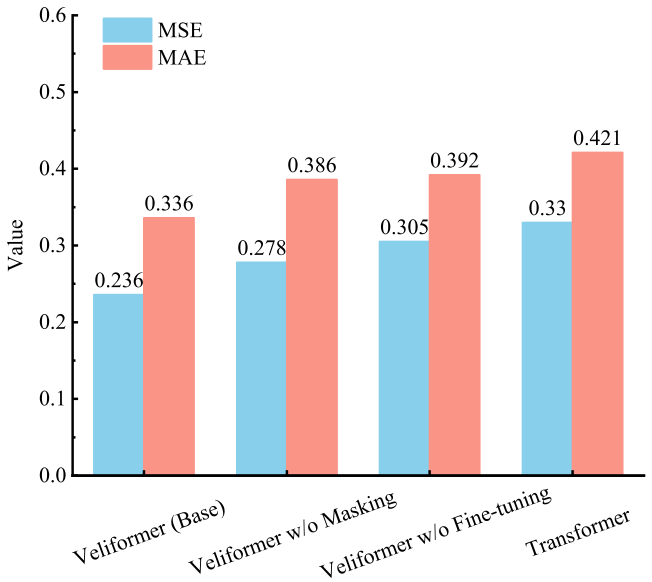


**Fig. 13.** Ablation Test on Veliformer. Veliformer (Base) is the original model, w/o means "without".

## CRediT authorship contribution statement

**Yangdi Huang:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Lina Yang:** Writing – review & editing, Supervision, Resources. **Xinzhang Wu:** Writing – review & editing, Resources, Funding acquisition. **Yunxuan Dong:** Writing – review & editing, Formal analysis, Data curation, Conceptualization.

## Acknowledgments

We would like to express our sincere gratitude for the financial support provided by the following organizations: the Department of Human Resources and Social Security of Guangxi Zhuang Autonomous Region (Grant No. 202401950), the Department of Science and Technology of Guangxi Zhuang Autonomous Region (Grant No. 2024JJB170087), the Guangxi Science and Technology Major Program (Grant No. AA23073019), and the National Natural Science Foundation of China (Grant No. 62371144). Their generous support has been instrumental in facilitating this research.

## Declaration of competing interest

All authors disclosed no relevant relationships.

The author(s) declared no potential con icts of interest with respect to the research, author- ship, and/or publication of this article.

## Appendix A. Detailed OPF solution method

The Optimal Power Flow (OPF) problem described in Eq. (1) is solved using the primal–dual interior-point method. By introducing slack variables and adding a barrier function to the objective function, the inequality constraints are transformed into equality constraints. Thus, the solution of problem (1) is transformed into the solution of the following problem:

$$\begin{aligned} \min \quad & f(\mathbf{x}^{\tau}(\mathbf{t})) - \xi\left(\sum_{p=1}^{r}\ln s_{1p} + \sum_{p=1}^{r}\ln s_{2p}\right) \\ \text{s.t.} \quad & h(\mathbf{x}) = 0, \\ & g(\mathbf{x}) - s_1 - \underline{g} = 0, \\ & g(\mathbf{x}) + s_2 - \bar{g} = 0, \end{aligned} \tag{A.1}$$

where $s_1$ and $s_2$ are slack variables. $s_{1p}$ and $s_{2p}$ are the $p$th elements in the slack variable vectors $s_1$ and $s_2$ ($s_{1p} > 0$, $s_{2p} > 0$). $r$ denotes the number of inequality constraints. $\xi$ is the barrier parameter. Problem (A.1) is an optimization problem containing only equation constraints, so it can be solved by the Lagrange multiplier method. Its Lagrangian function is as follows:

$$\begin{aligned} L = {} & f(\mathbf{x}^{\tau}(\mathbf{t})) - \lambda^T h(\mathbf{x}) - z_1^T\left[g(\mathbf{x}) - s_1 - \underline{g}\right] \\ & - z_2^T\left[g(\mathbf{x}) + s_2 - \bar{g}\right] - \xi\sum_{p=1}^{r}\ln s_{1p} - \xi\sum_{p=1}^{r}\ln s_{2p}, \end{aligned} \tag{A.2}$$

where $z_1 > 0$, $z_2 < 0$, $s_{1p} \geq 0$, $s_{2p} \geq 0$, $\lambda \neq 0$. $\lambda$, $z_1$ and $z_2$ represent the Lagrange multipliers associated with the constraints. $x$, $s_1$, and $s_2$ are the original variables.

According to the theory of Fiacco and McCormick [23], if the barrier parameter $\xi$ monotonically decreases to 0 during the iteration process, the solution of problem (A.1) is the optimal solution of problem (1). The perturbed KKT conditions form a system of nonlinear equations, which can be solved by Newton's method. We linearize the KKT conditions to obtain:

$$\begin{aligned} & \nabla_x^2 f(\mathbf{x}^{\tau}(\mathbf{t}))\Delta x - \nabla_x^2 h(\mathbf{x})\lambda\Delta x \\ & \quad - \nabla_x^2 g(\mathbf{x})(z_1 + z_2)\Delta x - \nabla_x h(\mathbf{x})\Delta\lambda \\ & \quad - \nabla_x g(\mathbf{x})(\Delta z_1 + \Delta z_2) = -L_x, \\ & \nabla_x h(\mathbf{x})^T\Delta x = -L_\lambda, \\ & \nabla_x g(\mathbf{x})^T\Delta x - \Delta s_1 = -L_{z_1}, \\ & \nabla_x g(\mathbf{x})^T\Delta x + \Delta s_2 = -L_{z_2}, \\ & Z_1\Delta s_1 + S_1\Delta z_1 = -L_{s_1}, \\ & Z_1\Delta s_1 + S_2\Delta z_2 = -L_{s_2}. \end{aligned} \tag{A.3}$$

Where $S_1$ and $S_2$ represent the diagonal matrices of the slack variables $s_1$ and $s_2$, respectively. Similarly, $Z_1$ and $Z_2$ represent the diagonal matrices of the Lagrange multipliers $z_1$ and $z_2$, respectively.

The corrections for each iteration can be obtained by solving Eq. (A.3) $\left(\Delta w = [\Delta x, \Delta\lambda, \Delta s_1, \Delta s_2, \Delta z_1, \Delta z_2]\right)$, which is commonly referred to as the Newtonian direction. After obtaining the Newtonian direction, the variables are updated by the following equation:

$$\begin{cases} x = x + \alpha_p\Delta x, & \lambda = \lambda + \alpha_d\Delta\lambda, \\ s_1 = s_1 + \alpha_p\Delta s_1, & s_2 = s_2 + \alpha_p\Delta s_2, \\ z_1 = z_1 + \alpha_d\Delta z_1, & z_2 = z_2 + \alpha_d\Delta z_2. \end{cases} \tag{A.4}$$

Where $\alpha_p$ and $\alpha_d$ are the step sizes. In our implementation, these step sizes are updated using the Adam optimization algorithm as follows:

$$\begin{aligned} \alpha_p' &= \alpha_p - \eta\frac{\frac{\beta_1 m_{\alpha_p,t-1} + (1-\beta_1)\nabla_{\alpha_p} L_t}{1-\beta_1^t}}{\sqrt{\frac{\beta_2 v_{\alpha_p,t-1} + (1-\beta_2)(\nabla_{\alpha_p} L_t)^2}{1-\beta_2^t}} + \epsilon}, \\ \alpha_d' &= \alpha_d - \eta\frac{\frac{\beta_1 m_{\alpha_d,t-1} + (1-\beta_1)\nabla_{\alpha_d} L_t}{1-\beta_1^t}}{\sqrt{\frac{\beta_2 v_{\alpha_d,t-1} + (1-\beta_2)(\nabla_{\alpha_d} L_t)^2}{1-\beta_2^t}} + \epsilon}, \end{aligned} \tag{A.5}$$

where $\alpha_p'$ and $\alpha_d'$ denote the updated step sizes. $\beta_1$ is the decay rate of the first-order momentum term (0.9), and $\beta_2$ is the decay rate of the second-order momentum term (0.999). $\eta$ is the learning rate, and $\epsilon$ is a small constant ($10^{-8}$) to prevent division by zero. After updating the variables, the new values are used as the initial values for the next iteration until the optimal solution is obtained.

## Appendix B. Experimental setup and reproducibility details

This appendix provides additional implementation details for the key performance metrics and the Optimal Power Flow (OPF) case study to enhance the reproducibility of our results.

### B.1. Frequency-domain metric calculation

The frequency-domain metrics used in the high-frequency component analysis (Section 3.5) were calculated with the following parameters:

- **Spectral Similarity:** The similarity was computed based on the magnitude of the Fast Fourier Transform (FFT) applied to the entire prediction sequence on the test set. The raw time series, sampled at 1 Hz, was used directly without a windowing function to avoid introducing artificial spectral artifacts. For computational efficiency, the signal was zero-padded to the next power of two before the FFT was performed.
- **High-Frequency MSE/MAE:** To isolate the high-frequency components, a 5th-order Butterworth high-pass filter was applied to both the ground truth and predicted signals. The cutoff frequency was set to correspond to a period of 10 min. A zero-phase digital filtering approach was used to ensure that no phase shift was introduced by the filtering process, allowing for a direct point-wise comparison of the resulting high-frequency signals.

### B.2. OPF cost calculation parameters

The cost parameters for the OPF case study (Section 3.3 and Table 2) were configured as follows:

- **Currency and Time Scale:** All costs are presented in **U.S. Dollars (USD)** and represent the total system operational cost over the specified forecast horizons (15 min, 6 h, and 20 h).
- **Grid and Load Data:** The simulations were performed on the standard IEEE 39-bus and 118-bus systems with a grid frequency of 60 Hz. The system load profiles were based on the standard datasets accompanying these test cases. To isolate the impact of tidal forecast accuracy, the generation costs for conventional thermal units were modeled using typical quadratic cost functions, while real-time electricity market price volatility was not considered.
- **Tidal Generation Cost Model:** The components of the Levelized Cost of Energy (LCOE) framework, such as CAPEX and OPEX, were estimated using generalized values derived from the techno-economic analysis literature on tidal energy projects, as referenced in the main text [20,21].

## Data availability

The authors do not have permission to share data.

## References


[1] Huihui W, Alharthi M, Ozturk I, Sharif A, Hanif I, Dong X. A strategy for the promotion of renewable energy for cleaner production in G7 economies: By means of economic and institutional progress. J Clean Prod 2024;434:140323.

[2] Xia Y, Wang J, Zhang Z, Wei D, Cao Z, Li Z. A wind speed point-interval fuzzy forecasting system based on data decomposition and multiobjective optimizer. Appl Soft Comput 2024;165:112084.

[3] Tian J, Ooka R, Lee D. Multi-scale solar radiation and photovoltaic power forecasting with machine learning algorithms in urban environment: A state-of-the-art review. J Clean Prod 2023;426:139040.

[4] Deb M, Yang Z, Haas K, Wang T. Hydrokinetic tidal energy resource assessment following international electrotechnical commission guidelines. Renew Energy 2024;120767.

[5] Todeschini G, Coles D, Lewis M, Popov I, Angeloudis A, Fairley I, Johnson F, Williams A, Robins P, Masters I. Medium-term variability of the UK's combined tidal energy resource for a net-zero carbon grid. Energy 2022;238:121990.

[6] Coles D, Wray B, Stevens R, Crawford S, Pennock S, Miles J. Impacts of tidal stream power on energy system security: An isle of wight case study. Appl Energy 2023;334:120686.

[7] Shamsi M, Cuffe P. Prediction markets for probabilistic forecasting of renewable energy sources. IEEE Trans Sustain Energy 2022;13(2):1244–53.

[8] Su H-Y, Huang C-R. Enhanced wind generation forecast using robust ensemble learning. IEEE Trans Smart Grid 2020;12(1):912–5.

[9] Meng Z, Guo Y, Sun H. An adaptive approach for probabilistic wind power forecasting based on meta-learning. IEEE Trans Sustain Energy 2024;15(3):1814–33.

[10] Ma X, Jin Y, Dong Q. A generalized dynamic fuzzy neural network based on singular spectrum analysis optimized by brain storm optimization for short-term wind speed forecasting. Appl Soft Comput 2017;54:296–312.

[11] Chen Y, Samson SY, Lim CP, Shi P. Multi-objective estimation of optimal prediction intervals for wind power forecasting. IEEE Trans Sustain Energy 2023;15(2):974–85.

[12] Wang J, Wang K, Li Z, Lu H, Jiang H. Short-term power load forecasting system based on rough set, information granule and multi-objective optimization. Appl Soft Comput 2023;146:110692.

[13] Yang H, Wu Q, Li G. An ocean tidal energy point-interval forecasting system based on enhanced auxiliary feature, mode decomposition combined with compressive sensing and attention interaction. J Clean Prod 2024;475:143680.

[14] Zhang Y, Liu C, He J. Data-driven prediction of tidal stream power using deep neural networks and environmental variables. Renew Energy 2023;205:987–96.

[15] Al-Sumaiti A, Meki R, Rezk H. Deep learning-based short-term tidal height forecasting using LSTM networks. Appl Energy 2022;324:119685.

[16] Lin J, Jiang N, Zhang Z, Chen W, Zhao T. LMQFormer: A laplace-prior-guided mask query transformer for lightweight snow removal. IEEE Trans Circuits Syst Video Technol 2023;33(11):6225–35.

[17] Pöppelbaum J, Chadha GS, Schwung A. Contrastive learning based self-supervised time-series analysis. Appl Soft Comput 2022;117:108397.

[18] Huang G, Laradji I, Vazquez D, Lacoste-Julien S, Rodriguez P. A survey of self-supervised and few-shot object detection. IEEE Trans Pattern Anal Mach Intell 2022;45(4):4071–89.

[19] Bazrafshan K, Mohsenian-Rad H. Power system optimization modeling in GAMS. Springer; 2017.

[20] Johnstone CM, Pratt D, C.A. G, Burrows R. A techno-economic analysis of tidal energy technology. Renew Energy 2013;49:101–6.

[21] Coles D, Angeloudis A, Greaves D, Hastie G. A review of the UK and british channel islands practical tidal stream energy resource. Renew Sustain Energy Rev 2021;145:111095.

[22] Elghali A, Ziani E-H, Benbouzid M. Techno-economic optimal sizing design for a tidal stream Turbine–Battery system. J Mar Sci Eng 2023;11(3):679.

[23] Avriel M. Nonlinear programming. In: Mathematical programming for operations researchers and computer scientists. CRC Press; 2020, p. 271–367.

[24] Qiu T, Xie Y, Niu H, Xiong Y, Gao X. Enhancing masked time-series modeling via dropping patches. In: Proceedings of the AAAI conference on artificial intelligence, vol. 39, 2025, p. 20077–85.

[25] Goswami M, Szafer K, Choudhry A, Cai Y, Li S, Dubrawski A. MOMENT: A family of open time-series foundation models. In: International conference on machine learning. 2024.

[26] Coles DS, Angeloudis A, Greaves D, Hastie GDM, Lewis M, Mackie L, McNaughton J, Miles J, Neill SP, Piggott MD, et al. A national-scale seasonal and interannual tidal stream energy resource assessment for Great Britain. Renew Sustain Energy Rev 2021;147:111223.

[27] Gallego A, Olauson J, Agren O, Astrand P, Salcedo-Sanz S, Topper M. On the characterisation of tidal flows: A system-driven methodology for the analysis of ADCP measurements. Renew Energy 2020;153:100–11.

[28] Guillou N, Chapalain G, Thiébot J. Vertical profile of tidal currents: A review for tidal stream energy resource assessment. Renew Sustain Energy Rev 2020;127:109877.

[29] Qian P, Feng B, Liu X, Zhang D, Yang J, Ying Y, Liu C, Si Y. Tidal current prediction based on a hybrid machine learning method. Ocean Eng 2022;260:111985.

[30] Li M, García-Jalón I, Lopez J, Mader J, Corman D, Ferrer L. A robust quality control and imputation methodology for HF radar ocean current data. Ocean Dyn 2022;72(8):577–95.

[31] Weiß CH, Aleksandrov B, Faymonville M, Jentsch C. Partial autocorrelation diagnostics for count time series. Entropy 2023;25(1):105.

[32] Gu Y, Ren H, Liu H, Lin Y, Hu W, Zou T, Zhang L, Huang L. Simulation of a tidal current-powered freshwater and energy supply system for sustainable island development. Sustainability 2024;16(20):8792.

[33] Dora BK, Rajan A, Mallick S, Halder S. Optimal reactive power dispatch problem using exchange market based butterfly optimization algorithm. Appl Soft Comput 2023;147:110833.

[34] Hewage P, Behera A, Trovati M, Pereira E, Ghahremani M, Palmieri F, Liu Y. Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. Soft Comput 2020;24:16453–82.

[35] Hu C, Cheng F, Ma L, Li B. State of charge estimation for lithium-ion batteries based on TCN-LSTM neural networks. J Electrochem Soc 2022;169(3):030544.

[36] Zhang Y, Wang Z, Liu W, Li Y, Zhang J. A machine learning model based on GRU and LSTM to predict the $CO_2$ concentration in a layer house. Sensors 2024;24(1):244.

[37] Grzenda M, Wojcik P, Dziemiańczuk M, Zabolotny W, Gorawski M. Comparative analysis of deep learning models for real-world ISP network traffic forecasting. ACM Trans Knowl Discov Data 2025.
[38] Innocenti S, Matte P, Fortin V, Bernier N. Analytical and residual bootstrap methods for parameter uncertainty assessment in tidal analysis with temporally correlated noise. J Atmos Ocean Technol 2022;39(10):1457–81.
[39] Zhang Y, Yan J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: Int. conf. learn. represent.. ICLR, 2023.
[40] Wang Z, Yuan Y, Zhang S, Lin Y, Tan J. A multi-state fusion informer integrating transfer learning for metal tube bending early wrinkling prediction. Appl Soft Comput 2024;151:110991.
[41] Zeng A, Chen M, Zhang L, Xu Q. Are transformers effective for time series forecasting? In: Proc. AAAI conf. artif. intell., vol. 37, 2023, p. 11121–8.
[42] Lee-Thorp J, Ainslie J, Eckstein I, Ontanon S. FNet: Mixing tokens with Fourier transforms. In: Proc. NAACL-HLT. 2022, p. 4296–313.
[43] Perlin M, Bustamante MD. A robust quantitative comparison criterion of two signals based on the Sobolev norm of their difference. J Engrg Math 2016;101(1):115–24.
[44] Zhang H, Li B, Su S-F, Yang W, Xie L. A novel hybrid transformer-based framework for solar irradiance forecasting under incomplete data scenarios. IEEE Trans Ind Inf 2024.