基于监督式机器学习的近海波高预测

陈思宇 林育佳 梁颖鹏 袁梓昭 张文雅

摘要：近海有效波高的预测对海上军事活动、海洋能源的开采与保护、海洋渔业、港口经济等人类近海活动的许多领域有着重要的作用，随着计算机的发展，利用监督式机器学习的方法，可以有效弥补传统数值模型预测的计算高成本及低时效。前人的研究中长短期记忆网络 LSTM 已被广泛用于海温预测，但是在近海波高预测的使用研究尚少。而机器学习中，支持向量机 SVR 已被较广泛使用，但是对于模型建立的认识没有系统的探究。因此本文将对比长短期记忆网络 LSTM 模型与传统模型及支持向量机 SVR 两种机器学习模型在近海波高预测的效果。通过与传统预测模型对比，探究监督式机器学习方式预测波高的效果及限制。同时通过多组实验，改变不同输入、输出维度、预测长度，归纳最佳的模型输入输出方式，探究其在近海波高预测中的实用性。

关键词：波高预测、长短期记忆网络、支持向量机

1. 引言

21 世纪被称为是海洋世纪，海洋发展是 21 世纪经济社会发展的主要方向。其中近海有效波高的预测与海上军事活动、海洋能源的开采与保护、海洋渔业、港口经济等人类近海活动的许多领域有着不可分割的关系，因此建立一套精确且高效的近海波高预报模型是十分必要的。

目前用于海洋有效波高的预测方法主要有海洋数值模型和统计方法模型两种。海洋数值模型已发展到第三代海浪数值模式 WAM、WAVEWATCH Ⅲ 以及 SWAN 等[1-3]，数值预报[4]是建立在明确的物理过程上的，根据主要影响因素而建立的物理方程能够很好的将物理机制用数学公式来表达，但因为影响海浪波高的因素是众多的，又不可能将所有的影响因子都考虑在内，所以想要建立一套预测精度极高的波浪预测模型是有一定难度的。相对来说，统计方法的预报[5]不需要

预设物理条件，只需对大数据隐藏的规律进行分析，找出其中的关联性利用统计方法即可进行预报,与数值模型相比，统计方法的计算成本更低，能延长预报的时效，甚至能得到更高的预报精度。

因为近年来计算机突飞猛进的发展,计算能力大幅度的提升使得人工神经网络成为当下热点问题,已被应用到许多领域问题,其中包括海洋近海波高的预报。机器学习是计算机对大数据进行系统学习、归纳分析、总结推算的一门多领域交叉学科。机器学习可分为浅层学习和深度学习两类，深度学习是当前机器学习领域中最热门的分支[6]，可分为有监督式学习、无监督式学习、半监督式学习和强化学习四类。有监督式学习是较为常用的一种学习模式,人工神经网络[7-9]便是其中的一种，例如 ANN、RNN、LSTM 等神经网络。

在 2002 年,陈希[10]等人运用 BP 神经网络建立了南海硇洲岛海区风浪的预报方案，预报达到一定的精度，并针对网络容易产生振荡，易发生不收敛现象作了对比分析；2005 年，齐义泉等人[11]利用一个浮标的波高观测值研究了 ANN 改进海浪数值模型的精度的应用，模型修订之后的准确度达 80%以上，但是物理响应过程仍然不够，对极端事件的波高预测误差较大。2016 年，王红萍[12]利用小波神经网络对海浪谱进行预报，预报的有效时长达到 6 小时。

近年来，长短期记忆网络（long short-term memory, LSTM）被提出并被运用在海洋要素的预报研究中。2017 年，Qin Zhang 等人[13]基于 LSTM 来模拟海温的时间关系来做预测，利用全连通层将 LSTM 层的输出映射到最终预测，验证了 LSTM 用于海温预报的有效性；2018 年，朱贵重等人[14]通过 LSTM-RNN 建立了西太平洋研究海区的 SST 时间序列变化模型,通过加入其它物理参数使得模型准确性提高 31%并且使得模型结果可解释性提高；2019 年，高丽赋[15]运用 LSTM 在海浪波高预测在台湾海峡及其周围海域进行尝试，建立了预报相关系数最高达 0.96 的预测模型。2020 年，王国松等人[16]基于再分析数据和观测数据利用 LSTM 对沿海风速进行预报，取得一定的进展。

另外，支持向量机（Support Vector Machine, SVM）也是机器学习中较为热门的回归模型，它具有非常优秀的泛化能力。2009 年，Mahjoobi 等[17]将风速作为输入因子建立了有效波高的预测模型，并指出该方法优于人工神经网络，且计算时间相对较少。2019 年，金权等人[18]采用 SVM 对有效波高进行预测，取波

浪场和风场作为输入的特征向量，结果显示与再分析数据的相关系数达 99%。2020 年，王燕等人[19]基于 SVR 建立了在渤海的近海有效波高的预测模型，并发现将未来风速的信息作为模型输入可提高模型在 12h、24h 的预报精度。

总的来说，LSTM 的非线性拟合能力强，能提高模型结果的可解释性，一定程度上解决 RNN 梯度爆炸、梯度消失、长时间依赖性的问题，但 LSTM 用于近海波高预测的研究较少；而 SVR 已广泛用于海洋要素的预报，虽然其预测精度较令人满意，但仍有进一步优化的空间，所以本文将利用 LSTM 和 SVR 对近海有效波高的短期预测进行探索。

2 数据来源及处理

2.1 数据选取及处理

以汕尾站（115.210°E,22.45°N）为本次实验研究点。本文实验使用的数据主要有 1993-1995 年三年的汕尾站潮水位高度（数据来源：ftp://ftp.soest.hawaii.edu/uhslc）以及距离汕尾站最近的四个经纬度整数点（115°E,22°N、115°E,23°N、116°E,22°N、116°E,23°N）的十米风速、十米风向、平均波向、平均波浪周期、组合风浪涌浪的显著高度（以下简称波高）（数据来源：https://www.ecmwf.int/，数据集：CERA-20C Ocean Wave, Daily）

表 1 数据特征值

	十米风速	十米风向	波高	波浪方向	波浪周期	潮水位高度
数据位置		115°E,22°N 116°E,22°N	115°E,23°N 116°E,23°N			115.210°E 22.45°N
测量方式			浮标			未知
精度			10^{-9}			1
时间间隔			3h			1h

除潮水位高度数据外，数据区域网格精度为 1°×1°，处理方法为：取距离汕尾站最近的四个经纬度整数点（115°E,22°N、115°E,23°N、116°E,22°N、116°E,23°N）的数据进行以距离为权重的加权平均作为研究点的数值。十米风速、

十米风向数据由极坐标转化为直角坐标。

所有数据均进行归一化处理至-1 到 1 的范围，避免由于量级相差悬殊对预测结果准确性影响。

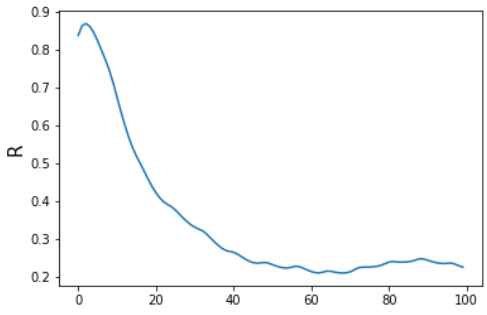
2.2 数据分析

2.2.1 数据的相关性

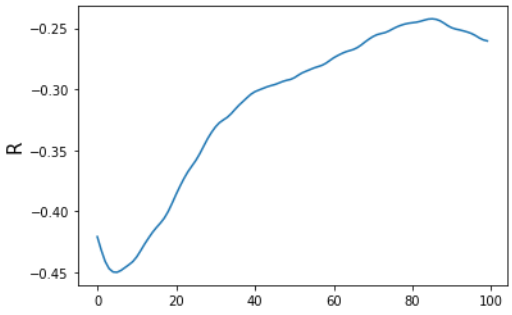
各项要素与波高的相关性程度有所差别，风及潮汐作用引起的波高变化需要一定时间的积累，因此各项因素对于波高的影响在时间上有一定滞后性，以及先前时刻的波高积累会对当前时刻波高有所影响。因此，需要对不同时刻的各项要素与当前时刻波高的相关性进行分析，为模型输入变量的选择提供参考。

图 1 表示：波高与十米风速、波高本身有较强的相关性，与十米风向、波浪周期、波浪方向、潮水位高度相关性弱。因此在选取预测模型变量时，需要优先考虑十米风速及波高本身两个变量。

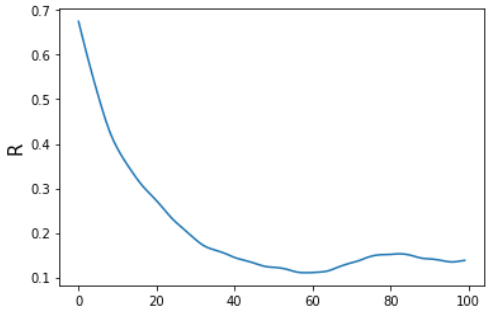
对于十米风速而言，随着时间前推，与波浪相关性先上升，而后下降，前 3h 的风速与波高相关性达到最高。波高与其本身的相关性随着时间前推而下降。由于实验获取的波高数据时间间隔为 3h，因此在建立预测模型时，仅需考虑前 3h 的十米风速和波高两个变量对预测结果的影响。数据的相关性分析为我们选择模型输入变量结构提供了重要参考。



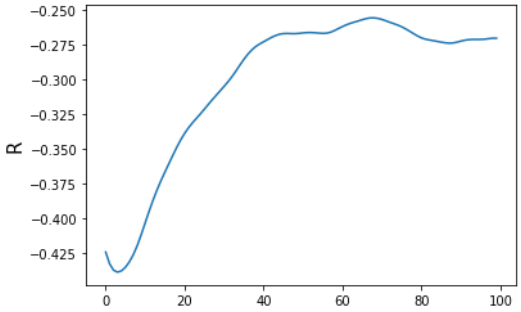
（a）波高与前面 x 时刻的十米风速相关系数



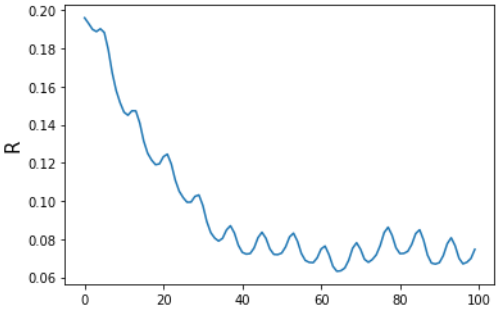
（b）波高与前面 x 时刻的波浪方向相关系数



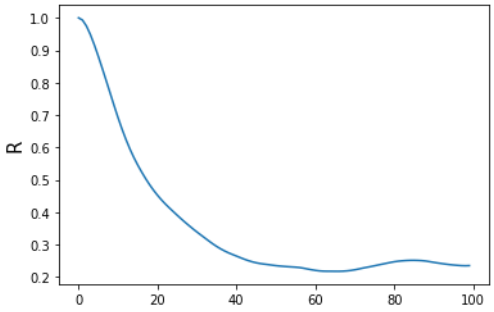
（c）波高与前面 x 时刻的波浪周期相关系数



（d）波高与前面 x 时刻的十米风向相关系数



（e）波高与前面 x 时刻的潮汐相关系数



（f）波高与前面 x 时刻的波高相关系数

图 1 波高及不同因素、时刻的相关系数

2.2.2 数据的统计特性

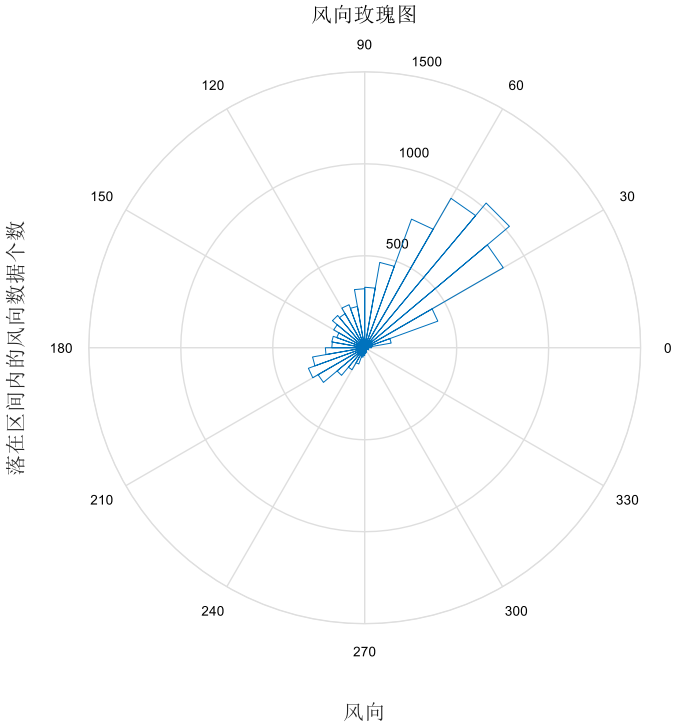
我们计算了每个数据变量的全体平均值，最低值和最大值，如下表 1 所示：

表 2 数据变量统计参数

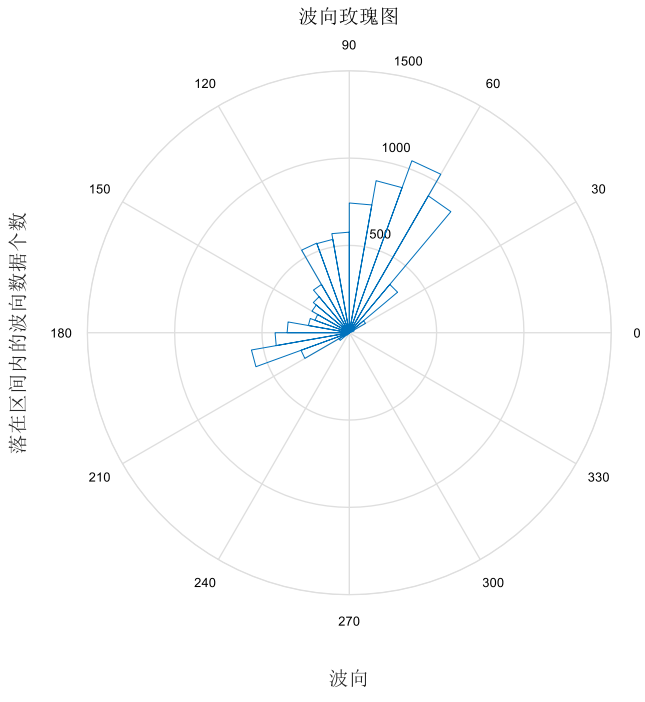
	风速 m/s	波高 m	波浪周期 s	潮水位 mm
平均值	4.501	1.016	5.525	1327.646
最小值	2.000	0.301	3.829	80.000
最大值	11.085	3.209	8.031	2870.000

注：在数据量级上由于单位不同，量级存在较大差别，在构建神经网络模型时需要归一化到同一量级。

对于风向和波向两个变量，由于变量的首尾连接性，我们选择用玫瑰图的方式来展示数据的统计特性，如下图 2 所示。可见，风向和波向都主要集中在东北方向，符合我们的常识认知，也说明了波浪主要是由风引起的。由于风向和波向一直都处于变动状态，某一个时刻的风可能削弱波高也可能加大波高值，需要考虑该时刻的风向和波向的相对关系，后面的实验结果也显示了考虑风向的影响的预测结果要优于不考虑风向的影响的预测结果。



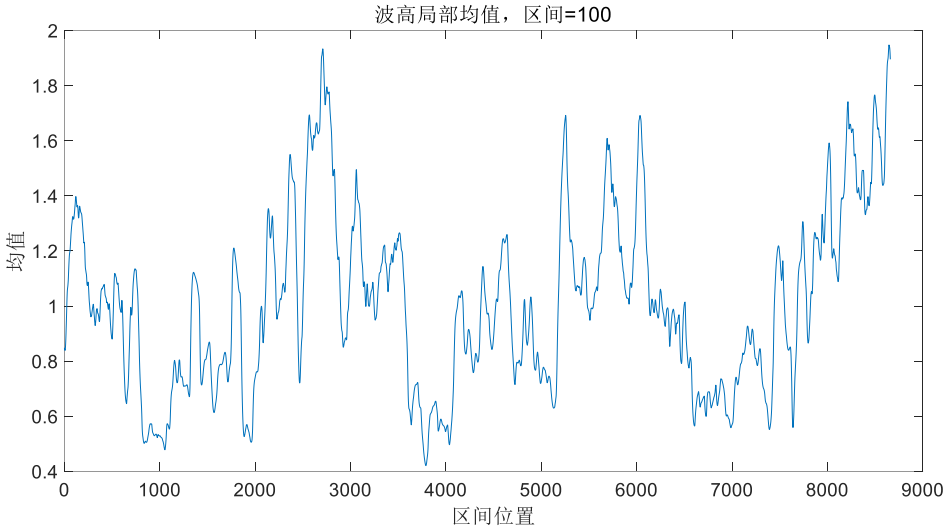
(a) 风向玫瑰图



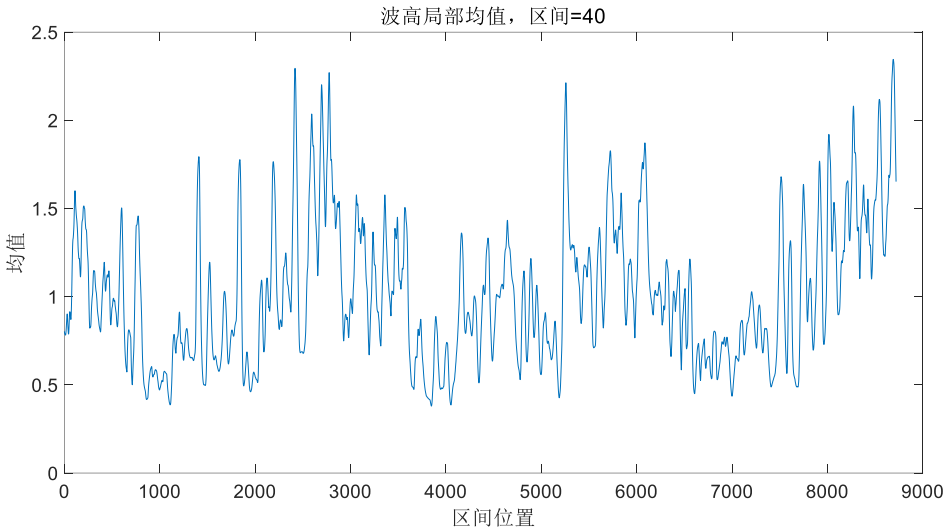
(b) 波向玫瑰图

图 2 风向、波向玫瑰图

我们在使用传统数值预测模型自回归滑动平均 ARMA 和灰色模型 GM（1，1）时，需要考虑数据的平稳性和平滑性。所以我们使用级比的方式来检验数据的平滑性，使用局部均值序列法和自相关函数法检验数据的平稳性。



(a)波高局部均值在区间位置上的变化（区间=100）



(b)波高局部均值在区间位置上的变化（区间=40）

图 3 波高局部均值在区间位置上的变化

从图 3 可以发现，局部均值随着时间推移是变化的，存在一个明显的波动趋势，可见波高序列是不平稳的。使用局部均值序列法检验平稳性存在一定的直观性，但是不够准确，我们进一步使用自相关函数法检验平稳性。

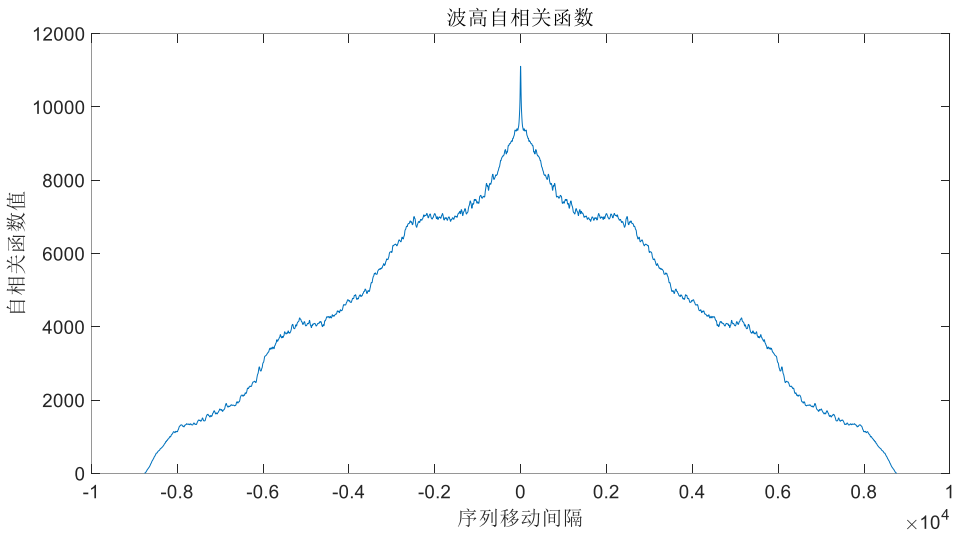


图 4 波高自相关函数值变化图

图 4 可以看出，波高自相关函数不是迅速下降的，而是缓慢下降的，所以波高不是一个平稳序列，而是一个有周期性的序列。为了消除平稳性，我们对数据进行一次差分，得到差分序列。并对差分序列的平稳性进行检验，检验结果如下图 6 所示：

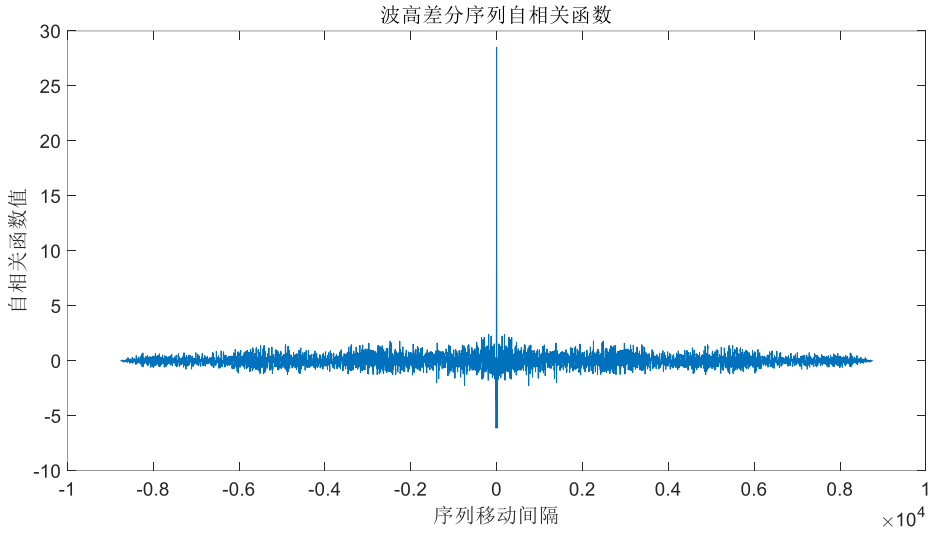


图 5 波高差分序列自相关函数

经过一次差分后，可以看见，数据已经平稳，自相关函数很快下降到零附近。因此，可以对一次差分序列使用 ARMA 模型，再使用累和法还原为波高序列。

下面我们使用级比法检验数据的平滑性。级比序列的最低值是 0.8909，最大值是 1.7024。级比区间是 0.8115>0.2，所以波高序列不是平滑的序列，存在

突然陡增现象。因此对于使用传统预测模型 GM（1，1）预测波高可能是不合适的。后面的实验结果也验证了这一点。

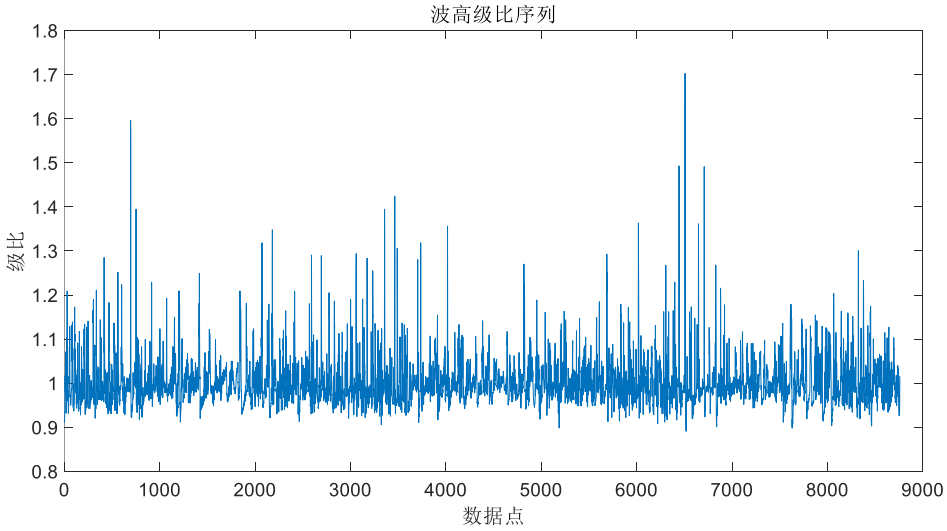


图 6 波高级比序列

2.3 人工神经网络

人工神经网络（ANN）是一种具有自学习性，自适应和很强的非线性函数逼近的能力的数学模型，是对真实神经网络的一种抽象。其中人工神经网络的单元被称为神经元。一般的神经元的结构如下：

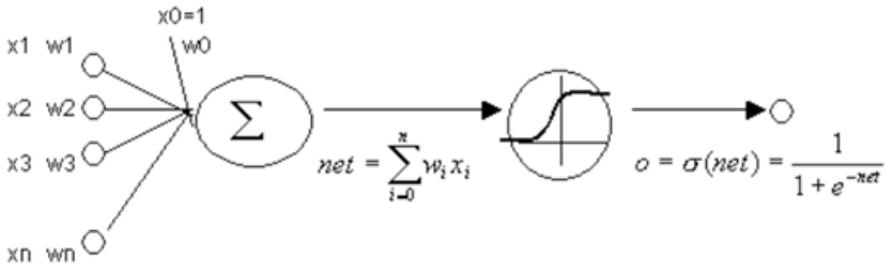


图 7 单个神经元结构

这里 x 为神经元的输入，w 为各输入的权重，加起来以后通过激活函数得到该神经元的输出。将若干个这样的神经元组合连接可以形成全连接神经网络如下图 8 所示：

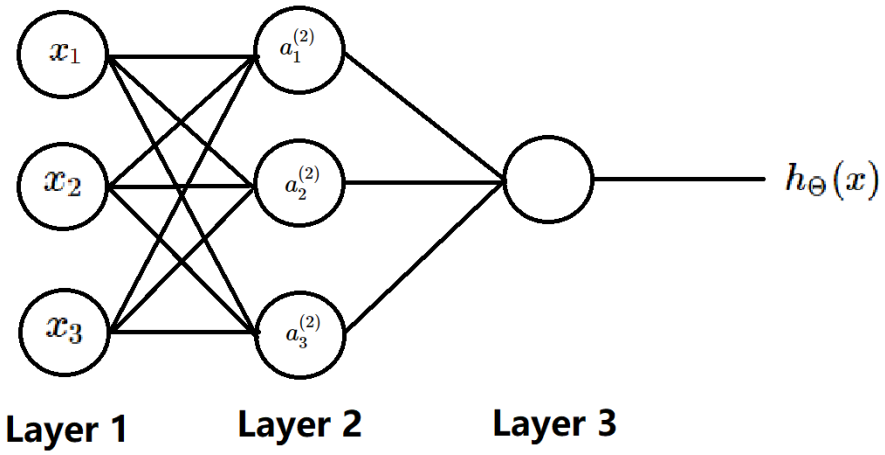


图 8 全连接神经网络结构

以上展现了一个三层神经网络，其中 Layer 1 是输入层，Layer 2 是隐藏层层，Layer 3 是输出层。本层的每一个神经元都与前一层的所有神经元相连接，这就是全连接神经网络。每一个输出都要通过激活函数，常用的激活函数有 Relu 函数，tanh 函数，sigmoid 函数等。通过对数据的传入和输出可以对神经网络的参数进行调整。本课题中采用 Adam 算法对神经网络的参数进行调整。Adam 算法过程如下：

步长：ε（默认为 0.001）

（1） 矩估计的指数衰减速率，ρ_1 和 ρ_2 在区间[0,1]内，默认为 0.9 和 0.999.

（2） 用于数值稳定的小常数 δ，默认为 10e-8.

（3） 初始参数 θ；

（4） 初始化时间步 t=0；

（5） 初始化一阶和二阶矩变量 s=0，r=0；

（6） while 没有达到停止准则 do

（7） 从训练集中采包含 m 个样本$\{x^{(1)},\cdots,x^{(m)}\}$的小批量，对应目标为 $y^{(i)}$

（8） 计算梯度：$g \leftarrow \frac{1}{m}\nabla_\theta \sum_i L(f(x^{(i)};\theta),y^{(i)})$

（9） t←t+1

（10） 更新有偏一阶矩估计$s \leftarrow \rho_1 s + (1-\rho_1)g$

（11） 更新有偏二阶矩估计$r \leftarrow \rho_2 r + (1-\rho_2) g \odot g$

（12） 修正一阶矩的偏差$\hat{s} \leftarrow \frac{s}{1-\rho_1^t}$

（13） 修正二阶矩的偏差$\hat{r} \leftarrow \frac{r}{1-\rho_1^t}$

（14） 计算更新$\Delta\theta = -\epsilon \frac{\hat{s}}{\sqrt{\hat{r}}+\delta}$

（15） 应用更新$\theta \leftarrow \theta + \Delta\theta$

（16） end while

除了神经网络的层数，激活函数，输入输出的神经元个数以外，神经网络的神经元也是可以调整的，对于随时间变化的波高预测，我们认为处理时间序列的神经网络 LSTM 能有更好的效果。LSTM 结构主要包含三个“门”的结构，分别是遗忘门，输入门和输出门。完整的 LSTM 过程如下：

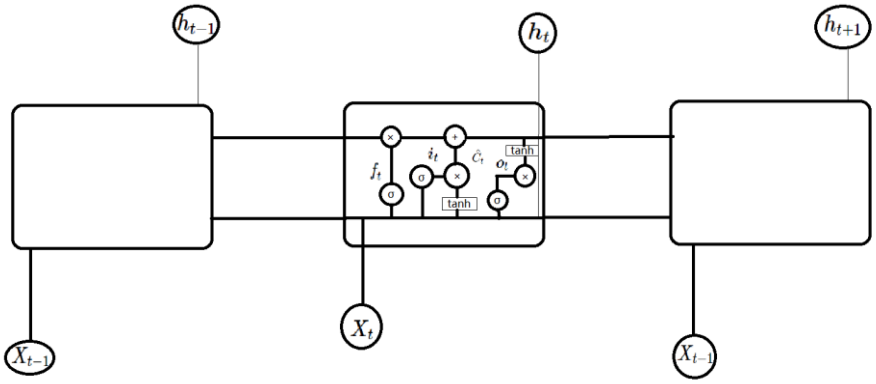


图 9 LSTM 结构图

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_i)$$

第一步，h_{t-1} 表示上一个 cell 的输出，x_t 表示当前 cell 的输入，σ 表示 sigmoid 函数。这一步通过一个遗忘门层完成，sigmoid 函数输出范围是(0, 1)，0 表示完全舍弃，1 表示完全保留。

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$

$$C_t = tanh(W_C * [h_{t-1}, x_t] + b_C)$$

第二步是决定让多少新的信息加入到 cell 状态中来。一个 sigmoid 函数决定哪些信息需要更新，一个 tanh 层生成一个向量，即备选的用来更新的内容。

$$\sigma_t = \sigma(W_O * [h_{t-1}, x_t] + b_O)$$

$$h_t = \sigma_t * \tanh(C_t)$$

最后要确定输出的值，先用一个 sigmoid 层确定 cell 状态的哪个部分将输出出去，然后把这个状态通过 tanh 函数进行处理并将它和 sigmoid 门的输出相乘。

2.4 支持向量机

支持向量机（SVM）是模式识别中广义肖像算法发展而来的分类器，通过推广至回归问题的 SVM 算法又称为支持向量回归（SVR）。设回归模型为$f(X) = \omega^T X + b$，若样本点与回归模型足够接近（小于 ε），则该样本不计算损失。则 SVR 即可表示成如下优化问题：

$$\max_{\omega,b} \frac{1}{2}||\omega||^2$$

$$s.t|y_i - f(X)| \leq \in$$

引入松弛变量ξ，ξ^* 表示 ε −不敏感损失函数的分段取值后可得：

$$\max_{\omega,b} \frac{1}{2}||\omega||^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*)$$

$$s.t|y_i - f(X)| \leq \in + \xi_i$$

$$f(X) - y_i \leq \in + \xi_i$$

$$\xi \geq 0，\ \xi^* \geq 0$$

引入拉格朗日乘子$\alpha, \alpha^*, \mu, \mu^*$可得到拉格朗日函数和对偶问题

$$L(\omega, B, \xi, \xi^*, \alpha, \alpha^*, \mu, \mu^*) = \frac{1}{2}\left|\left|\omega\right|\right|^2 +$$

$$C\sum_{i=1}^{N} - \sum_{i=1}^{N}\mu_i\,\xi_i - \sum_{i=1}^{N}\mu_i^*\,\xi_i^* + \sum_{i=1}^{N}\alpha_i[f(X_i)i - y_i - \in - \xi_i] +$$

$$\sum_{i=1}^{N}\alpha_i^*[f(X_i)i - y_i - \in - \xi_i^*]$$

$$\max_{\alpha,\alpha^*}\sum_{i=1}^{N}[y_i(\alpha_i^* - \alpha_i) - \in(\alpha_i^* + \alpha_i)] - \frac{1}{2}\sum_{i=1}^{N}\sum_{J=1}^{N}[(\alpha_i^* - \alpha_i)\,(X_i)^T X_j\left(\alpha_j^* - \alpha_j\right)]$$

$$s.t\sum_{J=1}^{N}[(\alpha_i^*-\alpha_i)=0,0\le\alpha_i\alpha_i^*,\ \alpha_i^*\le C$$

最终求得的 SVR 形式为

$$f(X)=\sum_{i=1}^{N}(\alpha_i^*-\alpha_i)X^TX+b$$

这样得到的 SVR 模型是线性的，若引入核函数把样本点映射到高维空间可以得到非线性的 SVR 模型。设映射函数$\phi(X)$表示从原始的特征空间映射至更高维的希尔伯特空间，则此时预测的 SVR 模型对应的超平面为$\omega^T\phi(X)+b=0$，定义映射函数的内积为核函数

$$\kappa(X_1+X_2)=\Phi(X_1)^T\Phi(X_2)$$

常见的核函数有线性核函数，多项式核函数，拉普拉斯核函数，Sigmoid 核函数等。

3 模型建立及实验结果对比

3.1 预测结果评价指标

在本次研究中，我们着眼于用于预测近海波高的机器学习的模型的建立，主要建立了两种模型，分别是 LSTM 与 SVR。通过对模型 MSE 等的分析评估模型的质量。主要使用的指标是：

相关指数(R^2)：

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}$$

均方根误差(RMSE)：

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

平均绝对误差(MAE)：

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

3.2 LSTM、SVR 模型的建立

训练集、验证集、测试集的数量比例为 8:1:1.

对于 LSTM 神经网络，在 Tensoflow 框架下，我们利用如下经验公式确定每层的神经元个数和神经网络的层数：其中，I 表示数据输入的维度，0 表示数据输出的维度。

$$Size = \sqrt{0.43IO + 0.12O^2 + 2.54I + 0.77O + 0.35}$$

$$N = [\sqrt{IO}]$$

隐藏层的激活函数用 Relu 函数，使用 Adam 算法进行优化，学习率定为 1e-3，为了避免过拟合，使用提前停止机制，在连续两次优化中损失函数没有变小的时候提前停止训练。每一轮训练结束后用验证集验证，每次训练最多 30 个 epochs。训练结束后计算 R^2，RMSE，MAE。

以上对于学习率的取定基于以下的实验事实：我们使用风速的横纵分解、潮水位作为输入数据，输出数据是波高。LSTM 的层数与每层的神经元数由经验公式确定。得出预测结果 R^2 与学习率的关系。下图的横坐标是学习率的对数，纵坐标是 R^2。可见，在一定范围内，随着学习率的上升，模型的预测效果上升，同时上升速度减慢，最后几乎趋近于一条水平线。所以在接下来的实验中，学习率全部取定为 1e-3。

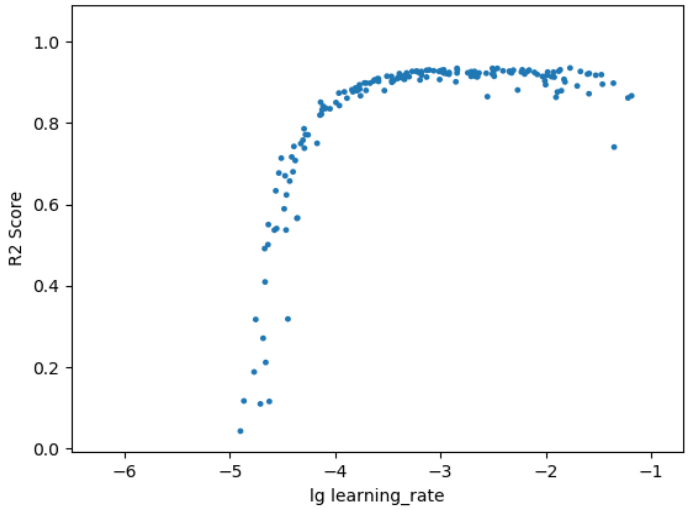


图 10 学习率与 R2 的关系

对于 SVR 算法，我们利用 sklearn 里的 SVM 库建立。经过前置的尝试训练，我们认为在本实验中所用的核函数应为线性核函数，ε 的选取经过多次尝试性训练比较（图 11），最终选定了 1e-2。最后利用多输出机器学习的 MultiOutputRegressor 将输出维度扩大。

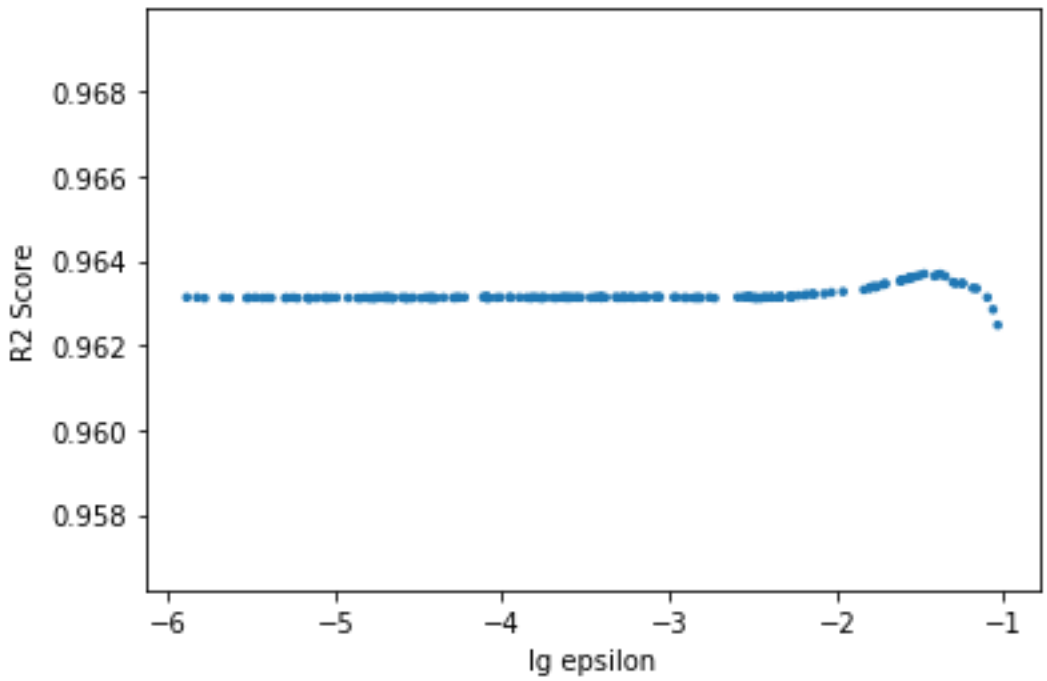


图 11 epsilon 与 R2 的关系

3.3 传统数值模型预测结果

3.3.1 滑动平均法

我们尝试了多种滑动长度，即用前面 x 个时刻的值作为下一个时刻的的值的预测，发现直接以当前时刻的值作为下一个时刻的预测值是最好的。通过计算，均方根误差只有 0.0486，R^2 有 0.9875。表 3 给出了不同的滑动窗口长度的预测精度。可以看见，随着滑动平均窗口长度的增加，预测精度是下降的，也就是说太久远的历史波高值对预测没有好处。

表 3 滑动平均法不同滑动窗口长度的预测精度值

窗口长度	均方根误差
20	0.2753

10	0.1937
5	0.1244
4	0.1076
3	0.0895
2	0.0699
1	0.0486

3.3.2 加权移动平均法

考虑到越远的历史时刻值对未来预测值的影响较低，较近的时刻对未来的影响较高，我们采用加权平均法来进行预测，权重采用相关系数。举例说，用前面三个时刻的波高值预测未来一个时刻，则计算波高 t 时刻的序列与波高（t-1）时刻，波高（t-2）时刻，波高（t-3）时刻的相关系数 R1，R2，R3，然后取R1/(R1+R2+R3)作为波高（t-1）时刻的权重，以此类推。

可以发现，同滑动平均法一致，都是窗口长度越短，预测效果越好，同时可以发现，加权比不加权的预测效果要好。

表 4 加权移动平均法不同滑动窗口长度的预测精度值

窗口长度	均方根误差
10	0.1848
5	0.1227
2	0.0698
1	0.0486

3.3.3 指数平滑法

（1） 一次平滑：

计算结果如表 5 所示，可以发现随着 α 值的增加，预测效果要好，而 α 值越大说明较近的波高值对未来的影响越重要，而较远的历史波高值对未来时刻的影响越低。反之亦然。这说明，不需要考虑太多历史时刻的波高值，只需取当前时刻作为下一个时刻的预测值即能获得较佳的预测效果。

表 5 指数一次平滑法不同指数 α 值的预测精度值

指数 α 值	均方根误差

1.0	0.0486
0.8	0.0583
0.6	0.0732
0.4	0.0993
0.2	0.1560

（2） 二次平滑：

计算结果如表 6 所示，这里的结果说明了同样的事实，也就是较近的历史时刻值的权重很高。但同时我们发现均方根误差比用当前时刻值作为下一个时刻预测值的情况要低，这说明并不是只考虑当前值作为预测值是最佳的，较远历史时刻值依然对未来时刻波高值存在一些影响。

表 6 指数二次平滑法不同指数 α 值的预测精度值

α 值	T+1 时刻预测值均方根误差	T+2 时刻预测值均方根误差
0.9	0.0309	0.0717
0.8	0.0332	0.0741
0.6	0.0413	0.0827

3.3.4 多元线性回归

我们构建自变量为波高前九个时刻、因变量为波高当前时刻的多元线性回归模型，并用拟合的结果进行预测，比较预测值和真实值的差异，计算得到的均方根误差是 0.0495（图 12）。

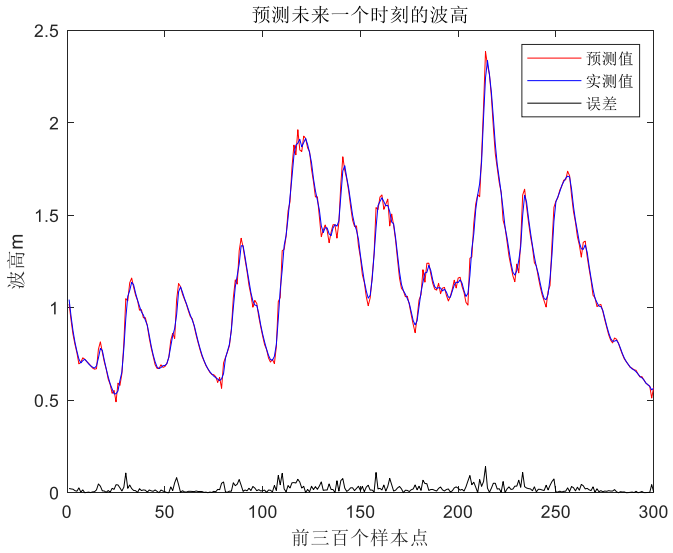


图 12 多元线性回归预测未来一小时波高图

我们将预测的未来一个时刻的波高值和现有波高值一起输入模型预测未来两个时刻的波高值，预测结果均方根误差为 0.0679（图 13）。

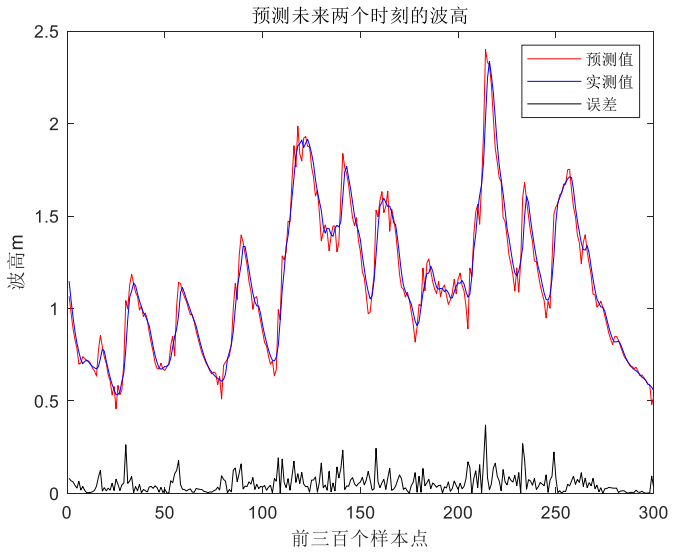


图 13 多元线性回归预测未来两小时波高图

3.3.5 灰色模型 GM（1，1）

用四十个已知数据预测十个未来数据，预测结果显示，预测值和真实值存在较大误差，未来一个时刻和两个时刻的预测结果和真实值的均方根误差达到了 0.18 和 0.2241。

3.3.6 ARMA 模型

首先，我们使用差分法消除波高数据的不平稳性。考虑了定阶的结果和数据的平稳性后我们选择 ARMA（2, 2）模型。用高阶的 ARMA 模型预测需要更多的样本数据，为了便于比较不同预测模型的优劣，我们依然采用历史 40 个波高数据预测未来 10 个时刻的波高值，另一方面可以避免数据平稳性不满足特征方程根的定解条件，无法求出 ARMA 的系数。预测结果如表 7 所示，预测未来一个时刻的均方根误差可以低至 0.0342。

表 7 未来十个时刻的预测均方根误差

1	2	3	4	5	6	7	8	9	10
0.034	0.080	0.130	0.180	0.229	0.274	0.318	0.361	0.407	0.454

3.3.7 传统数值预测模型小结

滑动平均，加权滑动平均，指数平滑，三种方法都指示出只采用前一时刻作为下一时刻的预测值效果最好，均方根误差可以低至 0.0309。灰色预测精度一般，预测未来一个时刻和两个时刻的均方根误差分别是 0.18 和 0.22。ARMA（2, 2）模型，预测未来一个时刻的均方根误差可以低至 0.0342。将风速和波高取对数后，以前一时刻波高和前一时刻风速为自变量，当前时刻波高为因变量，进行线性拟合，均方根误差是 0.0417。多元线性回归模型实验中，自变量是历史前九个时刻的波高值，Y 是当前时刻的波高值，均方根误差为 0.0495。

综上，预测未来一个时刻的波高时，0.0309 是最理想的预测均方根误差。

3.4 LSTM 模型实验结果

（1）实验一：

固定除了输入结构、输入输出维度以外的所有参数不变，得到不同输入结构下最好的拟合结果及对应的输入输出维度。下表 8 展示 LSTM 模型在不同输入模型下的拟合结果，IN、OUT 分别代表输入、输出维度。

表 8 LSTM 不同输入模型下的拟合结果

输入结构	RMSE最小值	RMSE平均值	RMSE标准差	R2最大值	IN	OUT
所有	0.041914358	0.237151043	0.039420136	0.993019	10	1
潮水位	0.478638513	0.576753091	0.03097038	0.089605	28	2
风速	0.143599383	0.288867289	0.048268807	0.918499	26	6
波高	0.039550001	0.231675655	0.039748926	0.993784	15	1
风向	0.436249763	0.514092895	0.026242265	0.248288	27	8
风速+风向	0.138821519	0.294535543	0.044905308	0.923289	22	1
风速+潮水位	0.143441578	0.27982657	0.037259665	0.918368	26	3
风速+波高	0.039823076	0.236747964	0.043234507	0.993687	25	1
潮水位+波高	0.040939041	0.234588374	0.042381917	0.993316	33	1
风向+潮水位	0.39545556	0.49417801	0.032144492	0.377484	36	2
潮水位+风向+波高	0.039542767	0.229462679	0.040436253	0.993764	30	1
风向+波高	0.038916947	0.23265074	0.04228923	0.99396	32	1
潮水位+风速+波高	0.040800281	0.230695509	0.033814712	0.993362	35	1
风速+风向+波高	0.038910429	0.239692775	0.043888271	0.993973	20	1
风速+风向+潮水位	0.136486509	0.293288822	0.048814985	0.925848	23	1

综合来看，风速+潮水位+波高的输入模型是最佳的。它拥有最小的均方根误差平均值和最小的标准差。其中 IN 与 OUT 表示取得最好 RMSE 时的输入维度与输出维度。

上面的结果显示，每个模型中取得最佳结果的一组实验的输出维度几乎都为 1，并且都有较长的历史数据输入但是和前面所做的传统预测相比并无优势，多元线性回归模型实验中，自变量是历史前九个时刻的波高值，Y 是当前时刻的波高值，R2 达到 0.998，均方根误差低至 0.0279，比 LSTM 预测的结果要好，所需要的输入数据也更少（只需九个历史时刻的波高数据即可，不需要风速、风向等数据）。ARMA 预测模型预测未来一个时刻的均方根误差可以低至 0.0342，也优于 LSTM。但是 LSTM 的预测结果比滑动平均法、加权滑动平均、灰色预测模型等算法较优。

下面我们探讨比较各个模型预测未来两个时刻的预测精度。表 9、表 10 给出了不同模型预测未来两个时刻的预测情况：

表 9 LSTM 不同模型预测未来两个时刻的预测结果

输入结构	RMSE最小值	IN	OUT
所有	0.091212534	19	2
潮水位	0.496031462	28	2
风速	0.172979163	29	3
波高	0.090305033	24	3
风向	0.436578437	27	8
风速+风向	0.167884645	10	3
风速+潮水位	0.169589381	18	2
风速+波高	0.093566298	25	2
潮水位+波高	0.092195088	27	2
风向+潮水位	0.403608591	36	2
潮水位+风向+波高	0.088335835	11	3
风向+波高	0.087452309	18	2
潮水位+风速+波高	0.094655741	15	2
风速+风向+波高	0.089952614	13	2
风速+风向+潮水位	0.168274808	32	3

表 10 传统模型预测未来两个时刻的预测结果

模型	ARMA（2,2）	二次指数平滑法（平滑指数α=0.9）	多元回归模型
RMSE（T+2）	0.080	0.072	0.0679

从上表分析，人工神经网络 LSTM 模型的对未来两个时刻的波高的预测效果并不比传统预测方法优越，均方根误差略低于传统方法。另外，LSTM 取得最好的预测结果时都具有较高的输入维度和较低的输出维度。

以下实验采用的模型均为风速+潮水位+波高。

（2）实验二：

图 14 分别展示输出维度为 1，3，6，10 时，输入维度与 R^2 的关系。

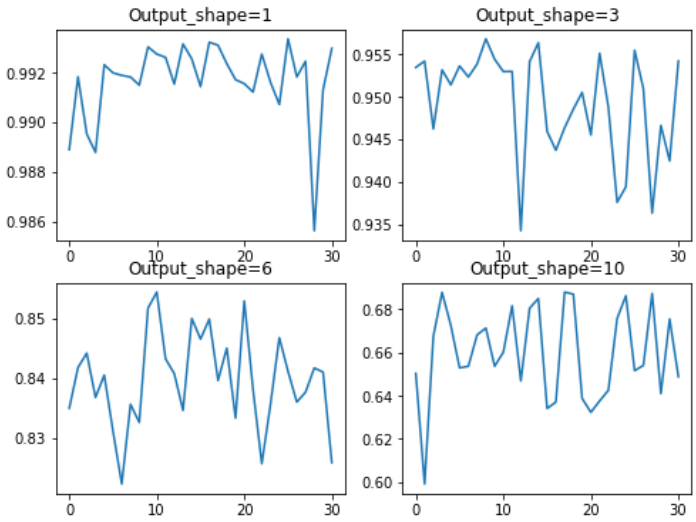


图 14 LSTM 输入维度与 R^2 变化规律

就预测长度为 1 而言，延长输入长度几乎对预测效果没有特别大的影响，变化幅度非常小。随着预测长度的不断变大，R2 的变化越发剧烈，没有明显规律。

图 15 对 Output_shape=1 做了具体分析，可以发现不同的 Input_shape 下的预测结果几乎重合。

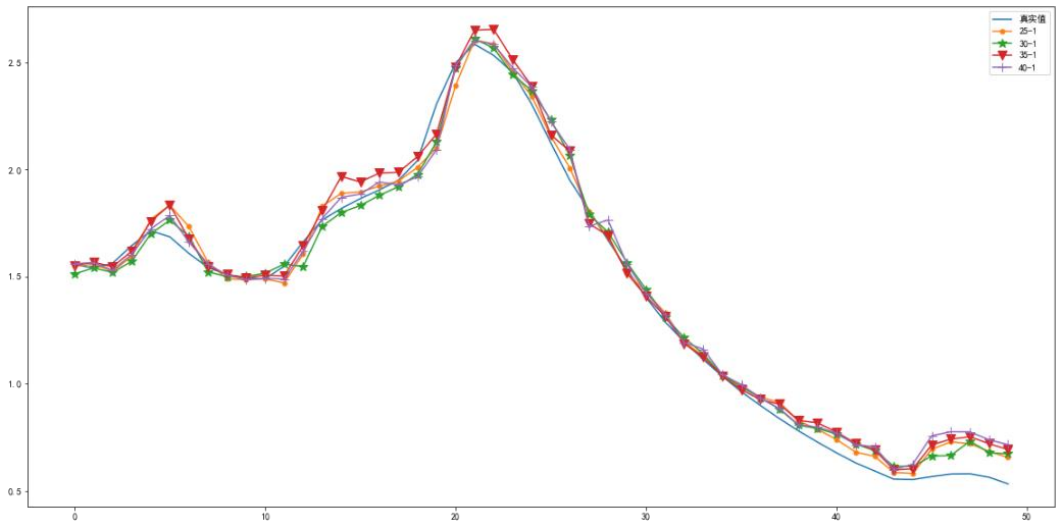


图 15 不同的 Input_shape 预测结果

（3）实验 4：

图 16 展示了在输入维度与序号不变的情形下，输出维度与 R^2 的关系。固定一个序号，当输入维度上升的时候，预测效果没有显著变化，这验证了实验二的规律。

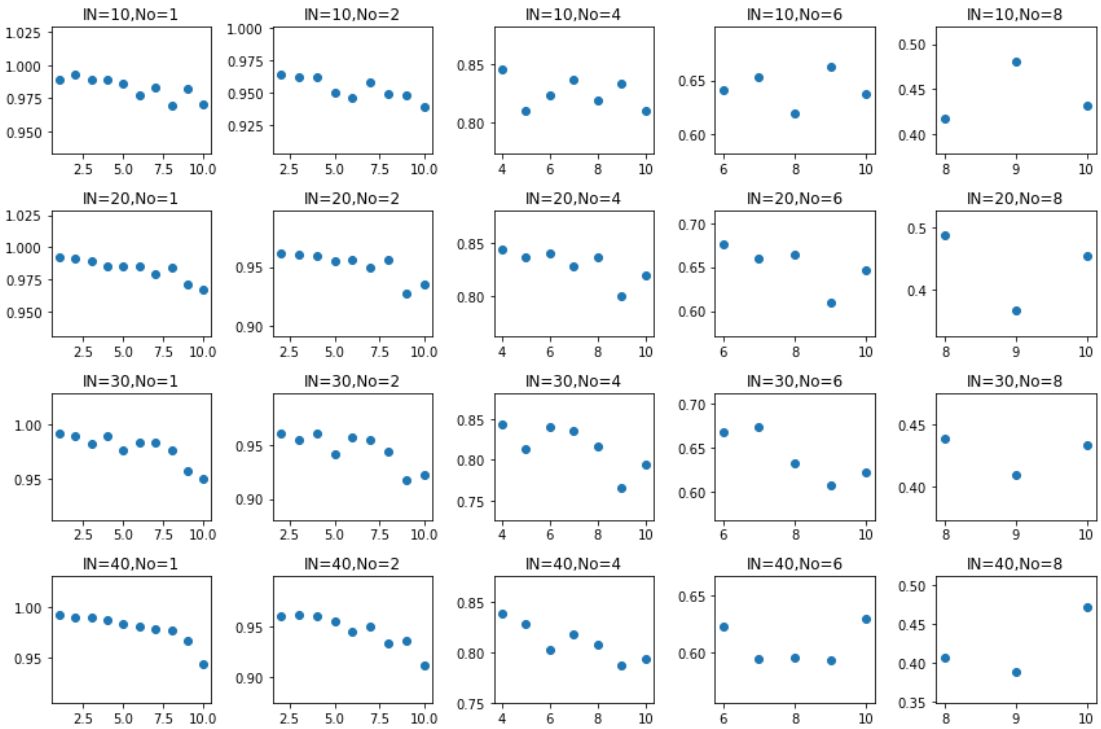


图 16 LSTM 输出维度与 R2 的关系

图 17 展示预测长度变化时的预测结果。可以看到，在变化剧烈时，预测结果只是被简单地平移了。这说明即使在短距离预测下，LSTM 的预测结果也不好。

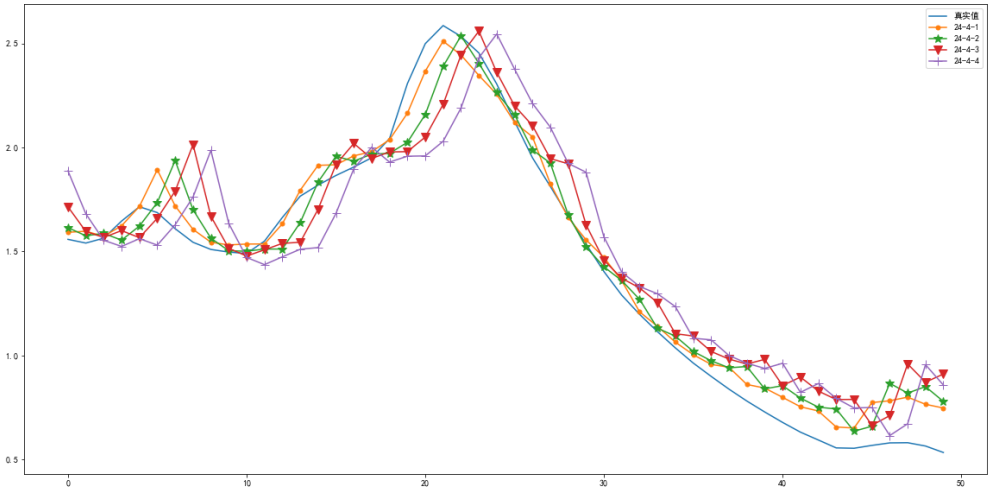


图 17 LSTM 不同预测长度与 R2 的关系

3.5 支持向量回归模型实验结果

（1）实验一：

改变不同的输入结构会得到不同的预测结果，展示如表 10 所示。可以看到，一旦添加波高这个因素，RMSE 便能显著提高。在此前提下，SVR 比 LSTM 的预测效果要好，同时也是所有模型中最好的。与 LSTM 显著不同的是，它的最值均在 OUT 较大时取到。

表 10 SVR 不同输入结构下的预测结果

输入结构	RMSE最小值	RMSE平均值	RMSE标准差	R2最大值	IN	OUT
风速+风向	0.514233043	0.525659676	0.004161366	0.133612	10	8
潮水位	0.475646164	0.495662934	0.010194951	0.262152	20	2
风速	0.248722309	0.301697467	0.02004151	0.798038	28	1
风向	0.543593807	0.546946632	0.00090822	0.028533	30	10
波高	0.030576154	0.18858403	0.025745482	0.99695	40	10
风速+风向+潮水位	0.492633856	0.500883736	0.001902761	0.20868	11	3
风速+风向+波高	0.03005879	0.183079998	0.024190569	0.997052	37	8
潮水位+风速	0.245541218	0.297591855	0.019314768	0.803168	40	5
风向+潮水位	0.487311959	0.502154777	0.006710071	0.226052	22	10
潮水位+波高	0.030447077	0.187122062	0.025289595	0.996976	39	10
风速+波高	0.030141891	0.186563172	0.025292275	0.997036	40	10
风向+波高	0.03050964	0.186645679	0.025077237	0.996964	34	10
潮水位+风速+波高	0.030150539	0.185505225	0.024957242	0.997035	35	10
潮水位+风向+波高	0.030472415	0.185916102	0.02484868	0.996972	35	10
所有	0.030024377	0.182603853	0.024128775	0.997059	40	10

取预测长度为 2 时，找到 R^2 取最值的位置。结果如表 11 所示

表 11 预测长度为 2 时 R^2 最值取到的位置

输入结构	RMSE最小值	IN	OUT
风速+风向	0.520963158	10	5
潮水位	0.480830555	39	2
风速	0.25799145	26	4
风向	0.546811548	16	7
波高	0.073456671	40	10
风速+风向+潮水位	0.493607923	10	5
风速+风向+波高	0.071973091	35	10
潮水位+风速	0.254511228	37	2

（接下表）

（接上表）

风向+潮水位	0.491153331	20	3
潮水位+波高	0.07301871	37	8
风速+波高	0.072336369	40	9
风向+波高	0.073221923	35	9
潮水位+风速+波高	0.072053307	37	8
潮水位+风向+波高	0.072994285	40	10
所有	0.071613381	36	8

根据两次实验所得结果，认为输入模型为风速+潮水位+波高+风向的拟合度最高，故下文探讨是均采用该输入模型结构。

（2）实验二：

图 18 分别展示输出维度为 1，3，6，10 时，输入维度与 R^2 的关系。在 Out 较小的时候，In 越多结果越精确，这也符合我们的认知；但 Out 取 6 时，已经看不出规律了；Out 取 10 时，甚至得到了相反的规律。

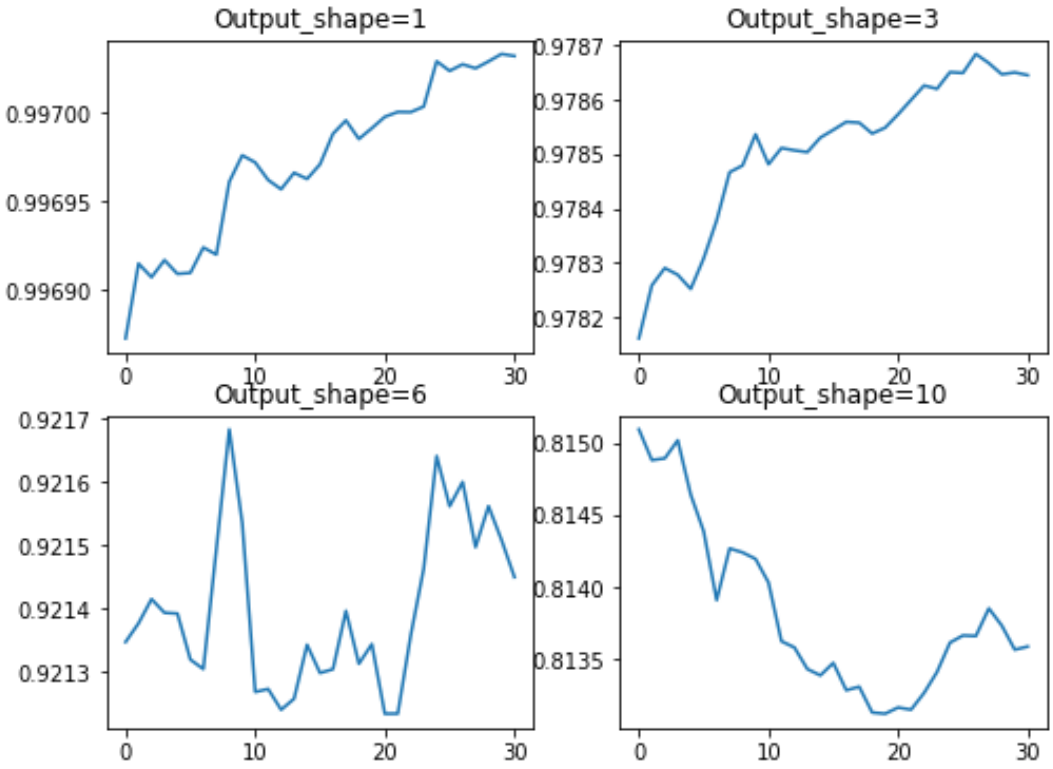


图 18 SVR 不同输入维度与 R2 的关系

（3） 实验三：

图 19 展示预测长度变化时的预测结果。可以看到 SVR 和 LSTM 有着相似的情况，在前半段数据变化剧烈时，预测结果只是被简单地平移，拟合度较差。但是在后半段数据相对平缓的情况下，拟合效果比 LSTM 要好。

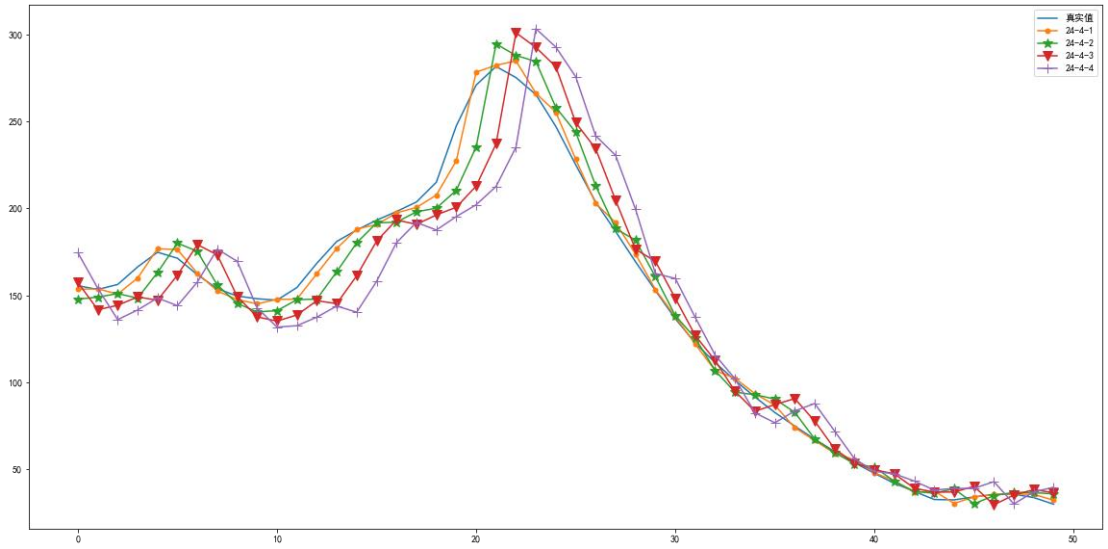


图 19 SVR 不同预测长度与 R2 的关系

4. 结论

（1） 传统预测模型中灰色模型和 ARMA 模型对数据序列的平稳性和平滑性等提出了要求，但是使用人工神经网络模型则不需要考虑数据的平稳性和平滑性。传统数值预测模型中移动平均法、加权移动平均法以及指数平滑法都显示只使用较近的历史数据比使用较长的历史数据能得出更好的预测精度，而人工神经网络的实验结果也表明，使用较远的历史数据反而会降低预测的精度。因此在实际应用中，不需要考虑太久远的历史数据对未来波高预测的影响。

（2） 对于 LSTM，波高预测的最佳输入结构为波高+风速+潮水位，对于 SVR，波高预测的最佳输入结构为波高+风速+潮水位+风向。与实验前的相关性分析结果有部分一致性，输入结构较少时，含有波高及风速的输入结果对应的结果精确较高。而潮水位及风向，在相关度分析中，与当前时刻的波高相关性较低，但是在模型建立中仍然有必要作为输入结构的一部分。因

此相关性分析仅能作为输入结构设定的参考，输入结构的组成也为我们探究影响波高的海洋要素提供指示。

（3） 对于 LSTM，输入维度和输出维度的比值需要较大，在预测精度较高的模型中，输出维度一般仅为 1。相对于 SVR 模型，精度较高的模型中，输入维度和输出维度比值较小，即输入维度较少，输出维度相对较多，与我们通常认为的输入越多输出越少对精确度越好不一致。但是在输出较少时，仍然满足输入越多，精确度越高的规律。因此在设定输入输出结构时，无能完全按照高输入低输出这一规律来提高精确度。

（4） LSTM 与传统预测模型在精准度上并没有体现出更大的优势，但是 LSTM 对于输入数据的平滑度要求较小，即在实际运用于数据平滑度较差的波高数据中，LSTM 仍然有一定的意义。在短周期预测中，SVR 的预测精度较高，高于 LSTM 及全部传统模型，而较长时间的预测中与 LSTM 存在相同问题，在数据波动幅度较大时拟合精度差，优势欠缺。结合 LSTM 和 SVR 的预测速度较快，成本较低的优势，在实际运用中，LSTM 和 SVR 更适合于即时性预测，如航海等方面。

参考文献：

[1]Group T W. The WAM model—A third generation ocean wave prediction model[J]. Journal of Physical Oceanography, 1988, 18(12): 1775-1810.

[2] Booij N, Ris R C, Holthuijsen L H. A third‐generation wave model for coastal regions: 1. Model description and validation[J]. Journal of geophysical research: Oceans, 1999, 104(C4): 7649-7666.

[3] Tolman H L. User manual and system documentation of WAVEWATCH III TM version 3.14[J]. Technical note, MMAB Contribution, 2009, 276: 220.

[4]王关锁，乔方利，杨永增．基于 MPI 的 LAGFD-WAM 海浪数值模式并行算法研究[J]．海洋科學進展，2007, 25(4): 401-407.

[5]杨德全, 郝日栩, 何健新, 何忠杰. 统计预报方法在海洋预报中的应用研究进展[J]. 海洋信息, 2019,34(02):1-9.

[6]胡越, 罗东阳, 花奎, 路海明, 张学工. 关于深度学习的综述与讨论[J]. 智能系统学报, 2019,14(01):1-19.

[7]Deo M C, Jha A, Chaphekar A S, et al. Neural networks for wave forecasting[J]. Ocean engineering, 2001, 28(7): 889-898.

[8]Makarynskyy O. Improving wave predictions with artificial neural networks[J]. Ocean Engineering, 2004, 31(5-6): 709-724.

[9] Makarynskyy O, Pires-Silva A A, Makarynska D, et al. Artificial neural networks in wave predictions at the west coast of Portugal[J]. Computers & geosciences, 2005, 31(4): 415-424.

[10]陈希, 沙文钰, 李妍, 张韧. 人工神经网络技术在海浪预报中的应用[J]. 海洋通报, 2002(02):11-15.

[11]齐义泉, 张志旭, 李志伟, 李毓湘, 施平. 人工神经网络在海浪数值预报中的应用[J]. 水科学进展, 2005(01):32-35.

[12]王红萍, 余义德, 张丹. 基于小波神经网络的波浪参数预报[J]. 舰船电子工程, 2016,36(12):80-84.

[13]Zhang Q, Wang H, Dong J, et al. Prediction of sea surface temperature using long short-term memory[J]. IEEE Geoscience and Remote Sensing

Letters, 2017, 14(10): 1745-1749.

[14]朱贵重, 胡松. 基于 LSTM-RNN 的海水表面温度模型研究[J]. 应用海洋学学报, 2019,38(02):191-197.

[15]高丽斌, 郭民权, 张少涵, 张振昌. 基于长短期记忆网络的波高预报[J]. 福建电脑, 2018,34(08):105-107.

[16]王国松, 王喜冬, 侯敏, 齐义泉, 宋军, 刘克修, 吴新荣, 白志鹏. 基于观测和再分析数据的 LSTM 深度神经网络沿海风速预报应用研究[J]. 海洋学报, 2020,42(01):67-77.

[17]Mahjoobi J, Mosabbeb E A. Prediction of significant wave height using regressive support vector machines[J]. Ocean Engineering, 2009, 36(5): 339-347.

[18]金权, 华锋, 杨永增. 基于 SVM 的海浪要素预测试验研究[J]. 海洋科学进展, 2019, 第 37 卷(2):199-209.

[19]王燕, 钟建, 张志远. 支持向量回归的机器学习方法在海浪预测中的应用[J]. 海洋预报, 2020,37(03):29-34.