DOI:10.11737/j.issn.1003-0239.2023.02.001

基于主成分分析和LSTM神经网络的海温预报模型

李竞时[1,2]，匡晓迪[1,2]，李琼[3]，何恩业[1,2]，张聿柏[3]，袁承仪[4]，张延琳[5]

(1.国家海洋环境预报中心,北京 100081;2.国家海洋环境预报中心 自然资源部海洋灾害预报技术重点实验室,北京 100081;3.山东省海洋预报减灾中心,青岛 266104;4.天津科技大学,天津 300222;5.辽宁省自然资源事务服务中心,辽宁 沈阳 110033)

摘 要： 利用荣成、海阳两站的自建浮标海温观测数据以及区域大气模式WRF(Weather Research and Forecasting)的气象数值预报数据,基于主成分分析(Principal Component Analysis,PCA)法和长短时记忆(Long Short-Term Memory,LSTM)神经网络,提出了适用于单站海表温度预报的PCA-LSTM海温预报模型。该模型可以提供24～120 h预报时效的海温预报,预测效果比数值模型和统计模型明显提高。

关键词： 主成分分析;长短时记忆神经网络;海温预报

中图分类号：P731.31 **文献标识码：**A **文章编号：**1003-0239(2023)02-0001-10

0 引言

海表温度(Sea Surface Temperature,SST,也称海温)是海水最重要的物理性质之一,对于全球气候、海洋生态、海气相互作用等有着巨大的影响[1-2]。海温的变化在很大程度上决定了海水密度的变化,海温是海洋动力场、声速场最显著的影响因子之一[3];海温对海水养殖、渔业资源分布和渔场位置的确定等也有着重要影响[4];此外,在海难事故中,海温与落水人员的幸存率也有着直接的关系。因此,海温预报在海洋生态环境保护、近岸经济建设、海洋渔业资源开发、国防建设和海洋研究等方面都有着极其重要的意义。

我国的海温预报研究工作始于20世纪60年代初。目前近岸海温预报方法主要包括经验预报方法、统计预报方法和数值预报方法等[5-7]。随着国家海洋经济的持续发展,传统的海温经验预报和海温统计预报已经逐渐难以满足我国近海海水浴场、滨海旅游度假区、水产养殖区、渔场等区域城市生产和生活活动中不断增长的预报需求。在近岸和浅水海域,受地形、潮汐和海陆作用等因素的影响,目前海温数值预报的精度与传统的海温预报方法仍存在一定差距[8]。将海温数值预报和中国近岸海域基础预报单元海温预报指导产品相结合,可以解决海温预报的覆盖问题,为我国近海海温预报提供较准确的指导产品。但由于观测资料无法覆盖所有预报点位,部分站点的预报误差较大[9]。

近年来,许多人工神经网络如误差反向传播神经网络(Back-Propagation Neural Network,BP)、卷积神经网络(Convolutional Neural Networks,CNN)、循环神经网络(Recurrent Neural Network,RNN)等已经在海温预报、潮汐预报[10-11]、海浪预报[12-13]、渔场预报[14-15]等海洋预报业务中得到了广泛应用,并且取得了出色的成果。匡晓迪等[8]利用BP神经网络方法,基于气象预报数据、台站观测数据和数值海温预报结果建立了定点海温精细化数值预报释用模型,并将定点数值预报的误差从2.2 ℃减少至0.7 ℃。WU等[15]使用BP神经网络方法结合经验模式分解(Empirical Mode Decomposition,EMD)建立

收稿日期：2022-11-25;**修回日期：**2022-12-16。
基金项目：自然资源部海洋环境信息保障技术重点实验室开放基金课题资助;海洋预警监测(海温预报释用服务试点)(SDGP37000000002021-02003589);国家自然科学基金(41606028)。
作者简介：李竞时(1993-),男,工程师,本科,主要从事海洋环境预报技术研究。E-mail:lijingshi@nmefc.cn

了非线性SST异常预测模型，与线性回归模型相比，预测模型与实际SST具有更高的相关性和较低的均方根误差（Root Mean Square Error，RMSE）。JIA等[16]利用长短时记忆（Long Short-Term Memory，LSTM）神经网络模型建立了东海未来5 d的SST预报模型，该模型能够较好地反映东海海温的季节性变化趋势。贺琪等[17]基于局部加权回归的周期趋势分解（Seasonal-Trend decomposition procedure based on Loess，STL）和LSTM神经网络建立了海温预测模型，并基于卫星遥感观测数据对SST进行预测，结果表明在STL分解基础上应用神经网络模型比单一的神经网络模型的预测精度更高。相比于传统的海温预报方法，人工神经网络的优势在于不依赖明确的物理过程，能够更好地提取数据的变化规律，因此对近岸海洋要素预报有很好的适用性。在开展海温预报时，我们使用的海洋水文、气象要素等数据多为包含大量前后相关信息的时间序列数据，这些数据之间也有复杂的关联性。一些人工神经网络，如RNN及其变种LSTM神经网络等能够考虑未来数据和历史数据的相关性，在连续性强的长时间序列训练中取得较好效果。考虑到海温观测和气象预报数据的特征数量、噪声影响以及时间特性，本文选取了荣成、海阳浮标的海温观测数据和区域大气模式WRF（Weather Research and Forecasting）的气象预报数据，采用主成分分析（Principal Component Analysis，PCA）法对其进行降维和去噪，选取训练要素建立训练数据集，并通过LSTM网络进行特征提取，构建了适用于荣成、海阳两站的PCA-LSTM海温预报模型。

1 数据和方法

1.1 海温观测数据

本文采用2019年1月1日—2021年9月30日荣成、海阳两个站点布设的浮标表层（海面以下0.5 m）日均SST观测数据作为研究对象，共计1 004个样本。对该观测数据进行质量控制，将缺测值（海阳站7个）和除荣成在受台风“烟花”期间以外单日SST突变超过2℃的异常值（海阳站3个，人工质控发现此类观测均为异常观测）替换为前一天的观测值。

1.2 气象数值预报数据

气象数值预报数据来自业务化运行的WRF结果。该模式网格的水平分辨率为10 km，垂向分为34层，时间步长为90 s，开边界采用了美国国家海洋和大气管理局（National Oceanic and Atmospheric Administration，NOAA）全球预报系统（Global Forecasting System，GFS）的预报结果。模式每日预报时效为120 h，逐3 h一次。本文选取了2019—2021年每日24 h的预报结果作为原始训练变量，对海面2 m的气温（$T2$）和比湿（$Q2$）、海平面气压（Sea Level Pressure，SLP）、纬向10 m风速（$U10$）、经向10 m风速（$V10$）共5个要素的日平均值进行研究，每个要素各1 004个样本。

1.3 主成分分析法

直接利用海温观测数据和气象预报数据进行海温预报时，要素较多且复杂，因此需要先对以上数据进行预处理。PCA法是一种统计方法，是将线性相关的原变量进行正交变换后转化为另一组线性无关的综合变量（即主成分），再根据主成分的累计贡献率不低于某值的原则，从中选取出重要程度靠前的综合变量，以达到对原变量进行降维和去噪的目的[18]。它能够消除各原变量之间的相关影响，在尽量减少原变量信息损失的前提下，用较少的主成分分别代表存在于原变量中的各类信息。通过PCA法对原始变量进行预处理后，可以尽可能多地保留原始变量信息，同时减少LSTM神经网络训练时的参数个数，提高模型收敛速度并减少模型训练时间。具体步骤如下：

①对由n维相关变量组成的原始变量集$X_{m\times n}$（m为样本数）进行z分数（z-score）标准化处理，得到均值为0、方差为1的标准化矩阵$\boldsymbol{ZX}$。

②基于标准化数据矩阵$\boldsymbol{ZX}$建立协方差矩阵$\boldsymbol{R}$，利用特征值分解方法求解标准化数据矩阵$\boldsymbol{ZX}$的特征值并将其从大到小排列，得到特征值λ_k（$k=1,2,\cdots,n$）、特征向量G_k与主成分F_k。

③根据方差贡献率和累计方差贡献率确定主成分。本文选取主成分的原则为单个成分的方差贡献率大于5%，累计方差贡献率超过95%，符合条件的p（$p<n$）个主成分F_k的特征值分别为

$\lambda_{k'}(k'=1,2,\cdots,p)$,从而将$n$维原始变量降维得到$p$维主成分向量。

1.4 LSTM神经网络

RNN是一种拥有短期记忆的神经网络,常用于处理序列数据[19],其网络结构如图1所示。RNN由多个链式连接的循环单元构成。$x_t(t=1,2,\cdots,t)$表示第t步时模型的输入;$s_t(t=1,2,\cdots,t)$为第t步时模型隐藏层的状态,根据当前输入层的输出$U\cdot x_t$与上一步隐藏层的状态s_{t-1}进行计算;$o_t(t=1,2,\cdots,t)$为模型第t步时的输出,只由模型当前的隐藏状态s_{t-1}决定;U、V、W为在模型中共享的线性关系参数。通过将上一时刻的输出作为神经元下一个时刻的输入,RNN能够保持数据中的依赖关系,因此非常适合处理时间序列数据。但是人们在实践中发现传统的RNN往往较难实现信息的长期保存,该网络同时也存在梯度消失和梯度爆炸的问题。

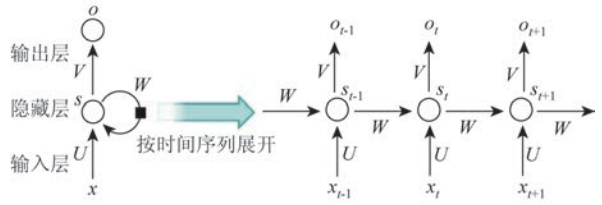


图1　RNN网络结构示意图

Fig.1　Structure of RNN

LSTM网络是RNN的一个变种,由HOCHREITER等[20]提出。LSTM网络通过将隐藏层中的RNN单元替换为LSTM单元,使该网络拥有了长期记忆能力,解决了传统RNN在处理长时间序列数据时存在的梯度消失和梯度爆炸问题。近年来LSTM网络在语音识别、翻译、图片描述等问题中都得到了广泛使用。

相较于传统的RNN单元,LSTM单元的结构更为复杂(见图2)。LSTM单元通过引入"门"机制改变贯穿隐藏层的细胞状态信息,从而实现了信息的长期记忆。每个LSTM单元中都包含输入门(input gate)、输出门(output gate)和遗忘门(forget gate)3类,以控制隐藏单元间的信息收集和传递。输入门用来决定什么细胞状态信息将会更新,遗忘门决定哪些信息将会得到保留,输出门则可以根据需要有选择地输出信息。公式如下:

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

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{3}$$

$$\tilde{c}_t=\tanh(W_c\cdot[h_{t-1},x_t]+b_c) \tag{4}$$

$$c_t=f_t\times c_{t-1}+i_t\times\tilde{c}_t \tag{5}$$

$$h_t=o_t\times\tanh(c_t) \tag{6}$$

式中:i、f、o、c分别代表输入门、遗忘门、输出门、细胞状态;W和b分别为对应的权重系数矩阵和偏置项;σ和tanh分别为Sigmoid激活函数和双曲正切激活函数;h_t为t时刻的输出;$\tilde{c}_t$为t时刻细胞状态更新值。

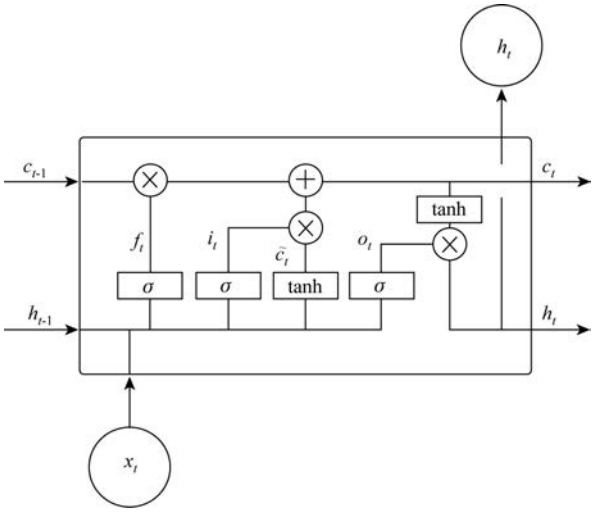


图2　LSTM单元

Fig.2　LSTM unit

1.5 PCA-LSTM海温预报模型

本文构建的PCA-LSTM海温预报模型的整体框架如图3所示,包括数据预处理、LSTM神经网络、结果输出等3个功能模块。数据预处理模块负责对输入的SST、$T2$、$Q2$等6个气象和水文原始变量x_1—x_6进行预处理,通过z-score标准化和PCA等方法(标准化、主成分分析系数均来自训练集)得到满足网络输入要求的降维后的各输入主成分F_1—F_4。LSTM神经网络模块为使用LSTM单元(见图2)搭建的单层循环神经网络,通过加入线性层实现模型的多变量多步预测。在结果输出模块中对LSTM神经网络输出的预测结果进行反标准化,得到24～120 h时效的海温预报。

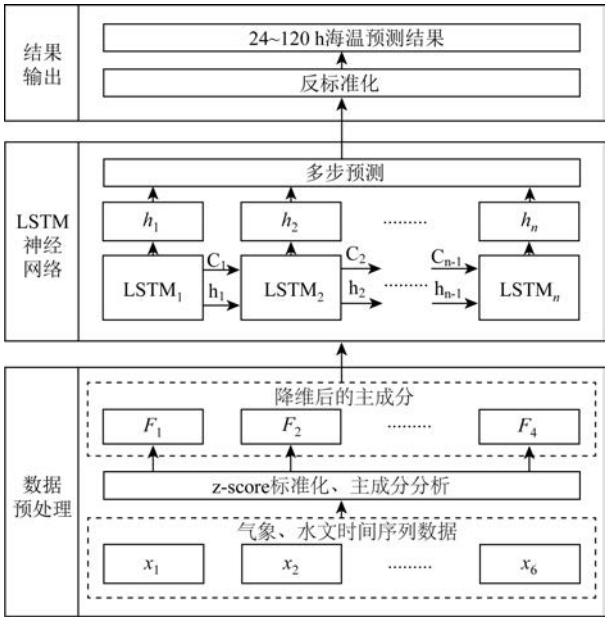


图3 PCA-LSTM海温预测模型整体框架

Fig.3 The overall framework of the PCA-LSTM SST forecasting model

2 实验和结果分析

2.1 数据预处理

本文计算了2019年1月1日—2021年9月30日荣成、海阳两站$T2_{(00\,h)}$、$Q2_{(00\,h)}$、$U10_{(00\,h)}$、$V10_{(00\,h)}$、$SLP_{(00\,h)}$、$SST_{(-24\,h)}$、$SST_{(00\,h)}$的皮尔森(Pearson)相关系数,从相关系数热力图中可以明显看出以上变量间存在明显的相关性(见图4)。荣成、海阳的$SST_{(00\,h)}$与$T2_{(00\,h)}$、$Q2_{(00\,h)}$、$SST_{(-24\,h)}$ 3个要素之间的相关系数分别达到了0.93、0.85、1.00和0.96、0.91、1.00的强相关水平,同时在显著性检验中所得到的p值<0.01。

对荣成、海阳两站的$T2_{(00\,h)}$、$Q2_{(00\,h)}$、$U10_{(00\,h)}$、$V10_{(00\,h)}$、$SLP_{(00\,h)}$、$SST_{(-24\,h)}$ 6个气象水文原始变量进行KMO(Kaiser-Meyer-Olkin)检验和Bartlett球形度检验。KMO检验的结果可以反映变量间的相关程度,取值在0~1之间,结果越接近1,说明变量间的相关性越强;当KMO统计量在0.5以上时说明原始变量适合进行PCA分析,该值越大,分析效果越好。Bartlett球形度检验可以反映数据的分布以及各个变量是否彼此独立,若各原始变量间彼此独立,则无法从中提取公因子,也就无法进行主成分分析;当Bartlett球形度检验的显著性p值<0.05时,说明原始变量呈球形分布,各个变量在一定程度上相互独立。

荣成、海阳两站气象水文原始变量的KMO检验结果分别为0.68、0.69,均大于0.5。Bartlett球度检验结果中的显著性p值均为0.00,小于0.05,说明上述原始变量间存在相关性,适合进行PCA分析。

将原始变量的前80%作为原始训练变量用于建立模型训练集和数据处理,后20%用于建立测试集。为避免测试集的信息泄露,进行数据预处理时使用的均值、方差、主成分系数等均仅由原始训练变量求得。对原始训练变量进行z-score标准化处理后,通过计算协方差矩阵和特征值,可以得到训练集中的各成分并进行主成分选取,计算结果见表1。

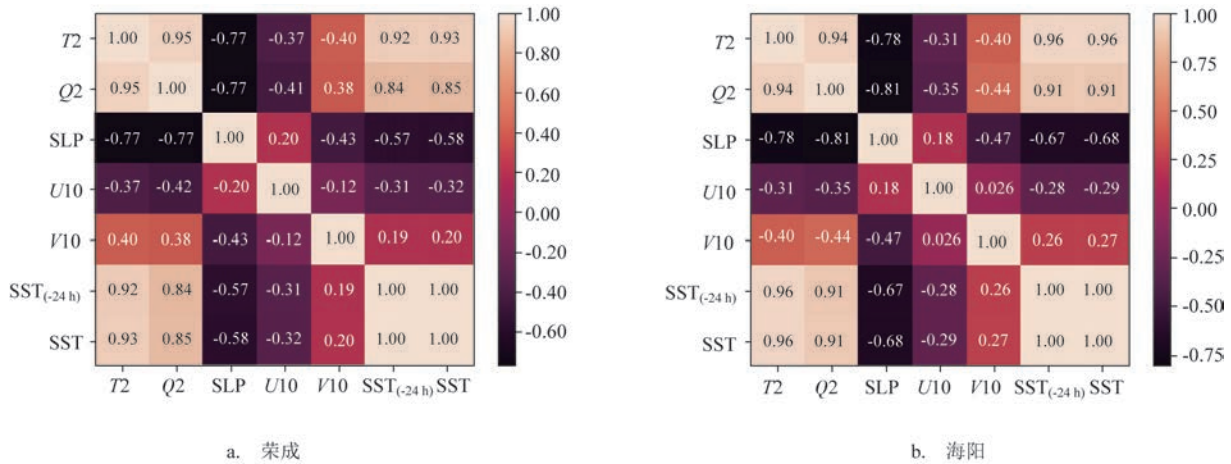


图4 相关系数热力图

Fig.4 Heat map of the correlation coefficient

表 1 各成分的方差贡献率

Tab.1 The variance contribution rate of each component

站名	成分	方差	方差贡献率/%	累计方差贡献率/%
荣成	F_1	3.75	62.5	62.5
	F_2	0.95	15.9	78.3
	F_3	0.81	13.5	91.8
	F_4	0.39	6.4	98.3
	F_5	0.09	1.4	99.7
	F_6	0.02	0.3	100.0
海阳	F_1	3.85	64.1	64.1
	F_2	1.04	17.3	81.4
	F_3	0.72	11.9	93.3
	F_4	0.32	5.3	98.6
	F_5	0.07	1.1	99.6
	F_6	0.02	0.4	100.0

从表1可以看到,荣成、海阳两站成分中的F_1、F_2、F_3、F_4的方差贡献率分别为62.5%、15.9%、13.5%、6.4%和64.1%、17.3%、11.9%、5.3%,累计方差贡献率达到了98.3%和98.6%。满足了上文所述的主成分选取原则,即单个成分方差贡献率大于5%、累计方差贡献率超过95%。因此本文选择F_1、F_2、F_3、F_4作为建立模型训练集的主成分,将6维的气象水文原始变量降低到了4维。

表2为成分矩阵,反映了各主成分和原始变量间的相关系数,能够体现主成分与各气象水文变量之间的相关程度。例如,方差贡献率最高的主成分F_1与主要载荷$T2_{(00\,h)}$、$Q2_{(00\,h)}$、$SST_{(-24\,h)}$呈正相关,而与$U10_{(00\,h)}$则呈负相关,即F_1主要反映海气状态,当气温升高、比湿增大时,径向风降低利于海温的升高,过去时刻海温较高则下一时刻海温也会相对较高;主成分F_2与$V10_{(00\,h)}$、$SLP_{(00\,h)}$呈负相关,反映了这两个要素对于海温的负面影响;主成分F_3与$V10_{(00\,h)}$、$SST_{(-24\,h)}$呈正相关,与$SLP_{(00\,h)}$呈负相关,说明当海温和纬向风都较高时,海平面气压的增加会导致海温降低;主成分F_4主要反映了$U10_{(00\,h)}$、$SLP_{(00\,h)}$、$SST_{(-24\,h)}$都较低时对海温的影响。

基于以上对荣成、海阳两站数据分析结果建立的模型训练集和测试集,时间步长均为15 d,包含的要素为主成分F_1、F_2、F_3、F_4。利用模型进行预报模拟时,若起报时间为T d,此时输入模型的每个样本为$T-14$～T d的各主成分,训练标签为标准化后的$T+1$～$T+5$ d(即预报时效为24～120 h)的SST。

表 2 成分矩阵

Tab.2 Component matrix

站名	成分原始变量	$T2_{(00\,h)}$	$Q2_{(00\,h)}$	$U10_{(00\,h)}$	$V10_{(00\,h)}$	$SLP_{(00\,h)}$	$SST_{(-24\,h)}$
荣成	F_1	0.505	0.498	−0.432	−0.221	0.247	0.447
	F_2	0.032	0.049	0.236	−0.641	−0.698	0.207
	F_3	−0.147	−0.083	0.064	−0.722	0.568	−0.351
	F_4	−0.164	0.067	−0.779	−0.075	−0.342	−0.489
	F_5	−0.101	0.834	0.351	0.112	−0.047	−0.396
	F_6	−0.828	0.206	−0.155	−0.030	0.097	0.488
海阳	F_1	0.495	0.497	−0.442	−0.176	0.261	0.462
	F_2	0.039	0.025	0.157	−0.773	−0.600	0.125
	F_3	0.176	0.059	−0.001	0.602	−0.686	0.365
	F_4	−0.208	−0.001	−0.835	−0.021	−0.293	−0.417
	F_5	−0.380	0.860	0.231	0.077	−0.075	−0.225
	F_6	−0.731	−0.095	−0.173	−0.050	0.102	0.643

2.2 训练结果分析

本文利用PyTorch深度学习框架构建了基于LSTM神经网络的LSTM海温预报模型、PCA-LSTM海温预报模型和用于对比检验的基于BP神经网络的BP海温预报模型。LSTM预报模型与PCA-LSTM海温预报模型的训练参数相同,其输入数据为未通过PCA法、仅通过z-score标准化方法标准化后的气象水文原始变量。BP神经网络是一种利用误差反向传播算法训练的多层前馈神经网络,是应用最广泛的神经网络模型之一。本文构建的BP神经网络为由一个输入层、两个隐含层和一个输出层组成的双隐层BP神经网络,输入数据与PCA-LSTM海温预报模型相同。各模型的具体训练参数见表3。

表3 各模型的训练参数

Tab.3 Training parameters for each model

模型	参数	值
PCA-LSTM/LSTM	隐藏层数	1
	隐藏层节点数	100
	初始学习率	0.001
	优化器	AdamW
	迭代次数	300
BP	隐藏层节点数	8
	初始学习率	0.01
	随机失活率	0.3
	优化器	SGD
	迭代次数	300

2.2.1 基线模型对比结果

为了证明PCA-LSTM海温预报模型的有效性,本文将该模型与基于标准化后的原始变量建立的LSTM海温预报模型和BP海温预报基线模型进行比较,时间范围为2021年3月13日—9月25日,结果见表4。从表中可以看到,PCA-LSTM和LSTM模型的预报效果明显优于BP模型。在海阳海域,PCA-LSTM模型和LSTM模型的预报效果差别较小;而在荣成海域,PCA-LSTM模型在48 h及以上预报时效中的预报结果更好,当预报时效为120 h时,PCA-LSTM模型的RMSE比LSTM模型最多减少了15.4%。

表4 与基线模型的RMSE对比结果(单位:℃)

Tab.4 Comparison result of RMSE with baseline models (unit:℃)

站点	预报方法	预测时效				
		24 h	48 h	72 h	96 h	120 h
荣成	PCA-LSTM	0.65	0.65	0.67	0.67	0.66
	LSTM	0.61	0.66	0.72	0.76	0.78
	BP	0.75	0.78	0.82	0.85	0.87
海阳	PCA-LSTM	0.42	0.46	0.49	0.52	0.54
	LSTM	0.40	0.45	0.49	0.52	0.55
	BP	0.55	0.58	0.63	0.66	0.70

2.2.2 与现有预报方法进行对比

基于荣成、海阳两站的测试集数据,对训练好的PCA-LSTM海温预报模型进行预报测试,并与经验预报、数值预报的预报结果进行对比,结果见表5和图5。利用中国近岸海域基础预报单元海温预报指导产品获得荣成北和海阳近岸海域的日均SST预报结果并作为数值预报结果,该产品采用偏差订正的海温预报释用方法,利用多个海洋站和近海浮标观测结果对区域海洋模式(Regional Ocean Modeling System,ROMS)的数据进行偏差订正,预报时效为24～120 h,使用的观测资料不包含荣成、海阳两站的自建浮标SST观测数据。经验预报方法为对数值预报结果进行人工订正,预报时效为24～72 h。

从结果可以看到,该模型在荣成、海阳两站的海温预报应用上取得良好效果,能够较好地反映出两地SST的变化趋势并做出预报。相比经验预报和数值预报,不同时段PCA-LSTM模型预报结果的平均误差(Mean Error,ME)、平均绝对误差(Mean Absolute Error,MAE)和RMSE均有较大幅度的减小,荣成、海阳两站24～120 h SST预报结果的MAE、RMSE分别达到了0.49～0.53 ℃、0.31～0.41 ℃和0.65～0.67 ℃、0.42～0.54 ℃。该模型在海阳海域的预报效果明显好于荣成海域,这是由于荣成海域在2021年7月末受到了第6号强台风“烟花”的影响,日均SST出现较大波动,而训练数据中并不包含此类极端天气状况,此时该模型的预报结

果的日变化较小,因而与真实结果产生较大误差。同时期海温数值预报和基于数值预报的经验预报方法也存在着较大的误差,很可能也是受该台风影响所致。在海阳海域,数值预报和经验预报方法出现较大的系统误差,这可能是因为数值预报在偏差订正时使用的观测资料距离该站点较远导致的。

表5　各预报方法的精度对比(单位:℃)

Tab.5　Accuracy comparison of each forecast method(unit:℃)

站点	预报方法	参数	预报时效/h				
			24 h	48 h	72 h	96 h	120 h
荣成	PCA-LSTM	ME	0.13	0.12	0.11	0.07	0.07
		MAE	0.49	0.50	0.52	0.53	0.51
		RMSE	0.65	0.65	0.67	0.67	0.66
	近岸基础单元	ME	−0.79	−0.82	−0.84	−0.85	−0.85
		MAE	1.15	1.17	1.22	1.23	1.24
		RMSE	1.48	1.53	1.59	1.61	1.62
	经验预报	ME	−0.58	−0.58	−0.57	—	—
		MAE	1.13	1.16	1.21	—	—
		RMSE	1.54	1.58	1.66	—	—
海阳	PCA-LSTM	ME	0.05	0.02	0.01	−0.02	−0.02
		MAE	0.31	0.34	0.36	0.39	0.41
		RMSE	0.42	0.46	0.49	0.52	0.54
	近岸基础单元	ME	−1.65	−1.68	−1.70	−1.70	−1.69
		MAE	1.67	1.70	1.72	1.73	1.71
		RMSE	1.83	1.88	1.92	1.93	1.92
	经验预报	ME	−1.69	−1.71	−1.72	—	—
		MAE	1.71	1.74	1.75	—	—
		RMSE	1.89	1.93	1.95	—	—

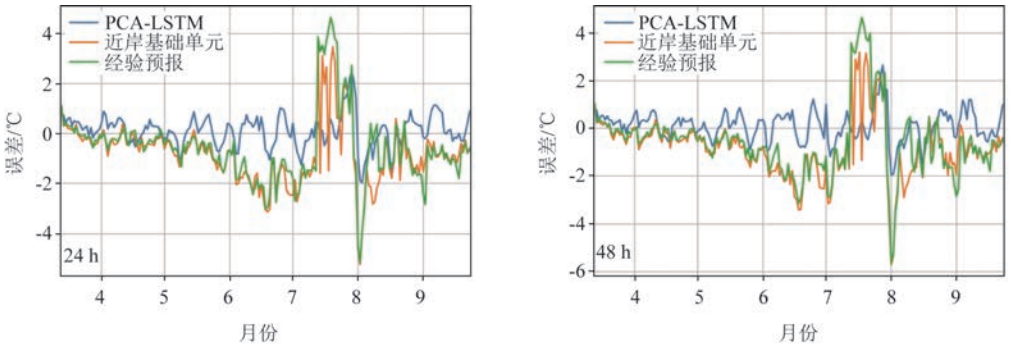
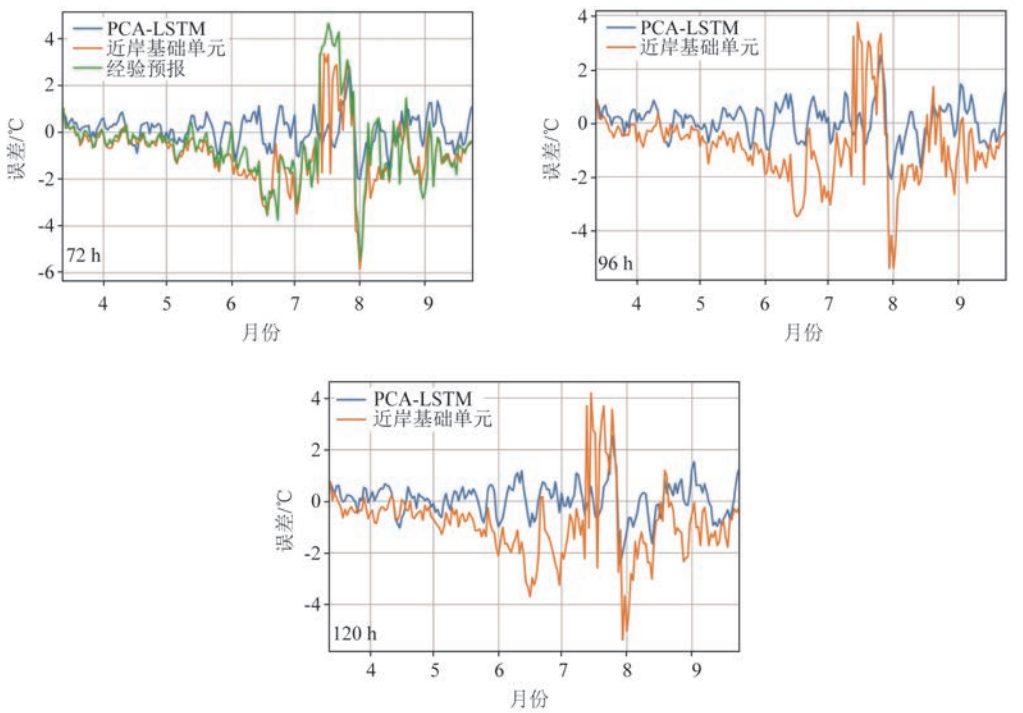


图5　与现有预报方法对比结果图

Fig.5　Comparison results diagram with the existing forecast methods



a. 荣成

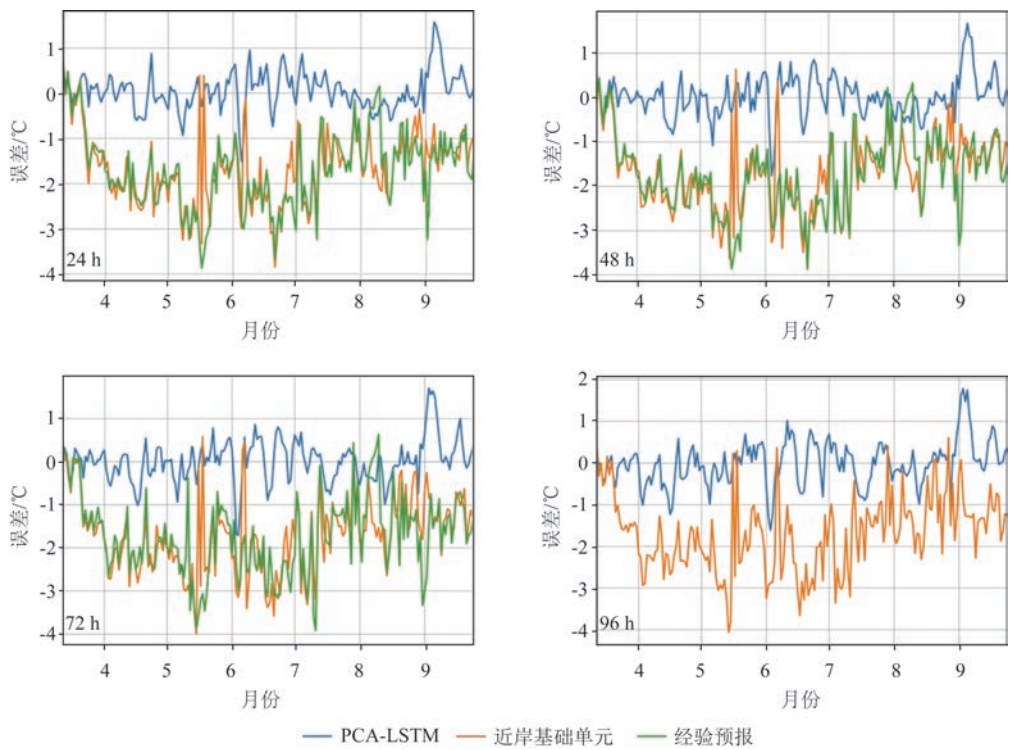
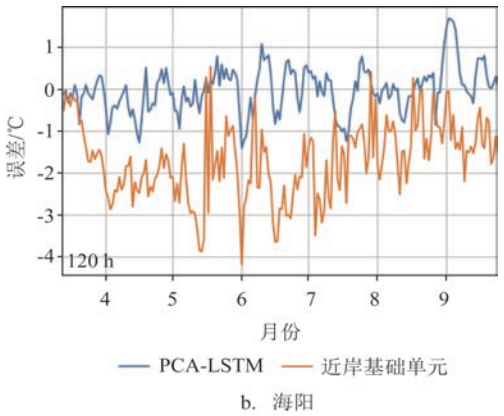


图5 （续）

Fig.5 （Continued）



b. 海阳

图5 （续）

Fig.5 （Continued）

3 结论

本文基于荣成、海阳两站浮标的SST观测数据和WRF模式的气象数值预报数据，利用PCA法进行数据预处理，结合LSTM神经网络构建了适用于单站海温预报的PCA-LSTM海温预报模型，并与基线模型和现有预报方法进行对比检验。结论如下：

①基于LSTM神经网络建立的海温预报模型的预报效果优于BP神经网络。使用PCA法可以在尽可能保留原始变量有效信息的同时，减少神经网络的输入变量数，提高预测精度。

②本文构建的PCA-LSTM海温预报模型在荣成、海阳两站的海温预报应用中取得了较好的效果，相比经验预报和数值预报，预报准确度有了较大幅度的提升，荣成、海阳两站24～120 h时效的MAE、RMSE分别达到了0.49～0.53℃、0.31～0.41 ℃和0.65～0.67 ℃、0.42～0.54 ℃，该模型能够较准确地对SST进行预测。

③由于总样本数较少，训练集中并不包含诸如台风等极端天气下的SST数据，因此该模型对受到台风“烟花”影响的荣成站的预报结果和海阳站有一定差距。在未来工作中，可以对该模型进行改进，通过增加训练样本的数量、在预报序列中加入未来时刻的气象预报数据，或使用非线性的数据分解方法对各要素进行处理等，使其能够更加精确地预测各种状况下的SST。

参考文献：

[1] WEBSTER P J, CLAYSON C A, CURRY J A. Clouds, radiation, and the diurnal cycle of sea surface temperature in the tropical western Pacific[J]. Journal of Climate, 1996, 9(8): 1712-1730.

[2] CASEY K S, CORNILLON P. Global and regional sea surface temperature trends[J]. Journal of Climate, 2001, 14(18): 3801-3818.

[3] 刘伯胜, 雷家煜. 水声学原理[M]. 哈尔滨: 哈尔滨工程大学出版社, 1993.
LIU B S, LEI J Y. Principles of underwater acoustics[M]. Harbin: Harbin Engineering University Press, 1993.

[4] 陈新军. 渔业资源与渔场学[M]. 北京: 海洋出版社, 2004: 289-306.
CHEN X J. Fishery resources and fishery field studies[M]. Beijing: China Ocean Press, 2004: 289-306.

[5] 张建华. 海温预报知识讲座: 第一讲 海水温度预报概况[J]. 海洋预报, 2003, 20(4): 81-85.
ZHANG J H. Sea temperature forecast knowledge lecture: the first lecture seawater temperature forecast[J]. Marine Forecasts, 2003, 20(4): 81-85.

[6] 李燕, 张建华, 刘钦政, 等. 单站海温短期预报自动化[J]. 海洋预报, 2007, 24(4): 33-41.
LI Y, ZHANG J H, LIU Q Z, et al. The automation of single sea station's surface sea temperature short term forecasting[J]. Marine Forecasts, 2007, 24(4): 33-41.

[7] 刘娜, 王辉, 凌铁军, 等. 全球业务化海洋预报进展与展望[J]. 地球科学进展, 2018, 33(2): 131-140.
LIU N, WANG H, LING T J, et al. Review and prospect of global operational ocean forecasting[J]. Advances in Earth Science, 2018, 33(2): 131-140.

[8] 匡晓迪, 王兆毅, 张苗茵, 等. 基于BP神经网络方法的近岸数值海温预报释用技术[J]. 海洋与湖沼, 2016, 47(6): 1107-1115.
KUANG X D, WANG Z Y, ZHANG M Y, et al. An interpretation scheme of numerical near-shore sea-water temperature forecast based on BPNN [J]. Oceanologia et Limnologia Sinica, 2016, 47 (6): 1107-1115.

[9] 王兆毅, 李云, 王旭. 中国近岸海域基础预报单元海温预报指导产品研制[J]. 海洋预报, 2020, 37(4): 59-65.
WANG Z Y, LI Y, WANG X. Development of forecast guidance product for sea temperature of basic forecast units in the Chinese coastal waters[J]. Marine Forecasts, 2020, 37(4): 59-65.

[10] 李启华, 吉海鹏, 张高. 基于神经网络的港口潮汐预报研究[J]. 广州航海高等专科学校学报, 2007, 15(1): 1-4.
LI Q H, JI H P, ZHANG G. Study on tidal prediction of harbor bared on neural network[J]. Journal of Guangzhou Maritime College, 2007, 15(1): 1-4.

[11] 秦思远, 李进军, 龙冰心, 等. 基于GPOS-BP神经网络模型的潮汐预报[J]. 海洋信息, 2020, 35(2): 1-5.
QIN S Y, LI J J, LONG B X, et al. Tide forecast model based on GPOS-BP neural network[J]. Marine Information, 2020, 35(2): 1-5.

[12] KUMAR N K, SAVITHA R, AL MAMUN A. Regional ocean wave height prediction using sequential learning neural networks [J]. Ocean Engineering, 2017, 129: 605-612.

[13] 朱智慧, 曹庆, 徐杰. 神经网络方法在上海沿海海浪预报中的应用[J]. 海洋预报, 2018, 35(5): 25-33.
ZHU Z H, CAO Q, XU J. Application of neural networks to wave prediction in coastal areas of Shanghai[J]. Marine Forecasts, 2018, 35(5): 25-33.

[14] 朱浩朋, 伍玉梅, 唐峰华, 等. 采用卷积神经网络构建西北太平洋柔鱼渔场预报模型[J]. 农业工程学报, 2020, 36(24): 153-160.
ZHU H P, WU Y M, TANG F H, et al. Construction of fishing ground forecast model of Ommastrephes bartramii using convolutional neural network in the Northwest Pacific[J]. Transactions of the Chinese Society of Agricultural Engineering, 2020, 36(24): 153-160.

[15] WU A M, HSIEH W W, TANG B Y. Neural network forecasts of the tropical Pacific sea surface temperatures[J]. Neural Networks, 2006, 19(2): 145-154.

[16] JIA X Y, JI Q Y, HAN L, et al. Prediction of sea surface temperature in the East China Sea based on LSTM neural network [J]. Remote Sensing, 2022, 14(14): 3300.

[17] 贺琪, 查钺, 宋巍, 等. 基于STL的海表面温度预测算法[J]. 海洋环境科学, 2020, 39(6): 918-925.
HE Q, ZHA C, SONG W, et al. Sea surface temperature prediction algorithm based on STL model[J]. Marine Environmental Science, 2020, 39(6): 918-925.

[18] 李艳双, 曾珍香, 张闽, 等. 主成分分析法在多指标综合评价方法中的应用[J]. 河北工业大学学报, 1999, 28(1): 94-97.
LI Y S, ZENG Z X, ZHANG M, et al. Application of primary component analysis in the methods of comprehensive evaluation for many indexes[J]. Journal of Hebei University of Technology, 1999, 28(1): 94-97.

[19] SCHMIDHUBER J. Deep learning in neural networks: an overview[J]. Neural Networks, 2015, 61: 85-117.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

SST forecasting model based on principal component analysis and LSTM neural network

LI Jingshi[1,2], KUANG Xiaodi[1,2], LI Qiong[3], HE Enye[1,2], ZHANG Yubai[3], YUAN Chengyi[4], ZHANG Yanlin[5]

(1. National Marine Environmental Forecasting Center, Beijing 100086, China; 2. Key Laboratory of Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Ministry of Natural Resources, Beijing 100081, China; 3. Shandong Marine Forecast and Hazard Mitigation Service, Qingdao 266104, China; 4. Tianjin University of Science and Technology, Tianjin 300222, China; 5. Liaoning Natural Resources Affairs Service Center, Shenyang 110033, China)

Abstract: Using the sea temperature observation data of buoys at Rongcheng and Haiyang marine stations and the numerical forecast meteorology data of the regional atmospheric model Weather Research and Forecasting (WRF), and based on the Principal Component Analysis (PCA) and Long Short-Term Memory (LSTM) neural network, a PCA-LSTM sea temperature forecasting model suitable for the Sea Surface Temperature (SST) forecasting is proposed in this paper. This model can provide SST forecast for the following 24～120 hours, and its forecasting accuracy is significantly improved compared with the numerical model and statistical model.

Key words: principal component analysis; Long Short-Term Memory neural network; SST forecast