基于遗传算法-长短期记忆神经网络的月降水量预测研究

张星 关悦 党鑫鑫

指导老师：吴建生

广西科技大学

目录

表目录

图目录

摘要

桂林市因其特殊的地理位置和自然环境，导致旱涝灾害频繁发生。准确及时的降水预报对于防洪抗旱具有非常重要的意义，同时对减轻旱涝灾害带来的经济损失、保护当地人民生命财产安全具有重要意义。本文利用桂林市 1951-2019 年的月降水量数据，建立长短期记忆神经网络（LSTM）模型和基于遗传算法的长短期记忆神经网络（GA-LSTM）模型，对桂林市 6 月、7 月、8 月的降水量进行训练。具体的研究进程如下：

第一章中，首先介绍了以桂林市降水预报作为研究对象的原因，以及研究的背景意义等；其次，对遗传算法（GA）、长短期记忆神经网络（LSTM）以及降水的研究进程进行了论述；最后，交代了本文研究问题的思路。

第二章中，依次给出了遗传算法（GA）、长短期神经网络（LSTM）算法的定义，对遗传算法、长短期神经网络的相关理论进行了详细的介绍，同时，为提高预测精度，考虑到遗传算法的空间探索的全局性，用其优化长短期记忆神经网络模型的参数，并给出优化后模型详细的操作流程。

第三章中，为使建立的模型很好的学习数据中的规律，首先对桂林的降水数据进行预处理，将数据分为训练数据集和测试数据集；其次建立长短期记忆神经网络模型（LSTM）并分别对桂林 1951-2019 年 6 月、7 月、8 月的降水数据进行实验研究。

第四章中，为提高预报精度，用遗传算法优化模型参数，用优化过参数的长短期记忆神经网络模型分别对桂林 1951-2019 年桂林 6 月、7 月、8 月的降水数据进行实证分析，并将两种模型的得到的分析结果进行对比，从而对桂林未来三年降水量进行预测。

经过实证研究，基于遗传算法的长短期记忆神经网络模型（GA-LSTM）的拟合效果要优于长短期记忆神经网络模型（LSTM）的拟合结果。因此，本文利用 GA-LSTM 模型对桂林 2020-2022 年 6 月、7 月、8 月的月降水量进行了预测，同时也为桂林防洪抗旱给出了合理的建议。

关键词：遗传算法；长短期记忆神经网络；月降水量预测；防洪抗旱

一、引言

（一）研究背景

水是地球上所有生物生存的必不可少的物质，其珍贵性不言而喻。但随着我国社会经济不断发展，人们开展的活动不断增多，严重消耗着我国水资源，对水资源环境的影响也愈来愈大。气候变化对水资源有最直接的影响，近些年，随着异常气候大规模浮现，对人们的生产生活带来不同程度的损害。众所周知，降水量的大小直接影响该地区人们生产生活的各个方面，降水量过多易出现洪涝，泥石流等灾害，降水量过少会影响人们日常生活用水，易形成土地沙漠化或土地盐泽化，因此要想减轻气候变化带来的经济损失、保护人民的生命财产安全，减少降水量对环境的破坏，提升降水预测能力尤为重要。提高降水量预测的准确性可以减少因降水量的多少带来的不必要损失。对于降水量少的地区，合理利用好每一滴水，将一定量的水资源存储起来，以缓解水资源短缺问题，降低因水资源匮乏带来的经济损失。针对一些降雨量较多的地区，可适当地参考降水预测的结果，进而采取一些有效的防止洪灾的行动。因此，地区降水量的预测尤为重要，关乎到整个地区的生产生活与经济发展。桂林是一座著名的旅游城市，其位于湘桂南端，年平均气温在 19 摄氏度左右，无霜期长达 200 多天。在桂林，每当暴雨侵袭时就极易发生洪涝灾害，给当地的发展以及建设造成不可预估的破坏，准确预测桂林的降水量对当地的经济发展具有非常重要的意义。提高降水预报的及时性和准确性,可以减少自然灾害造成的经济损失,并加强地区水资源管理,对工农业生产具有重要指导意义。

（二）研究意义

地区的降水量对整个地区有着非常重要的影响，掌握地区降水量规律，准确预测未来降水情况，可以加强有关部门防涝抗旱能力，使人们生活得以保障；可

以科学指导工业、农业用水，为该地区的工业、农业生产规划提供理论依据，减少因气候变化带来的不必要损失。本文针对桂林市气象变化，采用遗传算法、长短期记忆神经网络方法，对桂林市过去 69 年 6 月、7 月、8 月的降水数据进行分析，利用预处理后桂林 6、7、8 月的降水数据与遗传算法优化的相关参数，建立基于遗传算法一长短期记忆神经网络的桂林降水预测模型，对未来三年月降水量进行预测，从而对桂林市月降水量的预测有一个积极地指导作用。

（三）研究现状

1.遗传算法的研究现状

遗传算法(Genetic Algorithm, GA)是一种比较流行的对各个参数进行提高准确率的算法，是在大自然生命发展过程的基础上发展而来的，它所具备的特征可以分为三类，分别为随机、全局和并行[1]。有关于遗传算法的起源是来自于计算机对生态系统的模拟，在经过进化论与遗传学说推出了遗传算法的概念。遗传算法首次是被 Holland 提出来的[2]，不过在最近几十年里，遗传算法在各个方面和各个领域都有了很大的发展，它强大的功能主要体现在优化参数、机器学习、数据挖掘等[3]。在一般情况下，使用遗传算法这种方式对最合适的解进行计算一般是在全局搜索中，但是仅仅对一部分内容实施搜索的情况下，就只能得到部分的一个解，且所得到的结论也未必是最好的，对于出现的这种现象，一些学者们便提出了自适应函数，让搜索的范围变为很多个的空间。接着是对各个自适应函数逐渐的进行改进，不断的对这些多个空间进行搜索操作，这种操作就不容易引起所要求的结果是局部解，从而使得遗传算法具备高效率、并行性、全局搜寻这些特殊的性能[4]。在对最优解进行计算时，类似于生物学方面的理论知识，首先要对寻优的参数转变成一个基因编码，目标的染色体会包含这些遗传基因的编码，接着使用类似于生物进化、遗传等类似的途径对其实行交叉、选择、变异等方式

对基因里面的有用信息进行互换，一直循环此操作直到符合研究的目的[5]。

2.长短期记忆神经网络模型的研究现状

长短期记忆神经网络（Long Short-Term Memory，LSTM）是指建立在时间递归基础上的一种神经网络，主要解决训练过程中梯度消失和梯度爆炸的问题[6-8]。其功能包括文字的翻译、机器人的操作、图像的分析、文档摘要等[9]。该模型首先是由 Hochreiter&Schmidhuber 引出的，之后由于众多研究者的探索，从而更加方便使用起来[10]。长短期记忆神经网络版本也有许多类型，包括 GRU。由谷歌的检测可以得到，长短期记忆神经网络比较特别的是 Forget gate，然后是 Input gate，最后是 Output gate[11]。其主要包括三个过程：第一步是忘记阶段，选择性的忘记输入进来的一些不重要的信息，保留剩下的一些重要内容；第二步称为选择性记忆阶段，对输入的一部分内容实行选择性的一个预测；第三步是输出，这个过程重点是找出能够作为现在输出的内容[12]。长短期记忆神经网络能够较好的处理长期依赖的现象，基本特点是能够记下较长时间的内容，对于该模型中这种比较特别的构成，很适合对时间序列中延迟较长的数据进行处理[13]。

3.降水预测研究现状

降水量预测是一项较为复杂的工作，因为预测降水量需要多方面的交叉学科知识。在进行降雨数据采集时，由于使用方法的不同，会对预测结果产生一定影响，导致预测精度低。因此，提高预报降雨量数值的精确度，是众多科研学者工作的重中之重。

20 世纪，统计学理论发展较快，受到计算机技术影响，很快便受到了一些学者的热爱，这也造成了将统计学的理论问题应用于对天气情况的预测等方面[14]。经过实证研究，这些统计方法开始对降水量进行计算和预测，解决低维、线性等问题，但存在的问题是如果单纯用统计学所学的方法，想要得到高精度的

预测结果是有一定的难度的[15]。对于出现的这种问题，众多学者提出了不同的预测方法，如模糊理论、灰色系统、浑沌系统、机器学习等，对降水预测的研究提出了新的途径[16]。由于对降水进行预测时涉及多个领域的知识，若仅考虑单一体系会存在不足，通过将计算机技术与其他学科联系起来，产生了大量新方法。上个世纪 60 年代，部分学者将统计学方法应用于中国各个区域降水量的预测[17]，并得到了很大的发展。目前，有关降雨量预测已经产生了很多方法，比如时间序列、支持向量机、遗传算法、长短时间记忆神经网络模型等[18]。

真实的降雨量数据具有一定的周期性变化，由于降水量的多少受多种因素的影响，并且这些因素变化不稳定，易使预测结果不准确，为了预防气象灾害带来不必要的损失，并合理利用降水提供的水资源，有必要建立精确度较高的降水预测模型。因此，提高地区降水量预测的及时性与准确性，关乎到整个地区生产、生活与发展，可以减少因降水量带来的灾害以及不必要损失。本文基于遗传算法——长短期记忆神经网络模型对桂林未来降雨量进行预测，提高降水预测的准确率，为桂林降水预测方面提供一定的指导作用。

（四）研究内容

地球表面大部分区域都是海洋，陆地只占地球表面一少部分，淡水资源在地球上对人们和其他生物来说特别珍贵，是所有生物生存所必须依赖的条件，地球上的水长期处于动态平衡状态，地表水的补给主要来源于降水，降雨到地表上的水分，一部分汇集成江河，一部分渗入到地下，还有一部分注入海洋，通过蒸发把水传输到大气层中，以此便形成了生态系统的水资源无限循环过程，降水能够不断补给水循环系统，提供人类得以生存的水资源，保证生态环境稳定平衡发展。

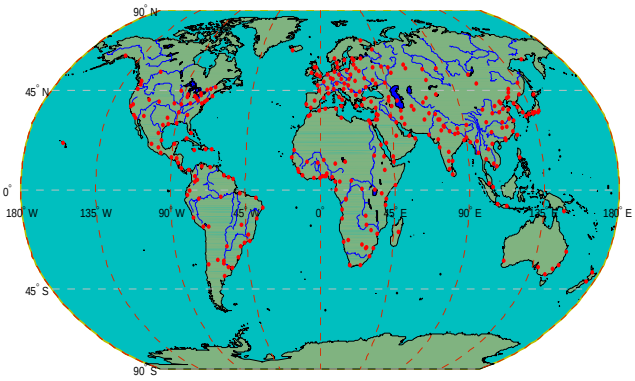

图 1　世界地图

中国位于地球上东半球的北半部，大部分在温带。广西在中国南部位置，南濒北部湾、面向东南亚，是中国唯一一个沿海自治区。广西由于它特殊的地理位置与气候环境，其境内也有众多河流，是中国年降水量比较多的地区之一。广西各地区年降水量均值为 1070mm 以上的[19]，但是绝大部分地区的降水量是在 1500mm 到 2000mm 之间的。受冷热气候的影响，广西一年中有三分之一是下雨的，其降水量占广西全年降水量的 70－85%[20]，降雨量过多极有可能发生洪涝灾害；及时预测降水情况可以给相关部门提供降水信息，使农业、水利等部门可以及时采取相应的防涝抗旱设施，尽可能地减少因气候带来的灾害损失。每年 10 月到次年 3 月降水量偏低，天气极其炎热时由于降水量偏少，地表温度过高，易引发森林火灾等自然灾害。因此，关于地区降水量预测的及时性与准确性，关乎到整个地区生产、生活与发展。

中中中中中中

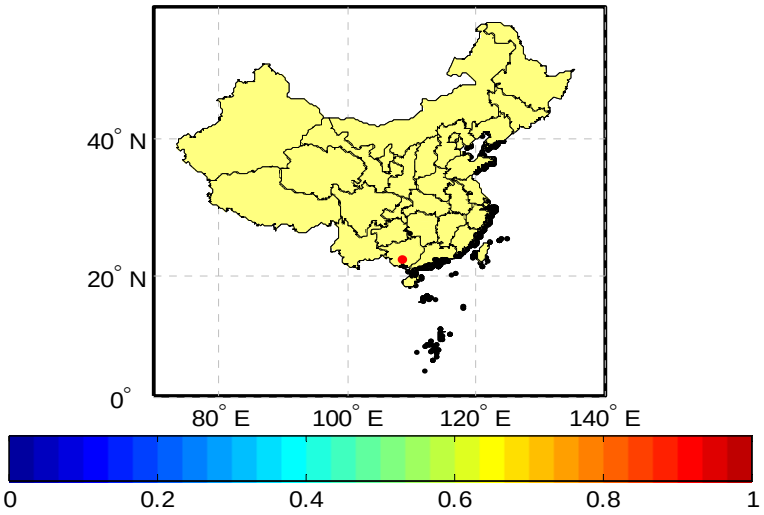

图 2　中国地图

本文使用基于遗传算法-长短期记忆神经网络模型进行降水预测，并对桂林未来三年的降水量进行预测，为相关部门针对不同自然灾害的发生作出有力的保护措施提供相应的数字依据，减少百姓的经济损失，促进经济稳固发展。研究内容如下:

1. 降水是否受众多物理量因子影响，而这些物理量因子影响程度不同，需要提取影响降水的有效因子，对桂林有关降水方面的数据进行预处理。

2. 模型存在一定不足，本文采用遗传算法进行参数寻优，得到预测降水的最优参数组合。

3. 针对预先处理后的有关降水的一些数据和遗传算法进行优化的有关参数来建立一种基于遗传算法一长短期记忆的有关桂林降水预测的模型，接着对桂林未来几年的降水量进行预测。

本文研究的流程图如图 3 所示：

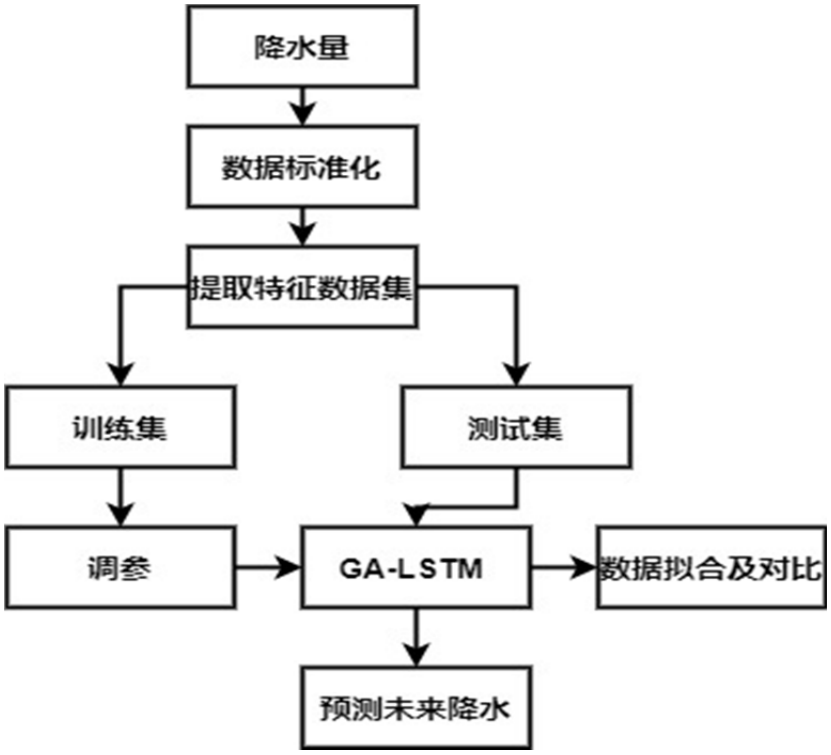

图 3　研究进程流程图

二、模型的理论依据

（一） 遗传算法

1.遗传算法的实现

几十年前，针对我们实际生活中的一些非线性系统的问题，尤其是涉及到人工智能优化的这个层次，传统的一些统计方法并不能很好的解决生活和经济中的某些问题，但是我们又不能不解决当下面临的这些难题。在这个背景下，遗传算法便得到了很快的发展和进步。考虑到遗传算法不仅能独立地存在于各个研究领域当中，还可以和其他常见的方法进行结合，从而进行分析和预测。于是遗传算法便广泛的应用于机器学习的各个方面，通过这种方式来更好并且更有效的解决生活以及社会经济中遇到的一些问题。

本文的遗传算法的实现过程如下：

① 编码与解码。

② 个体的适应度评估。适应度函数不是千篇一律的，不同的情况所对应的函数是不一样的，因此我们要提前制定好选择最佳适用度的规律。一般情况下适应度大的，遗传的机会也会较大。

③ 选择、交叉、变异，来生成下一个子代；

④ 根据现实中的情况来得出所预期的结果，进而来断定这个结果是否符合之前设定的条件，符合条件的话就可输出最适合的解，如果不符合的话就会循环迭代。GA 的流程图见图 4 。

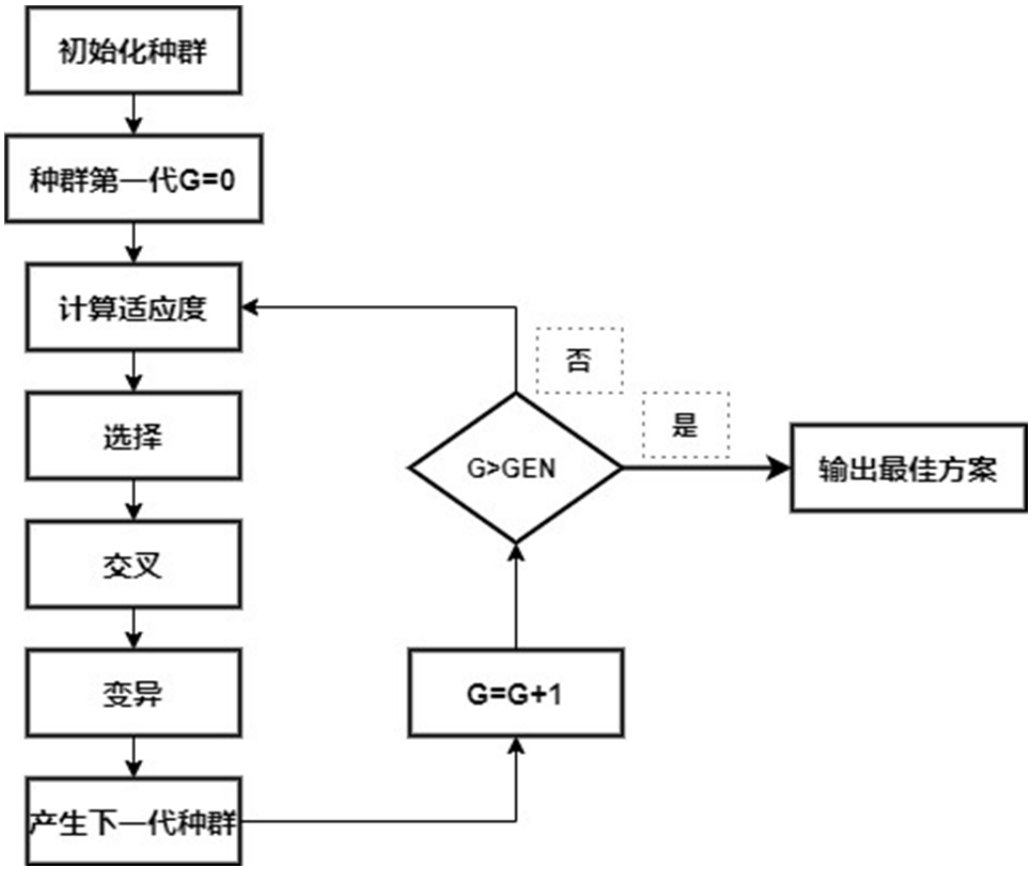

图 4 遗传算法

2.遗传算法对简单函数的优化

与此同时，为更好地表现遗传算法的优化流程，本部分内容中就详细地来介绍一下这个流程。

① 个体编码

遗传算法的概念就是针对一个个体单位的基因对其进行运算的过程。先把x_1和x_2进行一个符号串的编码，比如基因型 $\mathrm{X}=100010$,前三个数字代表x_1，后三位代表x_2，这就构成了个体的基因型，同样它也可用来表示出可行解的功能。另外每一个个体单位的 x 与 X 的表现型是通过编码和解码这个方式来进行相互转化的过程。

② 初始群体的产生

遗传算法它的对象并不是单一的个体，而是一个群体，这就需要事先来设定一些初始群体的点见表 1。另外最初的对象 $P(t)$ 是 $M=4$ 个单位组合而成的，其中个体单元是随机进行产生的。

如：110101，101011，011100，101001

③ 适应度

在本例中，选用函数值的大小来作为每个个体的适应度。

④ 选择运算

选择运算的过程就是一个优胜略汰的过程。具体过程如下：

*Step*1 第一步就是先计算出全部单位的适应度的总计 $F(1,2,\cdots,M)$ ；

*Step*2 接着对每个个体单位的适应度的值进行计算，也即每个个体的函数值的大小 f_i 除以全部个体的函数值的和 F ， $\frac{f_i}{F}$ 指的就是每个亲代传递到下一代子单元中的概率值的大小；

*Step*3 每个概率值组成了一个区域，并且 $p_1+p_2+\cdots+p_m=1$ ；

*Step*4 在 $[0,1]$ 之间随机产生了一些数字，确定它们分别位于哪些概率区域内，进而来确定各个单位被选中的次数。

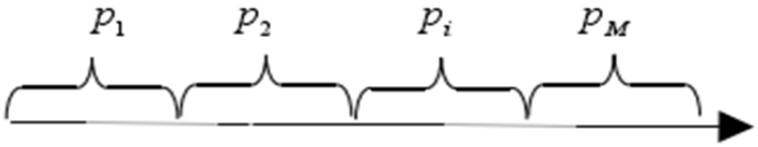

图 5　初始种群概率

表 1　选择算子

个体编号	初始群体	X_1	X_2	适值	占总数百分比%	选择次数	选择结果
1	011101	3	5	77	0.20	1	011101
2	101011	5	3	93	0.25	1	111001
3	011100	3	4	59	0.16	0	101011
4	111001	7	1	147	0.39	2	111001
总计				376	1		

⑤　交叉运算

交叉运算是指利用一个固定的概率值进而来交换其中两个单位之间的一部分的内容。

表 2　交叉运算

个体编号	选择结果	配对情况	交叉点位置	交叉结果
1	011101			011001
2	111001	1-2	1-2：4	111101
3	101011	3-4	3-4：5	101001
4	111001			111011

从表 2 中可以看出，新个体‘111101’、‘111011’和其父代的 p_i 都挺高。

⑥　变异运算

所谓的变异运算，就是指染色体上的基因值在一定的条件下作为催化剂按照一个相对来说比较小的概率来产生一个新个体的过程。详细的操作步骤为：第一步要确定每个单位变异基因的具体位置；然后再根据某一固定的概率针对变异位置的原有基因值取相反值。

表 3　变异运算

个体编号	交叉结果	交叉点	变异结果	子代群体 P（1）
1	011001	4	011101	011101
2	111101	5	111111	111111
3	101001	2	111001	111001
4	111011	6	111010	111010

接着对 $P(t)$ 实行了这样的一轮操作步骤之后就能得到新的下一代 $P(t+1)$。

表 4　新一代群体

编号	子群体	x_1	x_2	适值大小	占总数的百分比%
1	011101	3	5	77	12%
2	111111	7	7	245	39%
3	111001	7	1	149	24%
4	111010	7	2	155	25%
总计				626	1

根据表 4 中总结出的结果，进化后的对象的 p_i 的最大值、平均值有较大进步，最优解为 $x=[7,7]$。

（二）长短期记忆神经网络结构

1.长短期记忆神经网络概念

传统的人工神经网络（Artificial Neural Network，ANN）有着无法利用时序信息等之类的问题。在气象领域这个问题显得很明显，因为气象数据包含着特别多的时序信息问题。为解决这样的问题，循环神经网络（Recurrent Neural Network，RNN）对网络结构进行了一些改进，增加了一个环状结构进而来建立神经元自身的一个连接。这就会让时序信息的前一时刻的信息得以维持，接着对后续的网络信息产生作用。RNN 适合于对时间序列数据进行处理，并且已在很多领域得

以广泛应用, 例自然语言处理、情感分析、视频分类、语音识别和干旱预测等等。

但是虽然 RNN 在时序数据处理方面与简单的 ANN 相比有了一定进步, 但随着在实际应用中对 RNN 更加深入的研究, RNN 也同样暴露出了一些问题。梯度消失问题也就是指随着 RNN 网络按时间迭代的进行，历史时刻的值对网络隐含层的影响逐渐减小一直到近似为 0。

循环神经网络里会发生梯度消失这种情况，所以为了解决此种问题，Hochreiter et al 就列出了一个相对特别的 RNN 的一种情况——长短期记忆神经网络。其中长短期记忆神经网络最擅长处理的是具有较长时间依赖性的时间序列数据。本文利用了长短期记忆神经网络来探索这种方法在降水季节预测这一种长时间序列数据相关问题中的一些应用。长短期记忆神经网络这种方法目前已经在电力负荷、股票价格、PM2.5 浓度、船舶轨迹以及交通流量等等各个领域取得了较好的预测效果。

长短期记忆神经网络是利用改进后的一个记忆块来取代了原本 RNN 隐含层中的一个神经元，这就可以使得误差可以传递在一个较长的时间跨度内。

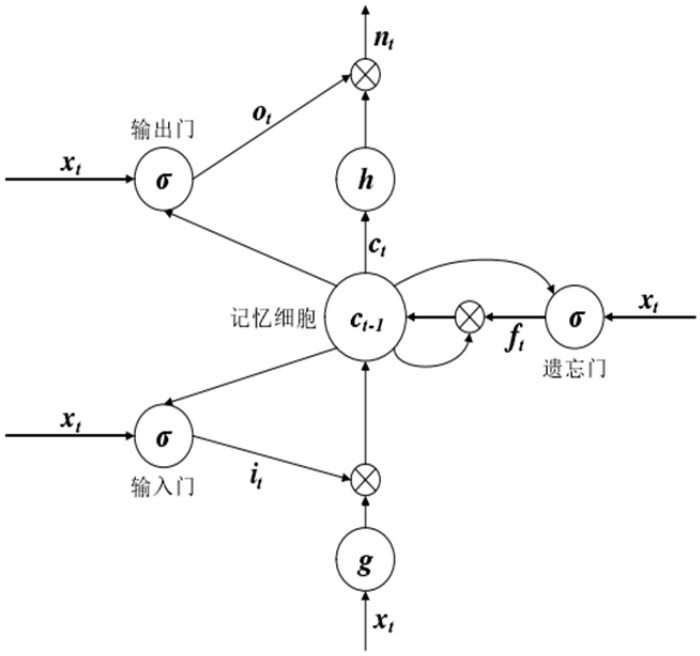

图 6　LSTM 结构图

2.长短期记忆神经网络原理及步骤

下面给出了一个很典型的长短期记忆神经网络的结构图, 按相关文献中所提及的一些方法, 具体计算过程如下所示：

$$i_t = \sigma W_{ix} x_t + W_{ih} h_{t-1} + W_{ic} c_{t-1} + b_i$$

$$f_t = \sigma W_{fx} x_t + W_{fh} h_{t-1} + W_{fc} c_{t-1} + b_f$$

$$D_t = \sigma W_{0x} x_t + W_{oh} h_{t-1} + W_{oc} c_{t-1} + b_o$$

$$c_t = f_t \times c_{t-1} + i_t \times \tanh W_{cx} x_t + W_{ch} h_{t-1} + b_c$$

$$n_t = o_t \times \tanh c_t$$

模型输入表达式为 $x = \{x_0, x_1, x_2, \cdots, x_t\}$， x_t 对应的是 t 时刻长短期记忆神经网络的输入参数； n_t 对应的是 t 时刻长短期记忆神经网络的输出参数； i_t、f_t、o_t、c_t 分别对应的 t 时刻的输入门、遗忘门、输出门、记忆细胞它们所包含的一个输出；W 、b 是指输入层和记忆细胞之间、记忆细胞和输出层之间的连接权值和神经元的偏置值。

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

与此类似, 与 Sigmoid 激活函数十分类似, tanh 函数它也是一种激活函数, 其值域为 $[0, 1]$。

$$\tanh = \frac{\sinh x}{\cosh x} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = 2 \times \sigma(2x) - 1$$

长短期记忆神经网络的训练过程可简化为以下四步：

*Step*1 计算长短期记忆神经网络的输出值；

*Step*2 计算记忆细胞在时间以及向上一层级的网络反方向传播的一个误差项；

*Step*3 基于误差项来逐个计算权重的梯度过程；

*Step*4 特定的选取梯度优化的算法来更新权重系数的矩阵，最终来完成长短期记忆神经网络的计算。

3.GA 优化的 LSTM 降水量模型

GA 一般是指计算机模拟达尔文生物进化论的生物科学研究算法，在 GA 里面有关于种群的遗传演化过程，在其中发现，种群进行遗传的主要载体就是染色体，通过借助多种随机操作，不断地演化出了一种新的解集种群，然后就根据个体的适应度和选择函数的取值来选择最优的种群个体，这就是 GA 中的优化最优解。

本文同样也利用 GA 对 LSTM 网络进行了一个寻优处理的过程，利用了 GA 比较强大的可以进行全局随机搜索的能力，得到 LSTM 网络中神经元个数、学习率和训练次数的最优组合。思路如下：

① 染色体编码

我们把 LSTM 网络中对应的隐藏层神经元的数目、学习率和训练次数分别作为 GA 的初始化对象，利用实数编码的这种形式对染色体所进行的一个过程叫编码。隐藏层神经元的区间大小为$[5, 40]$，学习率的区间大小是$[0.001, 0.1]$，训练次数对应为是$[50, 500]$。

② 适应度函数

适应度函数的选择会直接影响到 GA 优化后网络的性能，进而影响到降水量的提取效果。本文将遗传算法（GA）和长短期记忆神经网络（LSTM）相结合，建立基于 GA-LSTM 模型的对于桂林月降水量分析预测模型，利用遗传算法（GA）对长短期记忆神经网络（LSTM）的参数进行优化，得出学习率、隐层神经元个

数和训练次数的最佳组合，进一步提高模型解决非线性问题的能力。将优化过的参数与 LSTM 的模型相结合，对桂林未来的降水量进行预测。

③ 选择算子、交叉算子变及异算子

选择算子就是在当前种群中选择适应性较好的个体，也就是该个体符合目标所需要的优点作为亲本，并将好的遗传信息传递给子代。该选择算法具有高效的算法执行效率以及易于实现的优点，算法复杂度远远低于其他选择策略并且易于并行化，在进行选择的过程中不容易陷入到某个部分单位的最优点。交叉算子的运行过程用的是洗牌交叉算法，在进行该交叉算法之前，个体的上一代中就运用了一种关于函数洗牌的运算，也就是说如果随机数在$(0,1)$之间产生，并且该随机数的大小小于之前所给的交叉率的大小，这时候就实行交叉变换，也就是变异的部分操作。

4. 基于 GA-LSTM 模型的桂林月降水量分析

本文将 GA 与 LSTM 网络相结合，从而构建基于 GA-LSTM 的桂林月降水量的分析。首先采用 GA 对 LSTM 网络的超参数作寻优处理，得出学习率、隐层神经元数和训练次数的最佳组合，进一步提高模型的非线性映射能力；然后利用寻优的参数组合构建的 GA-LSTM 模型作为桂林月降水量的非线性变换函数；在此基础上应用非线性变换函数预测出桂林月降水量的情况。求得桂林月降水量预测的最好的估计，模型具体操作流程如下所示：

① 选择训练数据集。

② 利用 GA 优化 LSTM 网络参数。

a.将长短期记忆神经网络模型的参数进行优化，得出学习率、隐层神经元个数和训练次数的最佳组合，进一步提高模型解决非线性问题的能力，来执行种群初始化的过程以及染色体的编码与解码的操作过程。

b. 计算初始种群中各个单位适应度值的大小；

c. 对染色体进行选择、交叉和变异的操作；

d. 解码染色体、对种群内个体的适应度进行计算；

e. 不符合这个终止条件的话，那就需要重新返回 c 步；若符合遗传终止条件，则将 GA 求出的最优参数作为 LSTM 网络模型的最终参数；

③ 训练 GA-LSTM。

④ 提取桂林月降水量的预测情况。

（三）本章小结

本章节包括研究中所利用到的算法与其优化。降水季节预测所选取的算法有：遗传算法和长短期记忆神经网络以及被遗传算法优化后的长短期记忆神经网络，长短期记忆神经网络记忆块的特殊结构使得该算法更适合于时间序列预测。本章节中详细介绍了各方法的计算步骤及其优点，而利用不同方法的实际预测结果对比将在后续的章节进行讨论。

三、基于长短期记忆神经网络的桂林降水模型

（一）数据的收集与预处理

本文使用的桂林降水建模数据来自广西气象局。总数据为桂林 1951 年-2019 年 6 月、7 月、8 月总共 204 个。其中 1951 年-2012 年 6 月、7 月、8 月总共 183 个数据作为训练数据集合建立降水拟合模型，2012-2019 年 6 月、7 月、8 月的数据作为测试数据集合优化校验模型，同时，本文将进行 2020-2023 年桂林 6 月、7 月和 8 月降水量的预测，为相关部门针对不同自然灾害的发生作出有力的保护措施提供相应的数字依据，减少百姓的经济损失，促进经济稳固发展。

表 5　建立模型的数据特征

数据集	数据起止时间	样本数
训练数据	1951 年-2012 年 6、7、8 月	183
测试数据	2012 年-2019 年 6、7、8 月	21

桂林 1951 年-2019 年 6 月、7 月、8 月份的降水量如图 7、图 8、图 9 所示。因为每年的 6 月、7 月、8 月降水量比较多，这三个月最容易发生洪涝灾害，导致百姓的农作物减产，或者在本该多雨的三个月，降水量却不尽人意，导致经济遭到重创。从图中也可以看出，从 1951 年-2019 年桂林这三个月的降水趋势基本上可以排除掉白噪声序列，并且可以明显看出每年降水量的数量值，因此该数据在实验中可以直接使用。

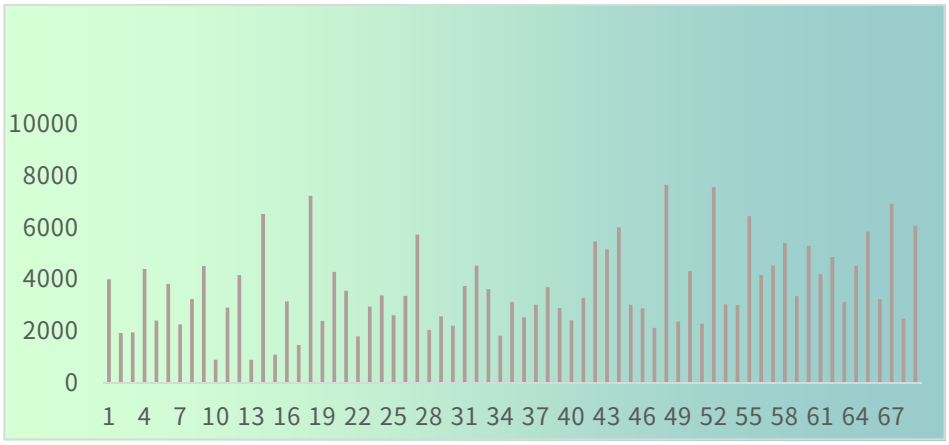

图 7　桂林 6 月份降水量时序图

从图 7 可以看出，69 年来桂林 6 月份降水量最多的时候达到了 7667mm，而最少的时候只有 895mm，二者的极差比较大，很容易造成洪涝或者干旱天气，造成经济作物的损失。

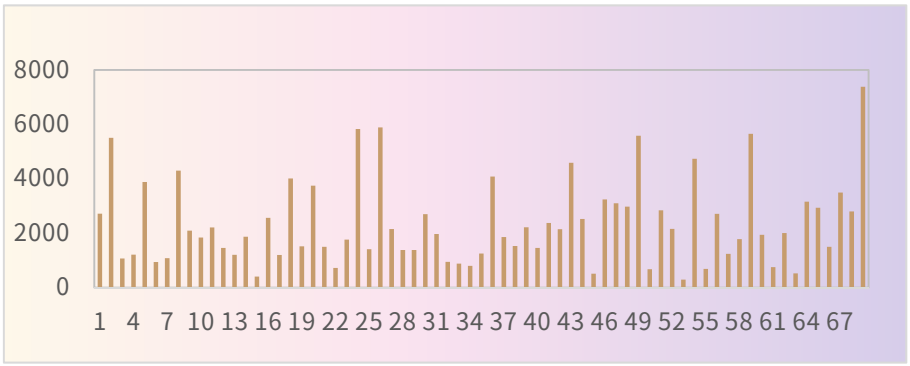

图 8　桂林 7 月份降水量时序图

从图 8 可以看出，7 月份的降水量最多的时候达到了 7381mm，最少的则有 293mm。

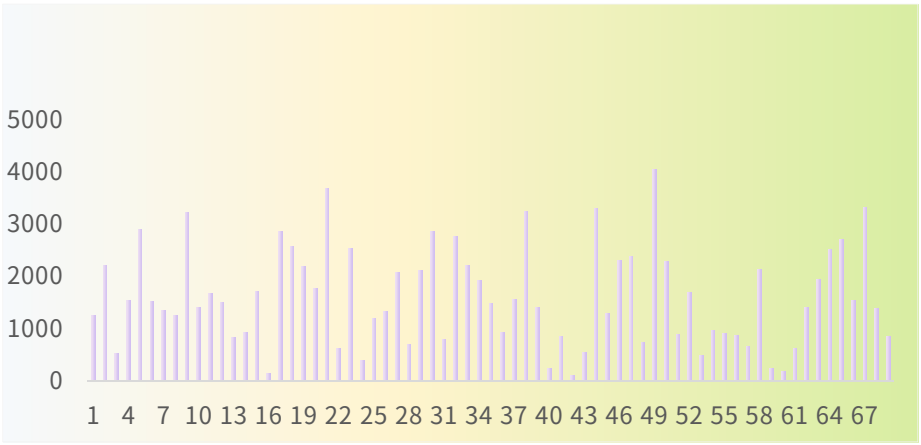

图 9　桂林 8 月份降水量时序图

8 月份降水量最多达到了 4085mm，最少的则为 115mm。可以看出，69 年来桂林的降水量分布极差比较大。

因为降水数据收集的过程中人为因素比较多, 为了在实验中能获得比较好的拟合效果，更能契合实际情况，本文将桂林月降水量数据进行标准化处理，使其在训练当中能更好的表现真实的降水情况。

（二）模型性能评价指标

同时为更加简单明了的观察模型对桂林月降水量数据的拟合的一个效果, 本

文建立了长短期记忆神经网络（LSTM）模型、基于遗传算法的长短期记忆神经网络（GA-LSTM）模型这两种模型。用他们分别对训练数据和测试数据进行拟合，然后用两个模型的对比的结果来检验模型预测的结果。本文主要利用以下几种方法来检验模型拟合程度。

1.均方根误差（Root Mean Square Error，RMSE）

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_t-\hat{x}_t)^2}$$

2.平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）

$$\mathrm{MAPE}=\frac{100}{n}\sum_{t=1}^{n}\frac{|x_t-\hat{x}_t|}{|x_t|}$$

其中，x_t 表示桂林市 4 月份降水的观测值；$\hat{x}_t$ 表示降水拟合值；评价指标 1 和 2 可以衡量降水实际观测值和模拟拟合值之间的偏差程度，其值越小，说明二者之间的偏差越小，拟合的效果就越好。

（三）基于 LSTM 模型的桂林月降水量分析

长短期记忆神经网络（LSTM）模型解决了网络单元以链式方式连接的传统递归神经网络梯度消失和爆炸的问题，有效地提高了模型学习的时间，在处理有关时间序列的预测和非线性映射的问题中，因其具有很强的记忆能力，因此该模型在解决非线性问题上体现出了较好的优势，也因此广泛的应用到各个领域当中，并且经验证在模型训练中取得不错的效果。

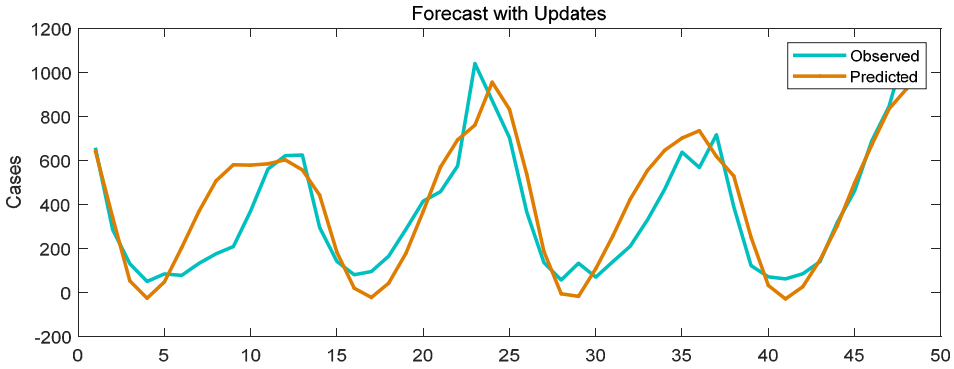

图 10　桂林 6 月份降水的拟合情况

由图 10 可以看出，桂林 6 月份降水量的拟合值和真实值的趋势大致是相同的，其中有拟合效果好的一段，也有拟合误差比较大的一段，拟合效果符合一般实验中数据的拟合情况。

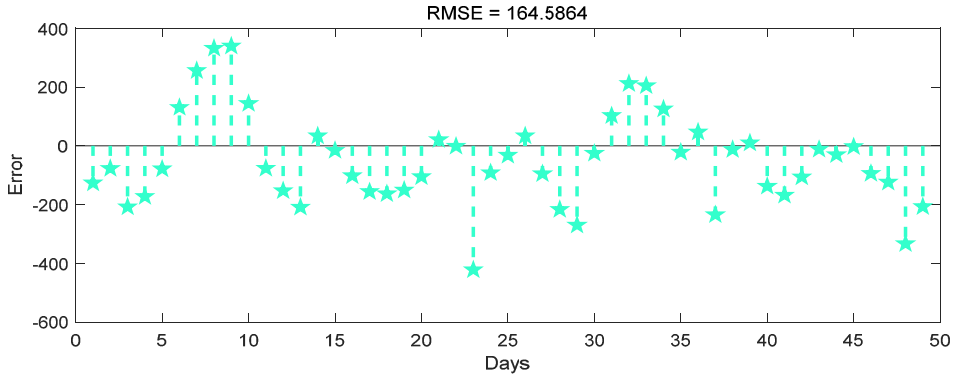

图 11　桂林 6 月份降水拟合残差

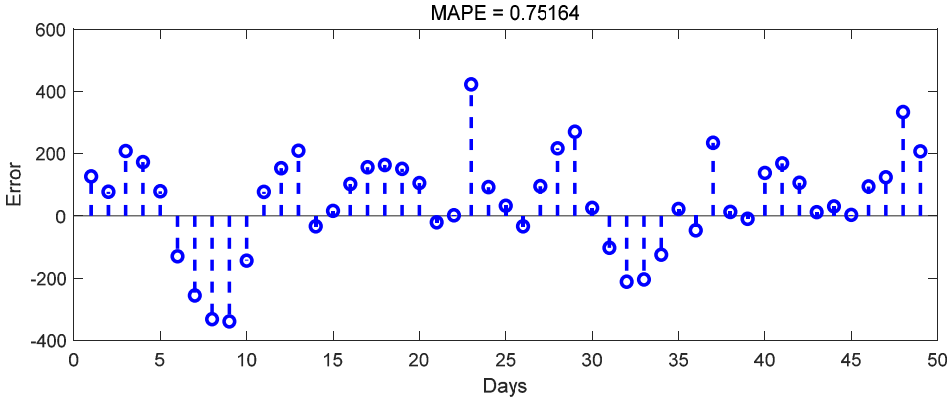

图 12　桂林 6 月份降水拟合误差

从图 11 和图 12 可以看出桂林 6 月份降水拟合的 RMSE 和 MAPE 的值分别为 164.5864 和 0.75164, 从数值结果来看拟合值和真实值之间偏差还是比较大的。

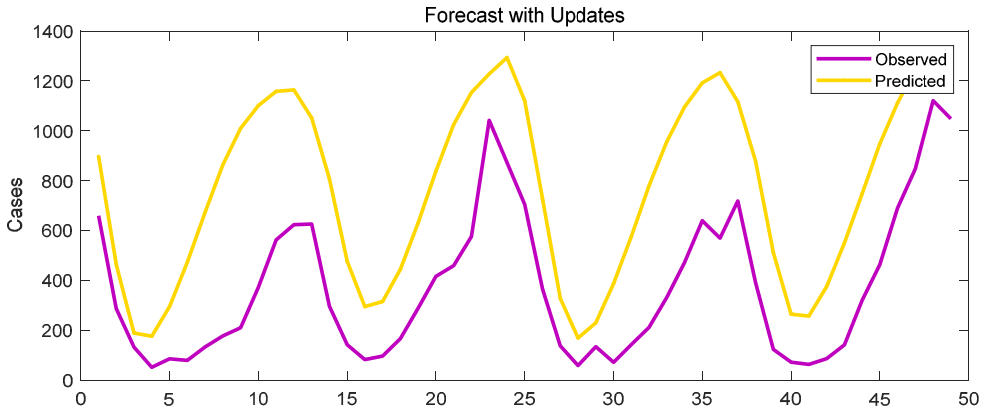

图 13　桂林 7 月份降水量拟合情况

由图 13 可以看出，桂林 7 月份降水量的拟合值和真实值的走势也是大致相同的，并且和 6 月份的走势相像，但是明显可以看出其拟合值和真实值之间的差值还是挺大的。

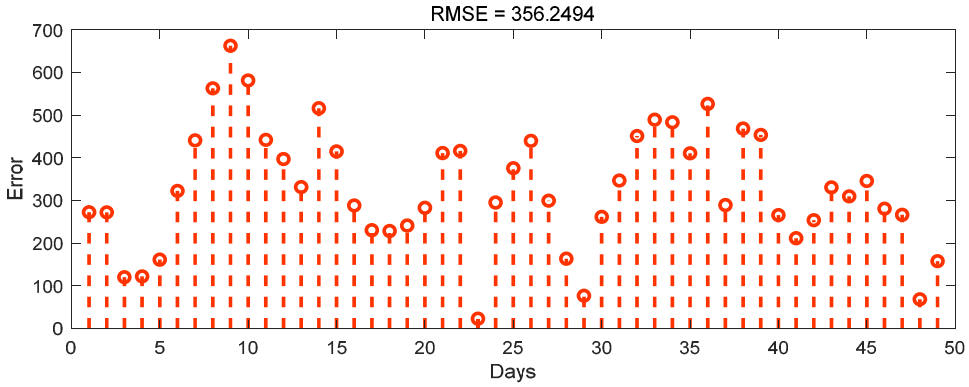

图 14　桂林 7 月份降水拟合残差

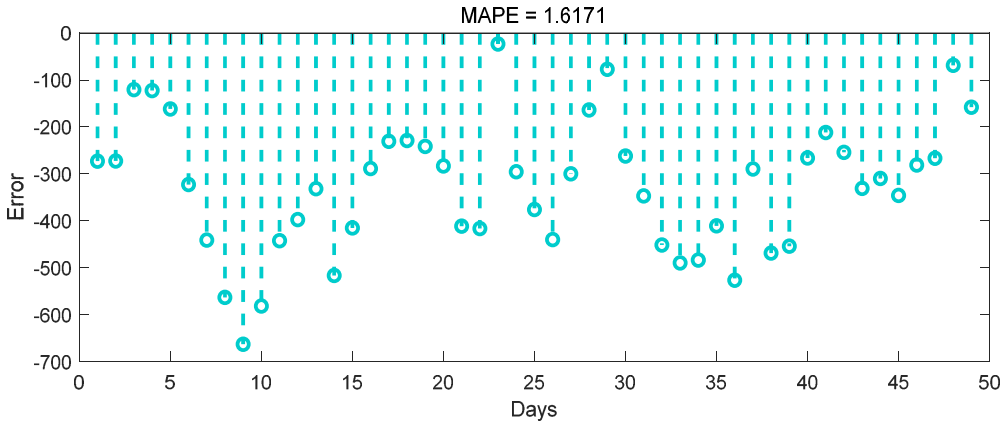

图 15　桂林 7 月份降水拟合误差

可以直接看出该模型在 7 月份的拟合效果误差较大的阶段多，从图 14 和图 15 可以看出桂林 7 月份降水量在 LSTM 模型训练下的均方根误差（RMSE）和平均绝对百分比误差（MAPE）分别为 356.2494 和 1.6171，很明显观测值和拟合值之间的误差要比 6 月份二者之间的误差大, 该模型在桂林 7 月份的训练中效果不是很理想。

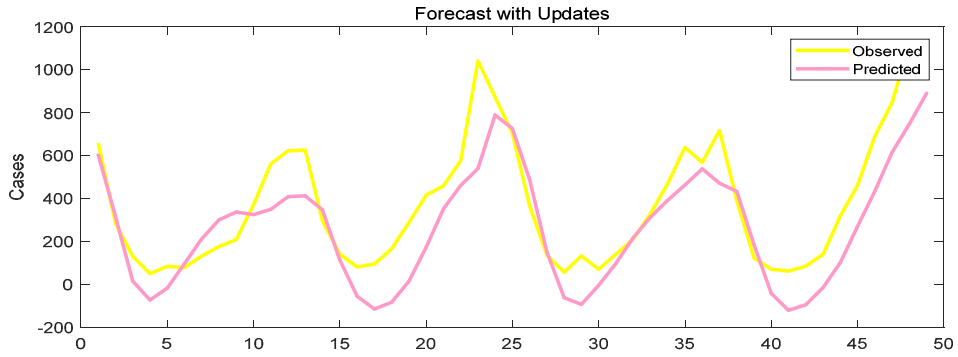

图 16　桂林 8 月份降水量拟合情况

由图 16 可以看出，桂林 8 月份降水量的拟合值和真实值的走势也是大致相同的，并且 6 月、7 月、8 月份这三个月的走势都很相像，也可以看出该模型训练的结果图中有拟合效果好的一段，也有拟合误差比较大的一段，拟合效果符合一般实验中数据的拟合情况。

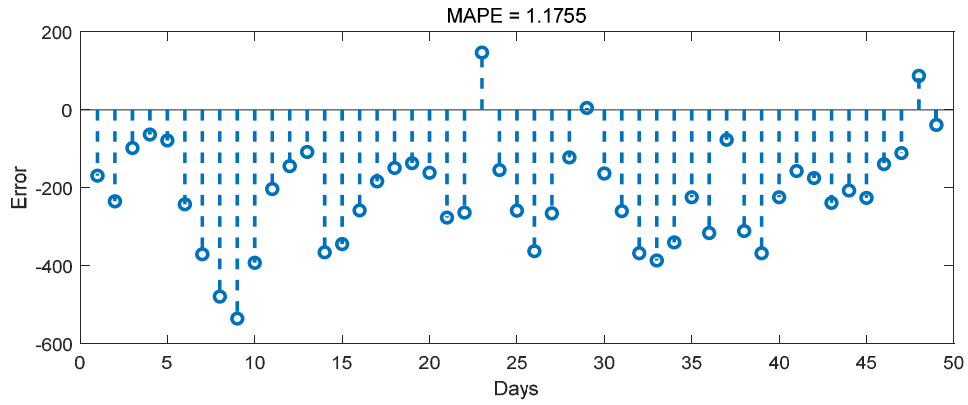

图 17　桂林 8 月份降水量拟合的残差

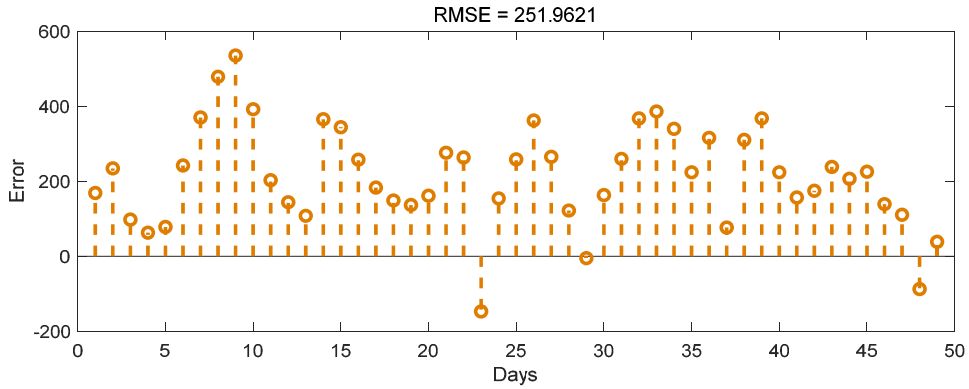

图 18　桂林 8 月份降水量的误差

从图 17 和图 18 可以看出桂林 8 月份降水量在 LSTM 模型训练下的均方根误差（MSE）和平均绝对百分比误差（MAPE）分别为 251.9621 和 1.1755，很明显观测值和拟合值之间的误差要比 6 月份二者之间的误差大, 但是比 7 月份拟合的效果要好一点。

四、 基于 GA-LSTM 模型的桂林月降水量分析

本文将遗传算法（GA）和长短期记忆神经网络（LSTM）相结合，建立基于 GA-LSTM 模型的桂林月降水量分析预测模型，利用遗传算法（GA）对长短期记忆神经网络（LSTM）的参数进行优化，得出学习率、隐层神经元个数和训练次数的最佳组合，进一步提高模型解决非线性问题的能力。将遗传算法优化过的参数直接运用到 LSTM 模型中，进一步对桂林的降水量进行拟合预测。

（一）基于 GA-LSTM 模型的实证分析

从图 19 可以看出，GA-LSTM 模型拟合的拟合值和真实值的趋势依然大致相同，并且可以很明显的看出二者拟合的程度更为接近，拟合值和真实值基本在一条线上的比较多。由图 20 和图 21 可以看出在 GA-LSTM 模型下训练的桂林 6 月份降水的均方根误差和平均绝对百分比误差分别为 117.2505 和 0.33723。

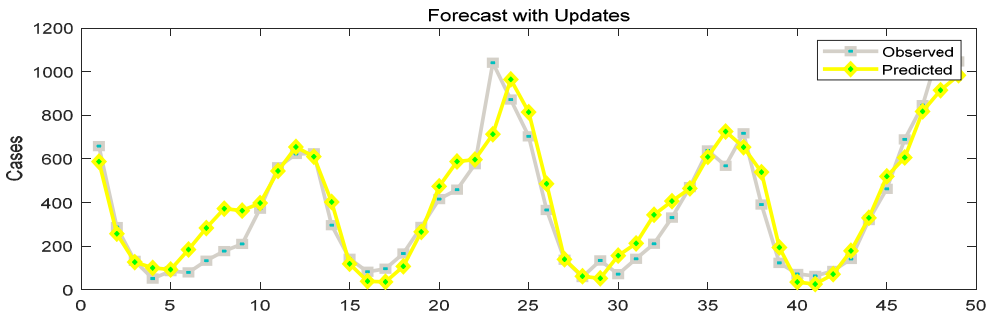

图 19　GA—LSTM 模型下桂林 6 月份降水量的残差

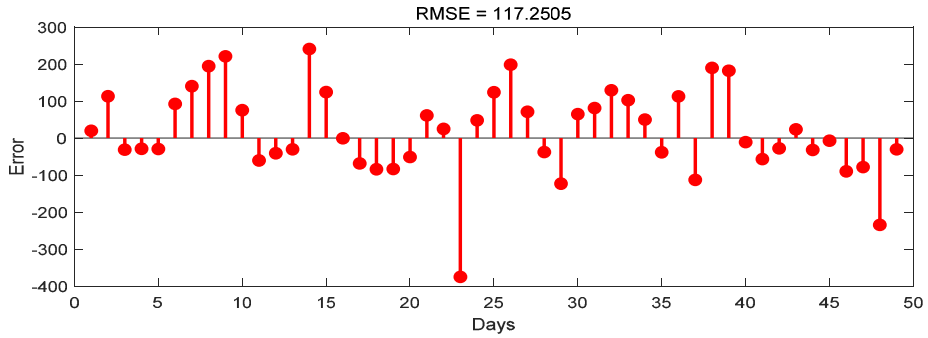

图 20　GA—LSTM 模型下桂林 6 月份的降水拟合情况

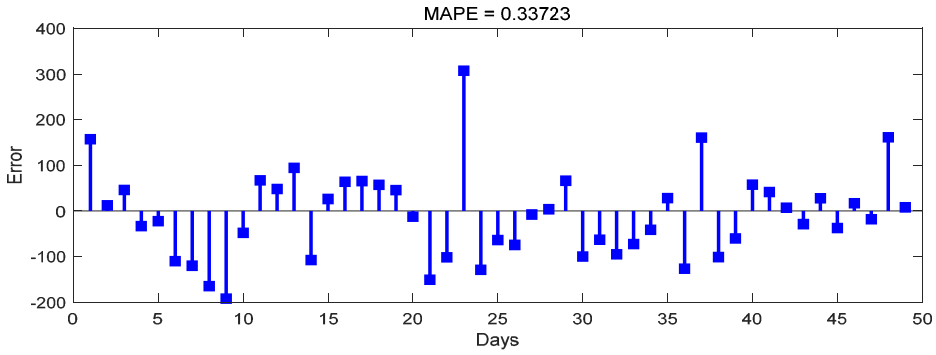

图 21　GA—LSTM 模型下桂林 6 月份降水拟合的残差

从图 22 可以看出，GA-LSTM 模型拟合的拟合值和真实值的趋势依然大致相同，并且可以看出二者拟合的程度比无优化参数的 LSTM 拟合的程度肉眼可见的更为接近。

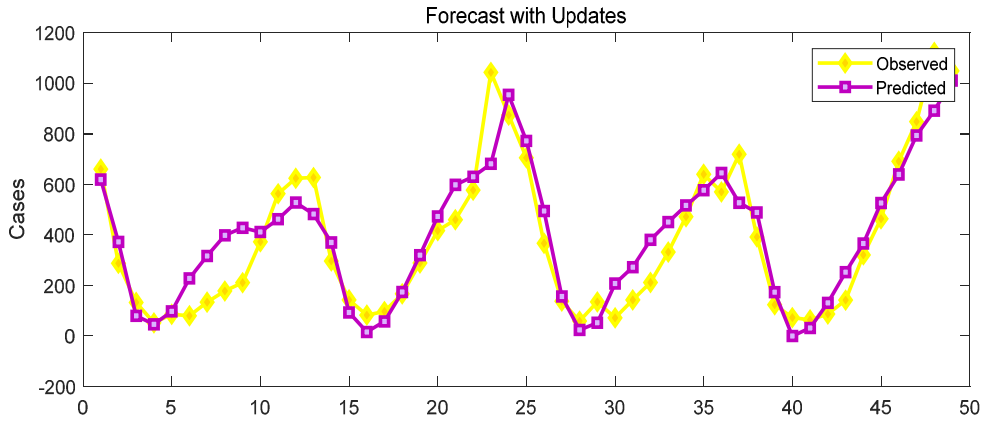

图 22　GA—LSTM 模型下桂林 7 月份的降水拟合情况

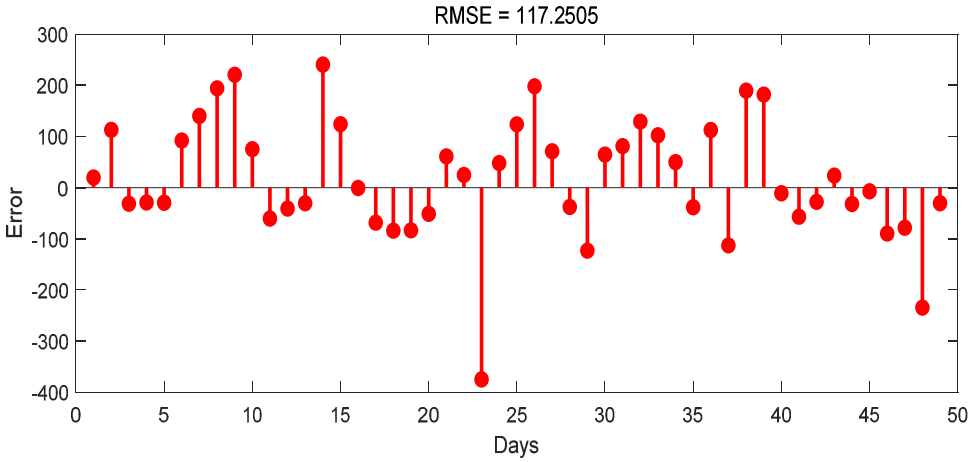

图 23　GA—LSTM 模型下桂林 7 月份降水拟合的残差

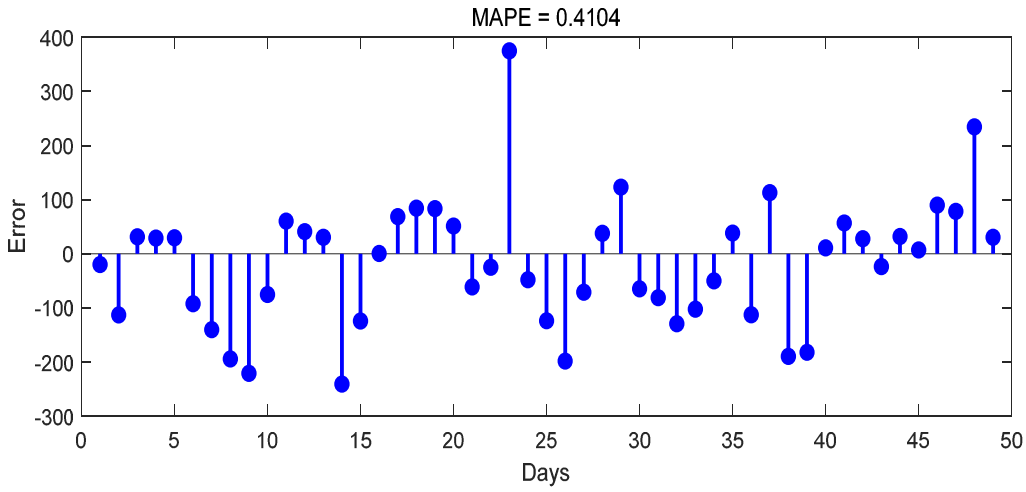

图 24　GA—LSTM 模型下 7 月份降水拟合的误差

由图 23 和图 24 可以看出在 GA-LSTM 模型下训练的桂林 7 月份降水的 RMSE 和 MAPE 的值分别为 117.2505 和 0.4104，该数据表明，GA-LSTM 模型拟合出来的效果真实值和拟合值之间的偏差明显的变小了, 并且在对 7 月份降水量进行拟合时，该模型比 LSTM 模型拟合的效果要好的多，该模型对 7 月份的拟合效果很好的体现了该模型的优点。

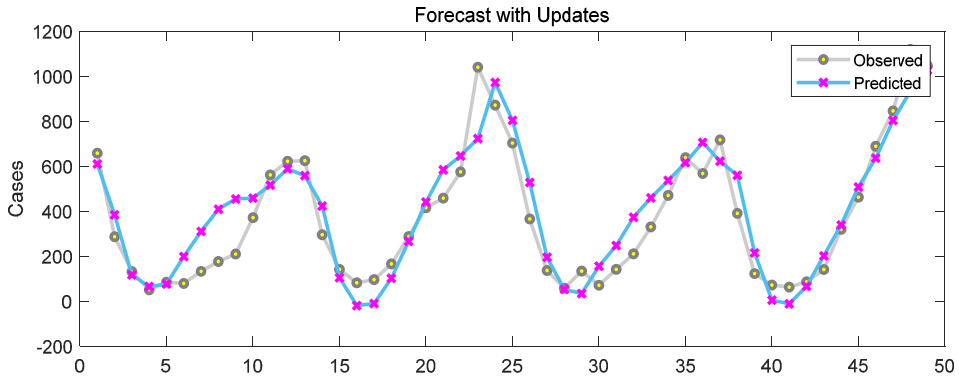

图 25　GA—LSTM 模型下桂林 8 月份的降水拟合情况

从图 25 可以看出，GA-LSTM 模型拟合的拟合值和真实值的趋势还是大致相同，并且可以看出二者拟合的程度也更为接近。由图 26 和图 27 可以看出在 GA-LSTM 模型下训练的桂林 8 月份降水的均方根误差和平均绝对百分比误差分别为 105.1552 和 0.36851，在对 8 月份的降水量进行拟合时，虽然拟合效果比

LSTM 模型的拟合效果好，但是其改进程度弱于 7 月份的降水预测模型。

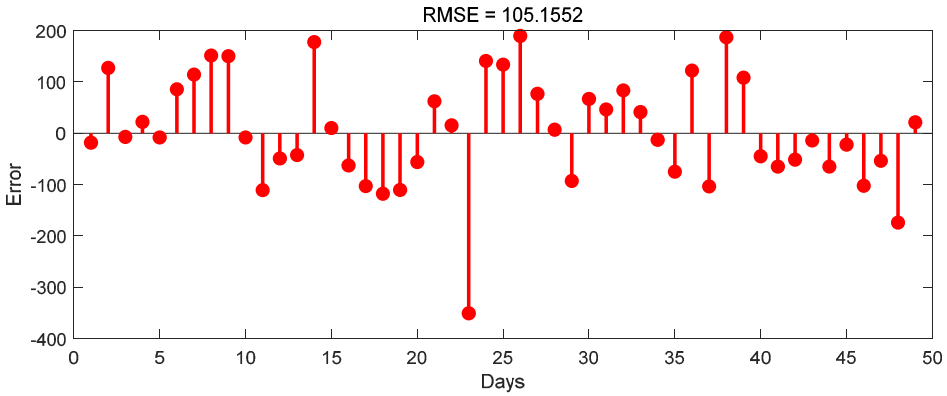

图 26　GA—LSTM 模型下桂林 8 月份降水拟合的残差

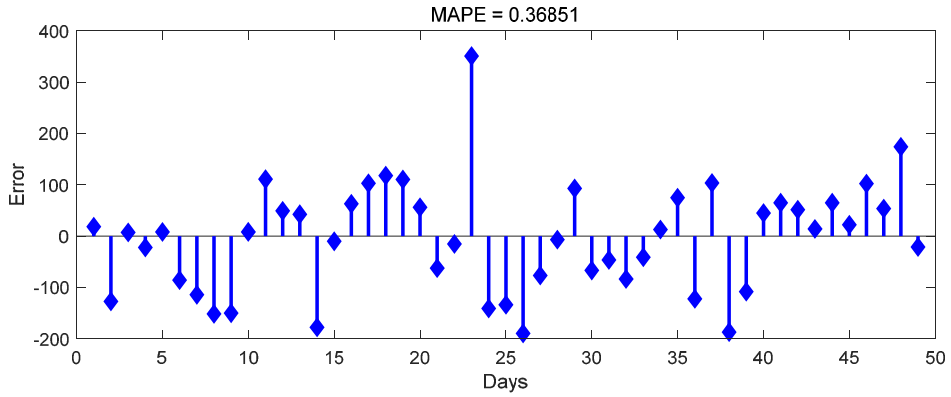

图 27　GA—LSTM 模型下 8 月份桂林月降水量拟合的误差

表 6 是 LSTM 模型和 GA-LSTM 模型对桂林 6 月、7 月、8 月降水量的拟合结果性能指标统计表。由表可以看出，在 LSTM 模型中，桂林 6 月份降水量拟合的 RMSE 为 164.5864，MAPE 为 0.75164，而在 GA-LSTM 模型中，桂林 6 月份降水拟合的 RMSE 为 95.1331，MAPE 为 0.33723，真实值和拟合值之间的偏差降低了 42.19%，也就是模型精度提高了 42.19%；桂林 7 月份降水量的 LSTM 模型的 RMSE 为 356.2494，MAPE 为 1.6171，GA-LSTM 模型下的 RMSE 为 117.2505, MAPE 为 0.4104, 真实值和拟合值之间的偏差降低了 67.08% ;同样的，在 8 月份中，LSTM 模型的 RMSE 为 251.9621，MAPE 为 1.1755，GA-LSTM 模型下的 RMSE 为 105.1552，MAPE 为 0.36851，真实值和拟合值之间的偏差降低

了 58.27%。由这些指标可以清楚的看出，GA－LSTM 模型训练出来的精度要比 LSTM 模型训练的精度高, 预测值的会更接近真实值。因此, 本文选用 GA-LSTM 模型对未来三年桂林 6 月、7 月、8 月的月降水量进行预测。

表 6　LSTM 和 GA—LSTM 的模型拟合结果性能指标统计表

时间	RMSE	MAPE	优化后 RMSE	优化后 MAPE
6 月	164.5864	0.75164	95.1331	0.33723
7 月	356.2494	1.6171	117.2505	0.4104
8 月	251.9621	1.1755	105.1552	0.36851

（二）模型结果分析

用遗传算法优化的长短期记忆神经网络模型对桂林 2020-2022 年 6 月、7 月、8 月份的月降水量进行预测，预测结果如表 7 所示。

表 7　GA—LSTM 模型对桂林未来三年降水量的预测

时间	6 月	7 月	8 月
2020 年	4237	3528	2560
2021 年	4520	3920	2811
2022 年	4802	3809	2735

由表 7 可以看出，桂林在 2020 年-2022 年 6 月份的降水量分别为 4237mm、4520mm、4802mm，这三年的降水量都在 4000mm 以上但是没有超过 5000mm, 说明未来三年桂林 6 月份的降水量还是比较稳定的, 没有出现特大降水也没有出现缺水的情况，但是相关部门也要做好防洪工作。2020 年-2022 年桂林 7 月份的降水量分别为 3528mm、3920mm、3809mm，这三年的降水量也比较稳定，都在 3500mm 至 4000mm 之间，同样的，2020 年-2022 年桂林 8 月份的降水量分别为 2560mm、2811mm、2735mm，稳定在 2500mm 至 3000mm 之间。

由此可见, 未来三年内桂林 6 月、7 月、8 月份的降水量几乎不会发生旱灾, 但是有关部门要做好防洪的准备，提前做好防洪工作。

五、总结和建议

桂林市由于它特殊的地理位置与气候环境，是旱涝灾害频发的地区，使该地的农业生产受到了严重影响。降水作为引起旱涝灾害的最主要的原因，研究降水量也就变得极其重要。本文从降水的角度出发，对桂林市过去 69 年 6 月、7 月、8 月的降水数据进行分析，并且在优化模型的基础上对桂林未来三年 6 月、7 月、8 月的降水数据进行预测，总结如下：

（一）总结

1. 本文提出的 GA-LSTM 模型可以以更高的精度来预测桂林未来的降水量。与普通的 LSTM 相比，该方法具有更好的特征选择过程，因此可以获得令人满意的效果。因为降水量可能受到众多线性的非线性的因素的影响，并且这些因素有的还是潜在的，无法对其进行分析的。而 GA-LSTM 模型可以自动选择合适的特征，不需要考虑其他因素对降水量的影响，从而简化了模型的结构，提高了学习的泛化能力。

2. 桂林 69 年来 6 月、7 月、8 月份的降水数据极差比较大，说明桂林出现极端天气的可能性比较大。用 GA-LSTM 模型预测的桂林未来三年 6 月、7 月、8 月的降水量基本上在 2500mm 到 5000mm 之间，降水还是比较均匀的，但是降水量相对来说还是挺多的。

3.基于遗传算法-长短期记忆神经网络模型在降水预测方面的文献比较少，本文有一定的参考价值。

（二）提出建议

在自然灾害来临之前做好防护工作对减少人民的财产损失至关重要。基于此，本文从政府管理、检测预警能力、水利工程建设、当地的自然条件等方面提出相关建议：

1. 加强防洪抗旱队伍建设，提高政府人员防洪抗旱技术水平

在桂林全市建立防洪抗旱服务中心，进行统一化的管理，提高防洪抗旱技术人员的技术水平和专业能力，改善防洪抗旱所用器具，完善当地对防洪抗旱灾害的预警能力和应对能力。

2. 加强气象观测，提高气象预警能力

加强对气象观测仪的监管，确保其可以不间断的工作，不定期的为气象观测人员进行技能培训，提高对气象部门研究资金的投入，利用最先进的技术对气象进行观测。

3. 加强水利工程的建设和完善，提高防洪抗旱能力

加强桂林市内水利工程设施的资金投入，保证其可以有效地储水放水，加强对当地水库的管理，确保水库一系列的设施可以正常运转，不定期的对水库工作人员进行培训，使其掌握先进的水库管理技术

4. 合理利用自然资源，保护自然环境

在现有基础上加强桂林市自然资源和自然环境的保护，最大力度的改善其空气质量，加强对现有企业排放的废水、废气的监管力度。

参考文献

[1]冉雨晴,吴玮,狄鑫.基于遗传算法优化 BP 神经网络的管网漏失定位模型研究[J].水电能源科学,2021,39(05):123-126+122.

[2]Lim Jong Yeon,Kim Tae Wan,Wang Xiao Yong,Han Yi. Evaluation of Compressive Strength of Sustainable Concrete Using Genetic Algorithm Assisted Artificial Neural Networks[J]. Materials Science Forum,2021,6248.

[3]韩露,史贤俊,林云,秦玉峰.基于遗传算法的贝叶斯网络模型测试配置方法[J].舰船电子工程,2021,41(05):139-142.

[4]刘志萍,杨华,周雨,詹华斌,曹瑜.基于遗传算法的江西七一水库来水流量新安江预报模型参数优化[J].气象与减灾研究,2020,43(02):149-154.

[5]潘迪. 基于遗传算法—支持向量机的柳州降水模型研究与应用[D].广西科技大学,2018.

[6]焦品博,王海燕,孙超,张桂臣.基于长短期记忆神经网络的船舶主柴油机性能预测[J].内燃机学报,2021,39(03):250-256.

[7]Feng Runhai. Uncertainty analysis in well log classification by Bayesian long short-term memory networks[J]. Journal of Petroleum Science and Engineering,2021,205.

[8]李文静,王潇潇.基于简化型 LSTM 神经网络的时间序列预测方法[J].北京工业大学学报,2021,47(05):480-488.

[9]党池恒,张洪波,陈克宇,支童,卫星辰.长短期记忆神经网络在季节性融雪流域降水-径流模拟中的应用[J].华北水利水电大学学报(自然科学版),2020,41(05):10-18+33.

[10]林琪凡,耿旭朴,谢婷,胡利平.基于长短期记忆神经网络的西太平洋暖池变化

预测[J/OL].厦门大学学报(自然科学版):1-11

[11]陶晔. 基于长短期记忆神经网络的气象预测研究[D].南京信息工程大学,2019.

[12]谭琼,廖青桃,李洪伟.基于长短期记忆神经网络的下立交积退水全过程预报[J].给水排水,2021,57(01):144-147.

[13]党池恒,张洪波,陈克宇,支童,卫星辰.长短期记忆神经网络在季节性融雪流域降水-径流模拟中的应用[J].华北水利水电大学学报(自然科学版),2020,41(05):10-18+33.

[14]朱文刚,李昌义,曲美慧,温晓培.深度神经网络方法在山东降水相态判别中的应用[J].干旱气象,2020,38(04):655-664+673.

[15]何文平,王柳,万仕全,廖乐健,何涛.旱涝预测的演化建模方法[J].物理学报,2012,61(11):548-555.

[16]杜懿,龙铠豪,王大洋,王大刚.基于机器学习方法的安徽省年降水量预测[J].水电能源科学,2020,38(07):5-7+41.

[17]赵国羊,涂新军,王天,谢育廷,莫晓梅.基于人工神经网络和支持向量回归机的干旱预测[J].人民珠江,2021,42(04):1-9.

[18]何慧,陆虹,覃卫坚,陆芊芊.人工神经网络在月降水量预测业务中的研究和应用综述[J].气象研究与应用,2021,42(01):1-6.

[19]陈飞盛,孙靖雯.广西汛期不同等级小时强降水时空特征分析[J].现代农业科技,2021(02):162-165+168.

[20]梁维亮,屈梅芳,何珊珊.两种雷达定量降水估测产品在广西区域的误差对比分析[J].气象研究与应用,2020,41(03):1-7.

致谢

值此论文撰写完成之际，回望这将近一个月里的日日夜夜，首先我要由衷感谢我的队员。从论文的选题、数据收集、实证模拟到最后的论文撰写与修改，我们不断讨论、分析、相互协作，在每个深夜里挑灯夜读不知疲惫。正是有了彼此的鼓励和支持，有了集体的智慧，我们的论文才得以在规定时间内顺利完成。

同时我们还要感谢吴建生老师、施业琼老师和胡波老师，感谢他们一直以来对我们的关心和精心指导。在论文选题、梳理框架等环节，老师们都为我们付出的宝贵时间和极大心血，为我们在编写论文的大方向上提供了宝贵的意见。当我们在实证研究中遇到棘手的问题时，老师也耐心指导，让我们少走了很多弯路。在此衷心感谢各位老师对我们的无私帮助！

其次，感谢各位老师和同学们在学习上对我们的帮助，在统计学领域，使我们的专业素养得到了很大提升，同时也加深了对本专业的认知。

最后，我们要感谢含辛茹苦培养我们茁壮成长的父母，在钻研课题遇到困难的时候给予我们恰当的安慰和开导，并且在精神和物质上给予了我们稳固的支持。在这里，我们向我们的父母和家人表示最诚挚的谢意。