中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

题目：基于时间序列的多模型全球气温预测

学年学期：2022-2023 秋季学期

课程名称：机器学习

课程编号：2131030176

课程序号：01

任课教师：刘杉

姓　　名：李菲、李韫琪、董津杉

学　　号：2020213063005、2020213063004、2020213063003

评分区域（由阅卷老师填写）：

结课成绩：

总评成绩：

提交时间：2023 年 1 月 8 日

作者

模型分工（按模型先后顺序）：

（1）SESP 模型、GF 模型：李菲

（2）ARIMA 模型：李韫琪

（3）LSTM 模型：董津杉

数据收集、处理、摘要部分、结论、模型现实意义与未来工作、文档编写及校正：共同完成

摘要

全球变暖一直是近年来困扰人们的问题，为提高人们对全球变暖问题的重视，我们小组选择建模预测未来全球温度水平。由于收集到的数据庞杂，本组首先对数据预处理，对所拥有的数据进行删除、插值、异常值处理等方法，以得到可以利用的数据，并根据不同地区、气候的地区得到年度和月度全球气温数据。

之后本组建立了 SESP 模型与 GF 模型，结果发现，SESP 模型的二次指数模型因预测不符合实际忽略，SESP 三次指数模型与 GF 模型得出的预测较为准确，表明全球年平均气温总体呈增长趋势，且增长速率较快。

在成功预测了全球年平均气温总体的未来趋势后，我们决定将工作细化到每个月的全球平均气温的预测，得到更加细节的未来气温数据。因此接下来建立了 ARIMA 模型，基于时间序列的预测模型，细化预测未来一年每个月份的世界平均气温。预测结果显示出全球气温在呈现整体逐年增长的趋势的同时，在每一年内呈现出了类似的周期性起伏变化。

但由于该模型受到逐年增长的大趋势的影响，我们并不能对其周期性变化进行较好的分析。因此我们引入了长短期记忆 LSTM 模型，排除大环境因素的干扰。通过只比较一年内的各个月份气温的变化，进一步分析月份与月份之间的联系，我们得出结论全球平均气温在每年 1 月与 12 月相对较低，而 6 月份则达到峰值，总体呈现先上升，后下降的趋势。

关键词：全球气温、灰色预测模型、SESP 模型、ARIMA 模型、LSTM 模型

目录

一、背景

瑞士时间 11 月 6 日，世界气象组织(WMO)发布了《2022 年全球气候状况》中期报告指出，近年来温室气体浓度在上升，热量在积累，过去八年预期成为气象记录中最热的八年，而气候变化的影响变得越来越严重。

按目前数据估计来看，2022 年全球平均气温会比工业化前的平均气温高出约 1.15℃。在罕见的“三重”拉尼娜现象的降温作用影响下，2022 年可能所有年份中前十暖的年份，虽然 2022 年不是第一，但不代表长期变暖的趋势停止了，按目前的增长趋势，再次出现有记录以来最暖的年份只是时间问题。

全球气候变暖是一种与自然有关的现象。正是由于温室效应的不断积累，导致地球大气系统吸收和排放的能量失衡，能量在地球大气系统中不断积累，导致气温上升，全球气候变暖。

在全球气温逐渐上升的今天，我们需要合适的模型分析并预测全球气温，以达到对未来气温掌握的目的，也能让人们看到全球变暖的现象，使人类对全球变暖这个现象做出响应措施。

二、数据的选择与处理

1. 数据筛选

我们收集到的原始气温数据涵盖了 1743-2013 年全球范围内 100 个城市的月度温度情况。由于部分城市在 1743-1850 存在较多时间跨度大的数据的缺失，影响后续数据分析。同时存在大量地理空间重合的气温观测站点，导致数据冗余。因此，为最大程度消除冗余数据，保证时间序列的连续性，本文结合全球气候类型和表单所提供数据的地域分布和温度，筛选出具代表性 7 个城市，分别是：印度新德里、巴西巴西利亚、法国巴黎、意大利罗马、中国北京、美国纽约、俄罗斯圣彼得堡，对他们 1851-2012 年的月度温度数据进行进一步地挖掘分析，在此基础上建立模型，实现对全球温度的未来预测。

最终筛选出的 7 个城市的数据具有全球代表性：

（1）所选地点能代表全球主要气候类型

所选 7 个城市涵盖了全球多数关键气候类型，包括热带季风气候、热带草原气候、温带海洋气候、地中海气候、温带季风气候、温带大陆性气候，

（2）所选地点在全球范围内达到基本均匀分布

所选 7 个城市包括南北半球和东西半球城市，纬度包括 60.27N、49.03N、42.59N、40.99N、39.38N、28.13N、15.27S，经度包括 2.45E、13.09E、29.19E 、77.27E、47.50W、74.56W、116.53E。

2. 缺失数据的处理

筛选后的数据仍然存在部分缺失值，但此处本文没有对缺失值直接进行简单的删除处理而是选择补全数据。因为 7 个城市中仅有印度新德里存在数据缺失，具体时期是 1855 年 3 月、1856 年 5 月、1858-1869 年，可见缺失值处在时间序列的中部。为确保时间序列的连续

性和后续数据分析数据的完整性和结果的准确性，我们进行了对空缺数据的平均值插值补全。由于温度数据属于稳定数据，在十年间隔中不会出现快速增长或下降的剧烈幅度变化，因此平均值插值补全方法具有可行性。

具体来说，我们综合缺失数据过去和未来的情况，用缺失数据的相邻年份的有值的月平均数据求取平均值并对缺失数据进行填充，公式如下：

$$t_2 = \frac{t_1 + t_3}{2}$$

其中 t_2 表示缺失月份平均温度，t_1 表示同地区前一非空缺值年份的同月份平均温度，t_3 表示同地区后一非空缺值年份同月份平均温度。如：t_2 表示 1855 年 3 月新德里温度，t_1 表示 1854 年 3 月新德里温度，t_3 表示 1856 年 3 月新德里温度。

得到修正后的数据如下表所示：

表 1 缺失颜色数据修补

dt	修复前	修复后
1855-03-01	--	22.489
1856-05-01	--	32.974

dt	修复前	修复后
1858-01-01	--	13.691
1858-02-01	--	16.551
1858-03-01	--	22.374
1858-04-01	--	28.307
1858-05-01	--	33.031
1858-06-01	--	33.059
1858-07-01	--	30.860
1858-08-01	--	29.191
1858-09-01	--	28.530
1858-10-01	--	25.208
1858-11-01	--	19.794
1858-12-01	--	14.549

注：表中“--”代表数据缺失，1858-1869 年缺失数据均由 1857、1870 年计算而得，此处展示 1858 年月平均气温情况。

3. 世界气温的表示

本文使用筛选和补全后的数据计算世界气温，年度世界气温计算公式如下：

$$\bar{t_1} = \frac{\sum_{\alpha=1}^{7}\sum_{\beta=1}^{12} t_{\alpha\beta}}{7 \times 12}$$

月度世界气温计算公式如下：

$$\bar{t_2} = \frac{\sum_{\alpha=1}^{7} t_{\alpha\beta}}{7}$$

其中 $\bar{t_1}$ 表示年度世界气温，$\bar{t_2}$ 表示年度世界气温。该年 $t_{\alpha\beta}$ 表示地区 α 第 β 月的月平均气

温。

下表为1851-2012年年度世界气温的节选展示：

表2 1851-2012年年度世界气温的节选

Year	$\bar{t}_1$
1860	12.821
1880	13.437
1900	13.713
1920	13.784
1940	13.067
1960	13.789
1980	13.554
2000	14.672

根据本文所选择的模型，我们仅使用到部分的世界月度气温，即使用2001-2012年的世界月度气温对2013年的月度世界气温进行预测。注意此处2001-01-01代表2001年1月的世界平均气温，其他月份同理。下表为2001-2012年世界月度气温的节选展示：

表3 2001-2012年世界月度气温的节选

Month	$\bar{t}_2$
2001-01-01	5.36
2001-02-01	6.37
2001-03-01	10.08
2001-04-01	14.56
2001-05-01	19.44
2001-06-01	21.46
2001-07-01	23.27
2001-08-01	23.07
2001-09-01	19.42
2001-10-01	16.35
2001-11-01	10.29
2001-12-01	5.14

2001年1月至2012年12月的月度世界平均气温情况绘制如图：

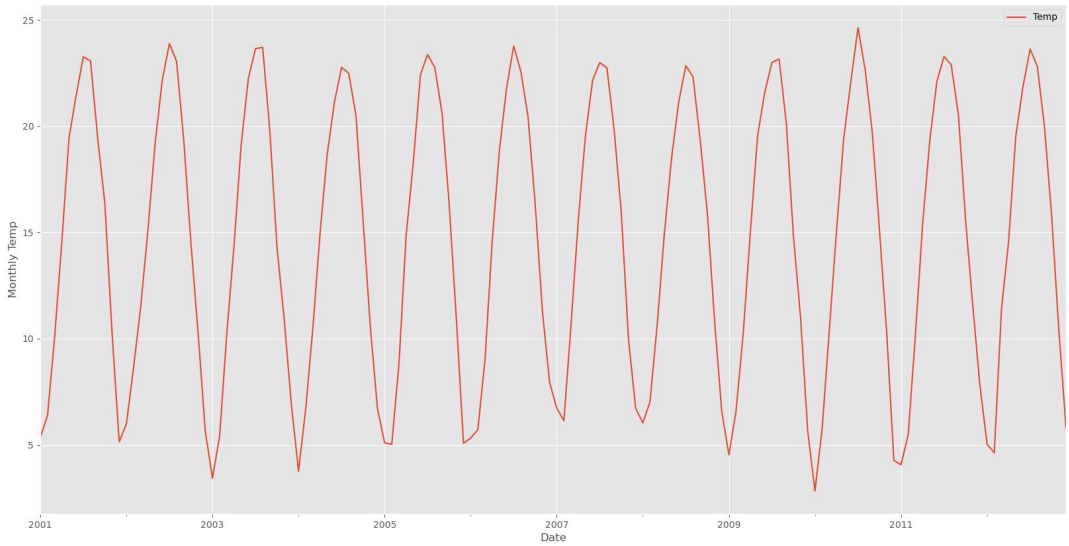

图 1 2001 年 1 月至 2012 年 12 月的月度世界平均气温图

4. 异常数据的处理

从所计算的世界气温数据来看，某些年份世界气温可能存在异常，注意此处数据异常并非由真实气候异常所导致，而是数据整理和计算过程导致的不可避免的异常偏差。为了保证最终结果的准确合理，必须对异常世界温度数据进行识别并剔除，空缺部分替换为平均值插值处理后的数据。

本文选择 Smoothed z-score 方法对异常数据进行识别，这种方法的主要思想是在一段历史时间序列中，基于数据的均值和标准差对下个时间节点进行预测，并和阈值比较，识别出异常数据，以实现进一步的平滑修正。

利用 Smoothed z-score 方法对世界气温异常值识别代码见附件，识别结果如下图：

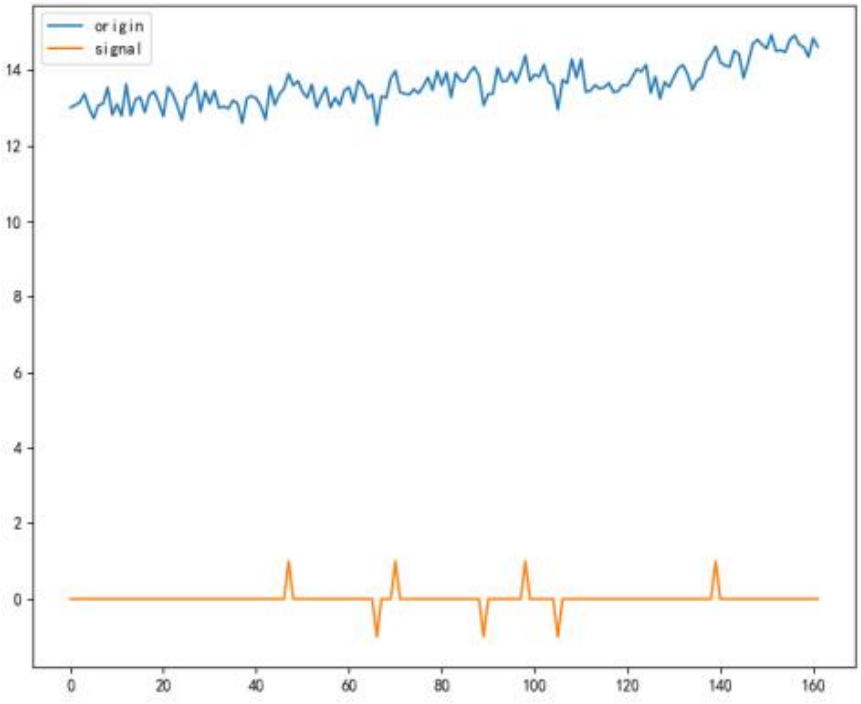

图 2 世界气温异常值识别图

结果显示1894年、1917年、1921年、1940年、1949年、1956年、1990年为异常点，对上述年份数据剔除并使用上文叙述的缺失数据补齐方法修正结果，最终结果见附件数据集。

三、建模，仿真与分析

（一）利用SESP 模型、GF 模型分析全球年平均气温趋势

1. 引入目的

基于历史数据，对全球温度水平的过去进行描述并且对未来进行预测，需要我们对全球不同年份的平均气温利用模型，进行数据处理与分析。原始数据有来自很多国家不同的100个地区的平均月气温，跨度从1851年1月到2013年9月。这些数据过于庞大且复杂，并不利于模型的建立以及模型对数据的描述与预测。所以我们用前文提到处理好的数据，结合搜索到的数据集合成1851年到2012年每年平均温度放入模型进行分析与预测。

由于研究的内容是温度随着时间的变化规律及根据规律预测，在准备建立模型的过程中更倾向于使用时间序列模型，故引入两个个模型：Gray Forecast 模型(简称为GF)、Simple Exponential Smoothing Prediction 模型(简称为SESP).其中，SESP 模型为基于时间序列的预测模型。

2. 模型原理

（1）SESP 模型

为研究全球每年平均气温随年份的变化规律，我们首先选用了SESP 模型。

SESP 模型以某个指标的实际值和预测值作为模型基础，引入一个简化的加权因子，作为平滑系数，从而可以求得平均数的时间序列预测法。它对离预测期较近的历史数据给予较大的权值，权值由近到远按指数规律递减的一种特殊的加权平均法。

指数平滑可继续划分为一次平滑，二次平滑和三次平滑，一次平滑法为历史数据的加权预测，二次平滑法适用于具有一定线性趋势的数据，三次平滑法在二次平滑法基础上再平滑一次，通常情况下使用三次平滑法较多。

(a) 一次平滑：

当时间数列无明显的趋势变化，可用一次指数平滑预测。其预测公式为：

$$y^{t+1} = \alpha Y_t + (1-\alpha)y^t$$

y^{t+1} ——t+1期的预测值;

Y_t ——t 期的实际值；

y^t ——t 期的预测值。每个 t 期为一年。

(b) 二次平滑

一次指数平滑法有明显的局限性：它只适用于水平型历史数据的预测，不适用于呈斜坡型线性趋势历史数据的预测。

解决步骤：

1.先求出一次指数平滑值和二次指数平滑值的差值；

2.将差值加到一次指数平滑值上；

3.再考虑趋势变动值。

$$F_{t+T} = a_t + b_t T$$

F_{t+T}为 t+T 期的预测值；

T 为 t 期到预测期的间隔期数；

a_t、b_t为参数。

(c) 三次平滑

三次指数平滑是在二次指数平滑的基础上再进行一次平滑，其计算公式为

$$S_t^{(3)} = \alpha S_t^{(2)} + (1-\alpha) S_{t-1}^{(3)}$$

（2）GF 模型

(a) 模型原理

灰色预测可以对含有不确定因素的数据进行预测。它通过鉴别数据之间不同因素的发展趋势的相异程度进行关联分析，寻找原始数据变动的规律，生成有规律性的数据序列，建立微分方程模型预测未来发展趋势状况。用等时距观测到的反映预测对象特征的一系列数量值构造灰色预测模型，预测未来某一时刻的特征量，或达到某一特征量的时间。

(b) 建模过程

定义 x(1)的灰导数为 $d(k) = x^0(k) = x^1(k) - x^1(k-1)$，令 z-1(k)为数列 x1 的邻值生成数列，即$z^1(k) = ax^1(k) + (1-a)x^1(k-1)$，于是定义 GM(1,1)的灰微分方程模型为 $d(k) + az^1(k) = b$，其中，a 称为发展系数，z-1(k)称为白化背景值，b 称为灰作用量。接下来我们得到如下方程组：

$$x^0(2) + az^1(2) = b$$
$$x^0(3) + az^1(3) = b$$
$$\cdots\cdots$$
$$x^0(n) + az^1(n) = b$$

按照矩阵的方法列出：

$$u = \begin{bmatrix} a \\ b \end{bmatrix}, Y = \begin{bmatrix} x^0(2) \\ x^0(3) \\ \cdots \\ x^0(n) \end{bmatrix}, B = \begin{bmatrix} -z^1(2) & 1 \\ -z^1(3) & 1 \\ \cdots\cdots & \\ -z^1(n) & 1 \end{bmatrix}$$

则 GM(1,1)就可以表示为 Y=Bu,接下来就是求 a 和 b 的值，可以使用线性回归或$(B^TB)^{-1}B^TY$（正规方程）按照最小二乘的原理来求出 a 和 b 的值。

3. 仿真结果及分析

（1）根据代码得出全球年平均气温趋势

（a）根据 SESP 模型得出结果

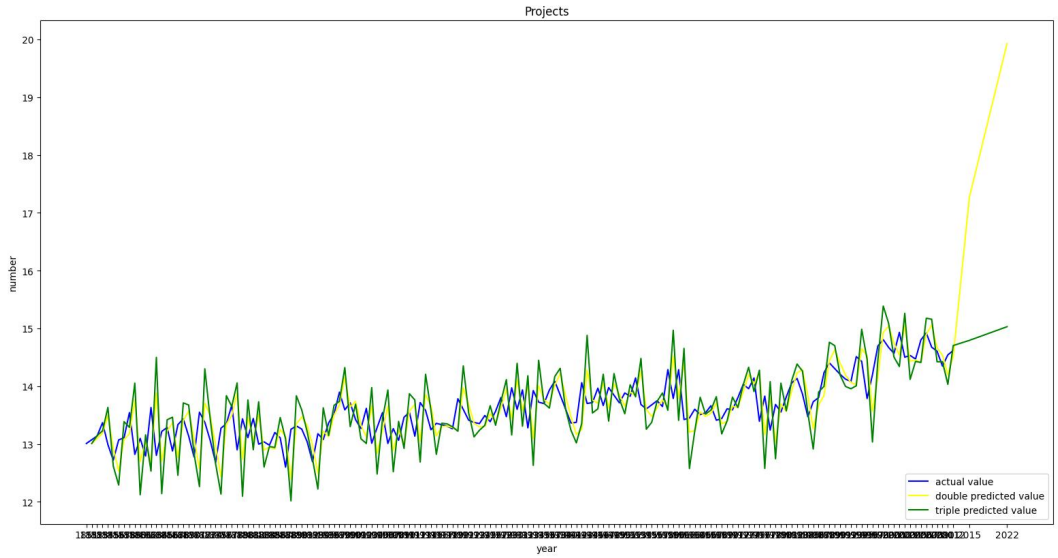

图 3 SESP 模型全球气温趋势图

上图为利用 SEMP 模型得到的分析结果。其中，蓝色线条为原始数据，黄色线条为二次平滑，绿色线条为三次平滑。不难发现，虽然气温的起伏波动较大，但总体数据呈上升趋势，尤其是到了 1950 年之后。根据图片对未来的预测，可得出以下观点：

1）虽然三次指数放大了数据的波动，但对数据走向的预测较为准确，比较符合大致走向。

2）二次指数比三次指数在过去数据的拟合性更好，但在对未来的预测上却增幅过大，根据图片数据走向可看出预测的可信度并不高。

3）结合二次指数与三次指数，可得出未来气温增幅在到 2015 年间涨幅会高于 2015 年到 2022 年涨幅，后续涨幅会稍趋于平缓。

（b）根据 GF 模型得出结果

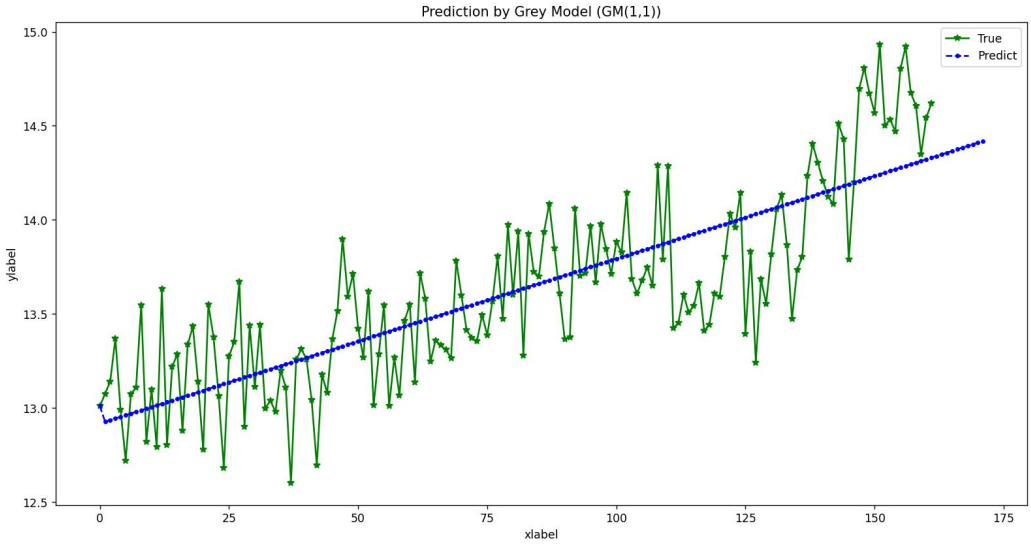

图 4 GF 模型全球气温趋势图

上图为利用 GF 模型得出的分析结果，图片中绿色线为真实值，蓝色线为根据过去数据拟合的预测直线根据图片，可得出以下观点：

1）此模型预测未来的规律，是完全源于对过去数据的分析数学规律并沿用得来的。

2）此模型的预测线是一条直线，满足线性关系，且是根据过去的气温水平进行拟合的，较为平稳。

（2）根据使用的模型预测 2015 年与 2022 年全球温度

首先，依据上面模型分析 1851 年至 2012 年整体趋势预测 2015 年与 2022 年温度如下（因为 SESP 模型有二次指数与三从指数两种预测，故写为两行，实际二者属于一个模型）：

表 4 SESP 模型与 GF 模型预测全球气温预测表

	SESP(double)	SESP(thrible)	GF
2015	17.27	14.79	14.35
2022	19.93	15.03	14.81

由上表可得出以下结论：

1）与 SESP 模型预测图得出结论相同，SESP 模型的二次指数增长率过快，偏离了原有曲线轨道，预测结果并不可信。

2）SESP 模型的三次指数虽然增长速率比二次指数平缓，但由于近年数据权重高，所以起始温度较大，导致预测出来温度都较高。

3）GF 模型的增幅稍大，但总体满足增长状况，预测较为合理。

根据以上结论来看，除了 SESP 的二次指数模型之外，其他模型的增长率宏观来看都并不高，而二次指数模型却严重偏离了原有温度涨幅轨道，是不可信的。

4. 模型评价

（1）SESP 模型

首先，根据此模型原理，如果序列的基本趋势比较稳，预测偏差由随机因素造成，则加权系数 α 值应取小一些，以减少修正幅度，使预测模型能包含更多历史数据的信息。而本序列的基本趋势较为稳定，所以应该选择较小的 α 值。但在代码修改与实践中发现，此模型修改 α 导致的总体结果差异不大。

对于过去数据的描述，本模型用二次指数与三次指数进行了数据放大的效果，并不能很直观的看出过去的趋势情况。

根据上面的图表数据也可以发现，SESP 模型在 1851 年到 2012 年的基础上的预测结果，二次指数斜率过大，三次指数斜率过小，可以看出这个模型并不适合预测未来全球气温水平。

（2）GF 模型

GF 模型对于过去的描述，也进行了拟合，拟合为了一条近似直线，可以看出来大致趋势，但对于未来年份的温度预测无法做到十分精准。

GF 模型对未来的预测，在全体数据分析的结果中抽取部分值放在下表中：

表 5 GF 模型预测误差表

索引项	原始值	预测值	残差	相对误差（%）
1866	13.287	13.016	0.271	2.039
1911	13.549	13.445	0.104	0.765
1930	13.973	13.627	0.346	2.474
2009	14.607	14.388	0.219	1.502

根据上面的表可以得出，在全体数据分析结果中，原始值与预测值误差控制在 3%以内，实际值和预测值的差距在可以接受的范围之内，如果用来观察目前情况下的全球气温未来趋势，还是非常有用的。

（二）利用 ARIMA 模型预测一年中每月世界平均气温

1. 引入目的

上述的 SESP 模型与 GF 模型已经成功预测了全球年平均气温总体的未来趋势，但该模型的预测是立足于年平均气温，也就是说并没有细化到每个月的全球平均气温的预测。因此接下的工作我们选择建立 ARIMA 模型基于时间序列的预测模型，希望使我们的工作进一步细化到对未来一年每个月份的世界平均气温的预测。

2. 模型原理

（1）ARIMA 模型

(a) Autoregressive 模型(简称为 AR)

AR 指自回归模型。自回归模型主要可以用来描述当前值与先前的历史值间的关系。也就是说利用 AR 可以实现通过变量本身的历史时间数据实现对未来一段时间的预测。值得注意的是，自回归模型的前提需要满足平稳性的相关要求。而且自回归模型在使用前，需要我们自行确定参数 p--一个阶数，用来表示在模型中将使用多少期的历史值来进行当前值的预测。p 阶自回归模型的公式如下：

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t$$

上式中 y_t 代表了当前值,u 代表了常数项,p 是阶数 γ_i 代表了自相关系数，ϵ_t 代表了误差。

然而自回归模型存在着一些不利的限制，包括：

1）自回归模型完全是在自身数据的基础上进行的预测；

2）必须具有平稳性和相关性，也就是说自相关系数小于 0.5 的情况，并不合适使用自回归模型。

(b) Moving Average 模型(简称为 MA)

MA 模型是指移动平均模型，该模型可以实现有效消除预测中存在的随机波动现象。移动平均法关注的是先前介绍的自回归模型中的误差项，将其进行累加，q 阶移动平均法公式如下：

$$y_t = \mu + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

(c) Autoregressive Moving Average 模型(简称为 ARMA 模型)

AR 模型和 MA 模型相结合，我们就得到了 Autoregressive Moving Average 模型:ARMA(p,q)，它指的是自回归移动平均模型，计算公式如下：

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

(d) ARIMA 模型

将 AR 模型、MA 模型和差分法结合，我们就得到了 ARIMA(p,d,q)，其中 d 是需要对数据进行差分的阶数。

ARIMA 的全称为差分整合移动平均自回归模型。ARIMA 中存在三个整数参数他们是：(p, d, q)，使用他们 ARIMA 模型参数化处理。ARIMA(p, d, q)的计算公式如下：

$$(1-\sum_{i=1}^{p}\phi_i L^i)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\varepsilon_t$$

其中 p 代表了自回归项数(它来自于 AR 部分)。p 参数的作用是，可以把过去的数值的影响加入进我们的模型里面。直观地举例说明，如果我们已经知道了过去 3 天的气候是温暖的，那这种情况我们可以说明天的气候可能也是温暖的。

d 参数的作用可以稳定我们需要的差异的数量。还是举例直观说明，如果我们已经知道了过去三天的温及其细微，那这种情况我们可以说明天的气候情况可能是与这过去的三天相同的。

q 参数主要是用来预测来自方程的滞后预测的误差数(也就是先前介绍的 MA 部分)。q 参数的主要作用就是让我们模型的误差实现了可以将其设置为过去的任何一个时间点所观测的误差值线性组合形式。

但值得注意的是仅仅是非季节性 ARIMA 模型用 ARIMA(p, d, q)表示，而本文的目标在于预测季节性的全球气温情况，因此我们需要的是季节性 ARIMA，它被表示为：

$$ARIMA(p,d,q)(P,D,Q)s$$

这里，(p, d, q)代表了上面描述的非季节性参数，注意(P,D,Q)遵循与之相同的定义，不同的是它是应用于时间序列的季节分量。项 s 代表了时间序列的周期属性。很明显在本文研究的问题中 s=12。而其他参数的设置其实在很大的程度是取决于已有的经验。也可以参考许多效果优异的实践来进行设置。需要说明的是，我们的工作是将网格搜索的预定义值范围内

的所有可能的参数值组合。我们使用了 AIC 值，AIC 可以在给定模型表示生成数据的过程中，对丢失的信息进行估计。过程中 AIC 值还权衡了模型拟合优度与模型复杂度。

3. 仿真结构及分析

（1）确定参数值

通过调用 stats 模型 s.api.tsa.statespace.SARIMAX，可以返回 AIC(赤池信息准则)和 BIC(贝叶斯信息准则)的值，这个步骤是通过最小化来选择出最佳的拟合模型。

季节性 ARIMA 的参数组合表示示例如下：

SARIMAX:(0,0,1) x (0,0,1,12)

SARIMAX:(0,0,1) x (0,1,0,12)

SARIMAX:(0,1,0) x (0,1,1,12)

SARIMAX: (0,1,0) x (1,0,0,12) ......

本文选择了数据系列的一个子集将其用作训练的数据，也就是前 11 年的月度气温数据。我们的目标是根据这些气温的输入来预测最后一年的气温情况。

关于确定 AIC 值和季节性 ARIMA 的参数组合的代码运行结果如下：

表 5 确定 AIC 值和季节性 ARIMA 的参数组合运行结果表

N	Tit	Tnf	Tnint	Skip	Nact	Proig	F
9	28	42	1	0	0	1.007D-04	6.290D-01

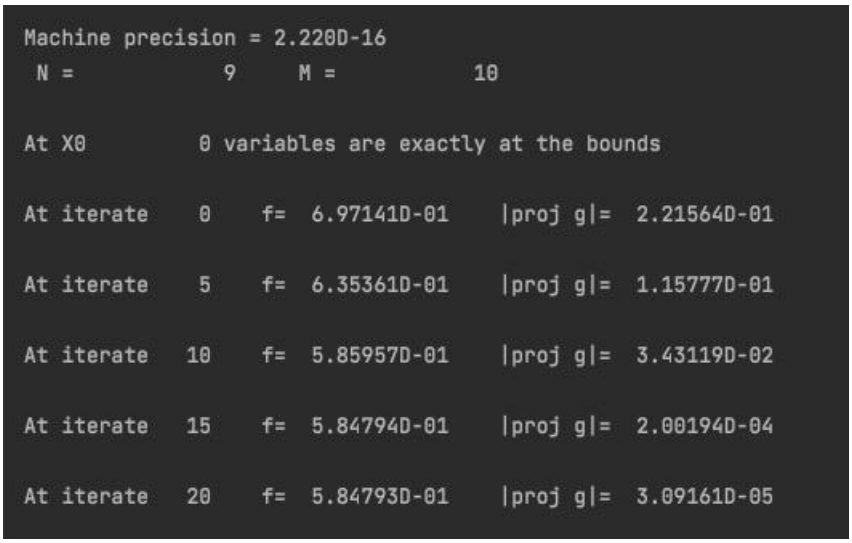

图 5 确定 AIC 值和季节性 ARIMA 的参数组合运行结果图

Tit 代表了总迭代次数，Tnf 代表了函数求值的总数，Tnint 代表了柯西搜索中探索的段的总数，Skip 代表了跳过 BFGS 更新的数量，Nact 代表了最终广义柯西点的活动边界数，Projg 代表了最终投影梯度的模，F 代表了最终函数值。

F = 0.62897435703533366

SARIMAX(3, 1, 1)x(3, 1, 1, 12) - AIC:184.04923025732808

因此本文选择模型的最小 AIC 值是 172.38547630797535 ，ARIMA 的参数组合取值为 SARIMAX(3, 0, 1)x(3, 1, 1, 12)。

（2）模型拟合和检验

接下来进行拟合模型：

表 6 确定 AIC 值和季节性 ARIMA 的参数组合拟合模型表

N	Tit	Tnf	Tnint	Skip	Nact	Proig	F
9	20	25	1	0	0	3.092D-05	5.848D-01

F = 0.58479347086354305

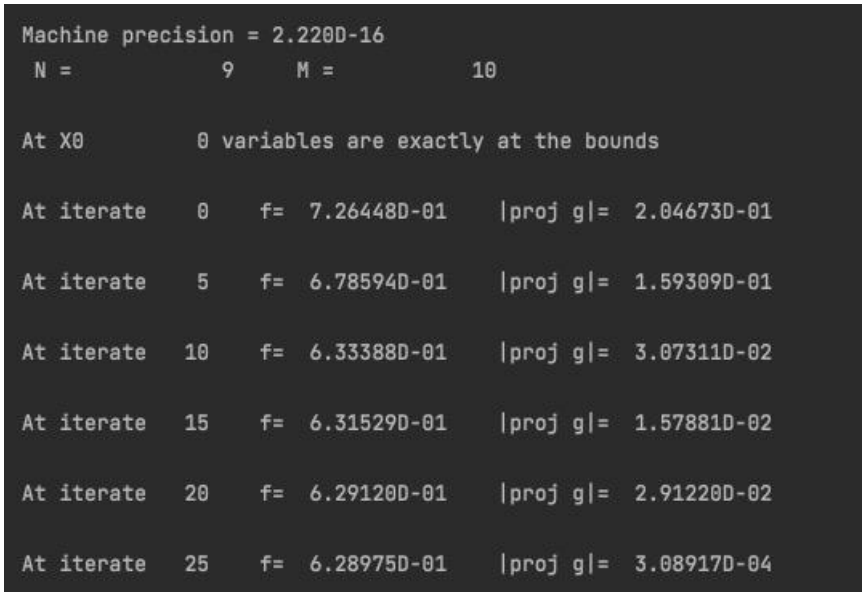

图 6 确定 AIC 值和季节性 ARIMA 的参数组合模型拟合图

模型被拟合后，我们还需要对模型进行检查，观察是否符合原有预期，以及模型结果是否违法所做的假设。我们计算了残差并绘图，见下图右下角。结果表明残差结果是不相关的而且也没有表现出任何明显的季节性，这可以通过左上图得知。除此以外，残差和大致正态分布与零平均值的结果呈现在右上角的图中。左下角的图则显示出残差(图中的蓝点部分)的有序分布大致遵循从了符合 N(0,1)标准的正态分布中采集的样本的线性趋势。这个结论也再次有力证明我们的残差是符合正态分布的。

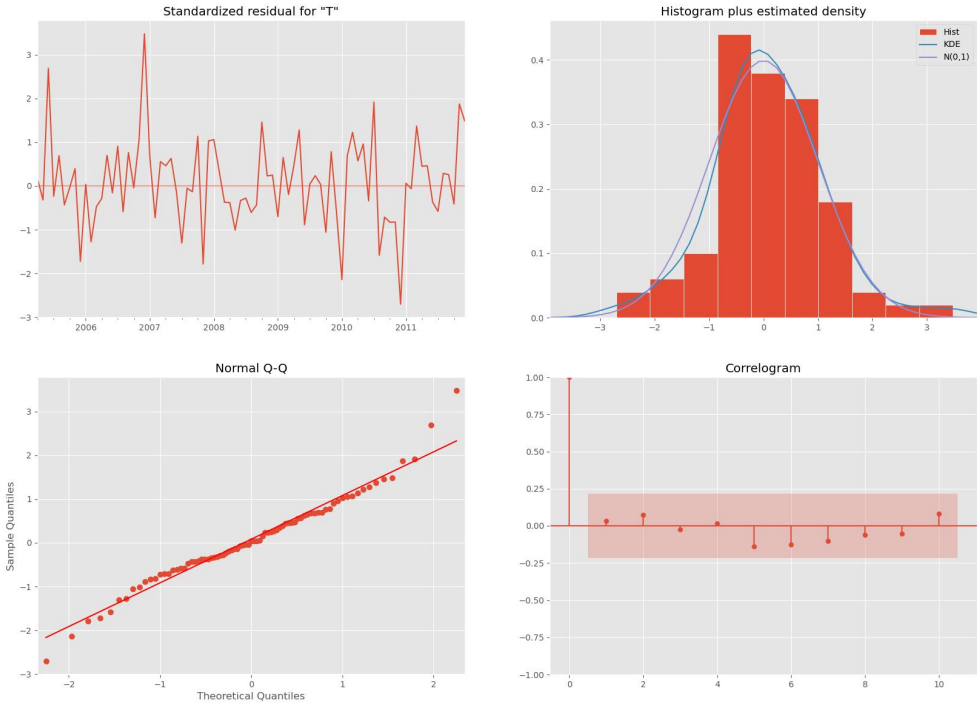

图 7 模型检验图

我们通过自相关函数 ACF 方法选择参数 p 与 q。因为有序的随机变量序列往往会和它本身相比较的自相关函数，反映出同一个序列处于不同时序里取值上存在的相关性 。ACF 的公式如下：

$$ACF(k) = \rho_k = \frac{Cov\ (y_t, y_{t-k})}{Var\ (y_t)}$$

ρ_k 的取值范围为[-1，1]，分子代表了 y_t 和 y_{t-1} 到 y_t 和 y_{t-k} 的相关系数，分母代表了方差。

（3）预测结果

现在就可以进行对未来天气情况的预测工作了。我们将使用下述三种方法实现结果的预测:

（a）在样本预测中，提前一步预测上一年度(2011 年)，具体来说，提前一步预测就是指用预测的每个点预测下一个点，预测结果在代码中计作 pred0。

（b）在过去一年(2011 年)的动态预测样本预测中。同样，该模型用于预测模型所基于的数据。预测结果在代码中计作 pred1。

（c）样本外数据的“真实”预测。在这种情况下，模型被要求预测它以前没有输入过的数据。也是我们的目标预测结果，这部分的预测结果在代码中计作 pred2。

我们成功预测出了 2013 年 12 个月的全球气温情况，见下表：

表 7 2013 年 12 个月的全球气温情况预测表

Month	Temp
2013-01-01	6.155790
2013-02-01	6.519149
2013-03-01	10.010506
2013-04-01	14.874503
2013-05-01	18.736440
2013-06-01	21.610584
2013-07-01	22.839593
2013-08-01	22.765785
2013-09-01	20.075409
2013-10-01	15.847618
2013-11-01	10.728590
2013-12-01	6.721220

把预测结果全部绘制如图：

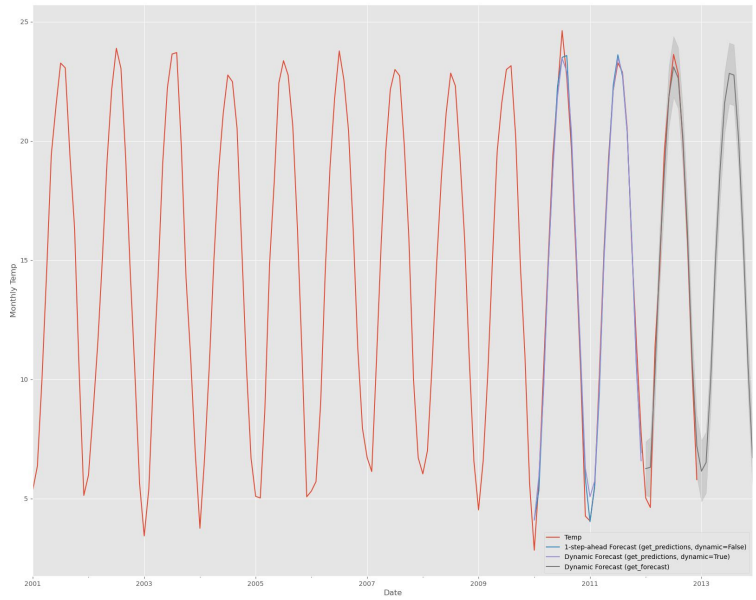

图 8 2013 年 12 个月的全球气温情况预测图

图中红线表示全球平均温度，蓝线紫线和灰色线分别代表了上述 a、b、c 三种预测结果。从图中可以看出，我们所构建的模型在时间序列建模方面的效果非常优秀。因为正如我们预期的那样，蓝色和紫色的线非常接近红线所指示的真实值。对于样本外预测结果的灰色线结果同样优秀，完美符合全球平均气温的预期。对于这样一个简单的时间序列，我们构造的 ARIMA 模型能相当准确地预测出数值。

4. 模型评价

代码算法评价指标使用的是：MAPE 和 MSE

（1）MSE 均方误差

为了量化 2012 年全球气温预测的准确性，我们计算均方误差指标，均方误差 MSE 具体来说就是利用模型预测出的参数的估计值和参数的真实值之间的差的平方的期望。MSE 是一

种常用的模型评价的度量--可以做到刻画预测值与被预测量之间的差异程度。因此可以说MSE是衡量我们的预测模型的平均误差的一种非常便捷的方案。MSE 可以评价数据的变化程度，注意，MSE的值越小，就意味着我们的预测模型描述实验的数据有着更加优秀的准确度。MSE公式如下：

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(observed_t - predict_t)^2$$

这里N表示测量次数，observed_t表了参数的真实值，predict_t代表了参数的预测值。

本模型对2012年全球月度气温预测的MSE为0.874，平均绝对百分比误差较小，表明模型预测准确度高。

（2）MAPE平均绝对百分比误差

先前的MSE均方误差是一种绝对度量的指标，因此都是尺度相关的。虽然它们能有效评价模型，并且广泛用于在同一个数据集上比较不同的方法。但相对于本文试图预测的时间序列的大小来说，MAPE会更有用。MAPE 指得是平均绝对百分比误差，它属于一种相对的度量。实际上MAPE是把MAD 尺度确定成为百分比的单位，而并不确定成变量的单位。平均绝对百分比误差其实属于相对误差度量值，具体来说MAPE是通过绝对值的计算，避免正误差与负误差之间相互抵消的情形，较优的比较出我们关于时间序列所建的模型的预测准确度情况。注意，MAPE的值越小，就意味着我们的预测模型描述实验的数据有着更加优秀的准确度。MAPE公式如下：

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

本模型对2012年全球月度气温预测的MAPE为9.40%，平均绝对百分比误差较小，表明模型预测准确度高。

（三）基于LSTM模型的单年各月气温预测

1. 引入目的

从上述ARIMA算法，可以看出，全球气温在呈现逐年增长的趋势的同时，在每一年内，也呈现出了类似的周期性起伏变化，但由于其受到逐年增长的大趋势的影响，我们并不能对其周期性变化进行较好的分析。

因此为了排除其大环境因素的干扰，只比较一年内的各个月份气温的变化，进一步分析月份与月份之间的联系，同时，还需要通过对以往的同月份数据的学习与分析，对未来进行预测。因此我们引入了长短期记忆LSTM模型，LSTM模型是一种特殊的RNN模型，其优化了RNN模型梯度弥散的问题，即RNN模型只有短时记忆，无法解决长期依赖问题，使得循环神经网络能够真正有效地利用长距离的时序信息。

2. 算法原理

长短时记忆网络模型（下面简称LSTM模型），其在循环神经网络（下面简称RNN模型）的结构基础上，通过增设门限（Gates）使得RNN模型短期记忆的问题得以解决，让循环神经网络能够应用于长距离的时序信息。

其中，LSTM模型增设门限（Gates）包含输入门限（Input Gate）、输出门限（Output Gate）、遗忘门限（Forget Gate）这3个逻辑控制单元，并将其各自连接到了一个乘法元

件上（见图几）。其中，信息流的输入、输出以及细胞单元（Memory cell）的状态，皆可通过设置不同的神经网络的记忆单元与其他部分连接的边缘处的权值来进行控制。

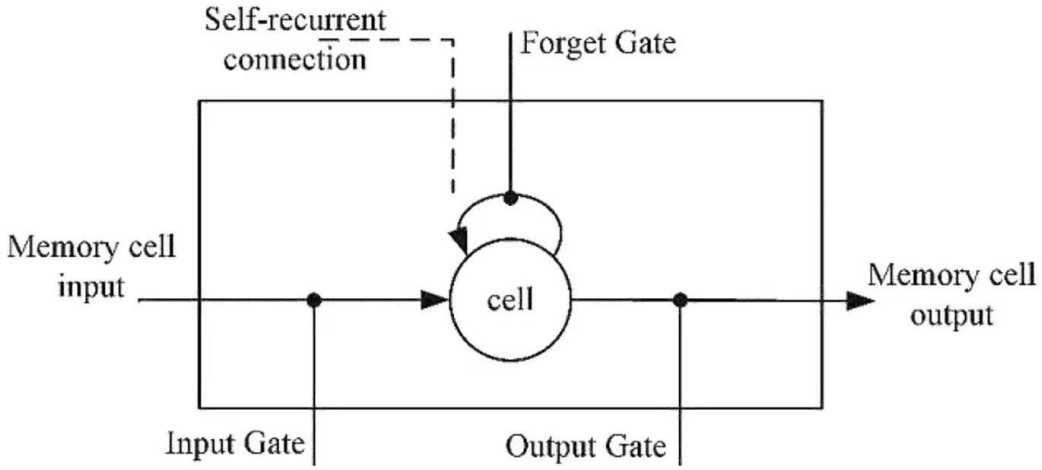

图 9 LSTM 结构图

在图 9 中，各个元件的功能如下：

（1）输入门限 Input Gate：控制信息向 Memory cell 的流入，记为 i_t；

（2）输出门限 Output Gate：控制 Memory cell 中的信息向隐藏状态 ht 的流入，记为 o_t；

（3）遗忘门限 Forget Gate：控制上一时刻 Memory cell 中的信息累积到当前时刻的 Memory cell 中，记为 f_t；

（4）cell：记忆单元，对神经元的状态有记忆功能，实现 LSTM 单元的存取、重置以及更新长距离历史信息的能力，记为 c_t；

由此，可对 t 时刻的 LSTM 神经网络做出如下定义：

$$
\begin{aligned}
f_t &= \mathrm{sigmoid}(W_f \cdot [h_{t-1}, x_t] + b_f) \\
i_t &= \mathrm{sigmoid}(W_i \cdot [h_{t-1}, x_t] + b_i) \\
o_t &= \mathrm{sigmoid}(W_o \cdot [h_{t-1}, x_t] + b_o) \\
\tilde{c}_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
c_t &= f_t \times c_{t-1} + i_t * \tilde{c}_t \\
h_t &= o_t \times \tanh(c_t)
\end{aligned}
$$

其中，W_* 表示对应门限的递归连接权重，sigmoid 与 tanh 表示 2 种激活函数。

在定义了相关参数后，按如下步骤对 LSTM 神经网络进行训练：

STEP1 将 t 时刻的数据特征传入输入层，经过激活函数后得到输出结果；

STEP2 将该输出结果与 t -1 时刻的隐藏层输出信息与存储在 cell 单元内的信息一并作为输入数据，输入到 LSTM 模型的节点中；

STEP3 通过输入门限、输出门限、遗忘门限和 cell 单元的处理，将数据输入到下一隐藏层（输出层）；

STEP4 将 LSTM 模型节点的结果传输给输出层神经元，以此计算反向传播误差并根据公式更新各权值。

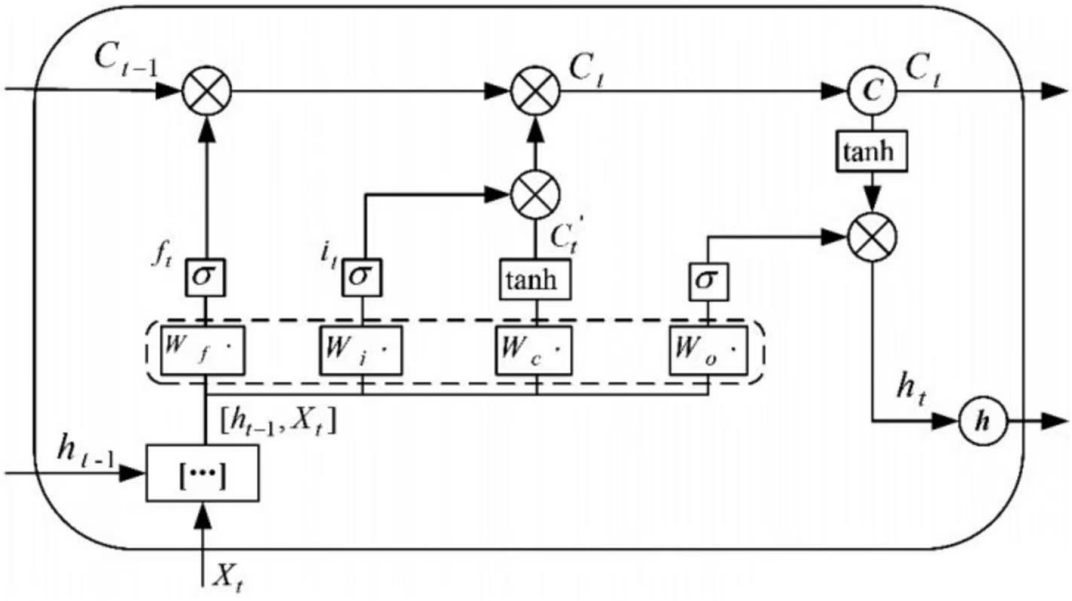

图 10 隐藏层 cell 结构图

3. 建模过程

在明确了 LSTM 模型的基本原理后，我们基于该模型，利用 Python 与 Keras 按如下步骤具体实现对全球气温单年各月份气温的模型构建：

STEP1 数据的导入

该部分主要实现将全球气温数据导入到模型中，以便做进一步的适应化处理与 LSTM 模型构建。

STEP2 数据的模型适应化

鉴于导入的数据无法直接应用到 LSTM 模型上，因此，在构建 LSTM 模型对数据进行学习与进一步分析之前，需要先对数据做如下适应化处理，以便数据能够更好适应模型：

（1）将时间序列转化成监督学习

对于时间序列，我们常把最后一个历史时刻 t-1 的观察值作为输入特征 X，把当前时刻 t 的观察值作为输出特征 Y，以此实现转换。

但对于本问题而言，我们需要实现对一组时间序列数据的转换，所以无法类比真正的监督学习，即使得输入与输出具有明确的一对一映射关系。特别是对于数据集的开始或结尾部分，我们并不能使他们同时在训练集中找到相应的对应关系。

为了解决该问题，对于最开始的输入特征，我们进行置 0 操作，其对应的输出就是时间序列的第一个元素。

（2）平稳时间序列

通过上述两个小节的分析可以看出，全球气温存在逐年上升的趋势，虽然该上升幅度较为缓慢，但仍可以认为该数据在时间序列上是非平稳的，即如果直接使用该数据进行 LSTM 模型的构建，其数据的上升趋势仍然会对针对单年份各个月份的气温数据与变化趋势的分析产生影响，因此，我们对数据进行差分操作，排除该因素对本节分析的影响。

（3）数据标准化

为了提升算法的收敛速度与算法最终效果，我们在数据导入模型前对数据进行标准化操作。同时，对于 2 个激活函数 sigmoid 与 tanh，由于其梯度最大的区间处于 0 值附近，因此当输入值过大或过小时，会很大程度上对这两个激活函数的变化造成影响，即由于数据绝对值较大，sigmoid 的导数 sig（1-sig）将会趋于 0，则使用梯度下降法进行优化的时候，梯度会趋于 0，sigmoid 函数的变化将会变得十分平坦，其优化速度变缓。

进一步对上述文字进行解释说明：由于在初始化操作时，一般使用 0 均值的的正态分布或小范围的均匀分布（Xavier），则必须排除输入数据在尺度上的差异，即不允许存在尺度相差较大的两种及以上的特征，例如（10000，0.001）。此时利用归一化操作则可以很好的消除数据在尺度上的差异，若直接输入数据信息，则将会导致激活函数的输入 $w_1x_1+w_2x_2+b$ 变的很大或者很小，从而致使激活函数梯度趋于 0。

同时，LSTM 模型的默认激活函数是 tanh 函数，tanh 的输出范围为[-1,1]，该范围也是时间序列数据的首选范围。因此，可以使用 MinMaxScaler 类，将数据集的取值范围映射到[-1,1]上。

在具体实现过程中，与 scikit 中其他用于转换数据的方法类相似，我们需要对 MinMaxScaler 类提供矩阵格式的数据信息，因此，在对数据进行转换之前，需要对 numpy 数组进行重塑。

STEP3 构建 LSTM 模型

在算法原理中，我们已经对 LSTM 模型的构建与运行流程做出详细介绍，因此此处不再赘述，具体实现过程详见附录代码部分。值得注意的是，在设置输入数据矩阵时，LSTM 层要求应以以下格式进行输入：

LSTM 层要求输入矩阵格式为：[样本，时间步长，特征]

由于训练数据集定义为 X 输入和 Y 输出的形式，因此在具体代码实现时，我们需要先将其转化为“样本/时间步长/特征”的形式。

长短期记忆网络 LSTM 模型作为递归神经网络 RNN 模型的一种特殊算法分支，该类网络的优点是它能学习并记住较长序列，且不依赖预先指定的窗口滞后观察值作为输入。 在 Keras 中，该优点被称为“stateful”，而在具体定义 LSTM 网络层时，使用“True”对“stateful”语句进行设定。

STEP4 LSTM 模型的效果评估

在构建 LSTM 模型后，仍需要对本模型的拟合预测效果进行评估。首先，我们对预测值与实际期望值曲线进行绘制，得到一个较为直观的对比图，并在此基础上，使用均方根误差 RMSE 对模型的预测误差进行量化，并以此为模型预测效果的评估标准。

RMSE 是对 MSE 的进一步优化，其可以使得数据在数量级上更为直观。我们通过如下公式得到最终建模结果的 RMSE 值：

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

其中，我们假设：

$$\text{预测值: } \hat{\mathbf{y}} = \{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n\}$$
$$\text{真实值: } \mathbf{y} = \{y_1, y_2, ..., y_n\}$$

RMSE 取值范围为[0,+∞)，其中，当所有预测值与对应的真实值完全吻合时，RMSE 为 0，即表明该模型可以完美对数据进行预测；而当模型预测效果与真实情况差距越大时，RMSE 也越大。

4. 仿真结果及分析

在对全球温度进行基于 LSTM 模型的算法建模后，我们得到如下训练与预测结果，其仿真结果如下：

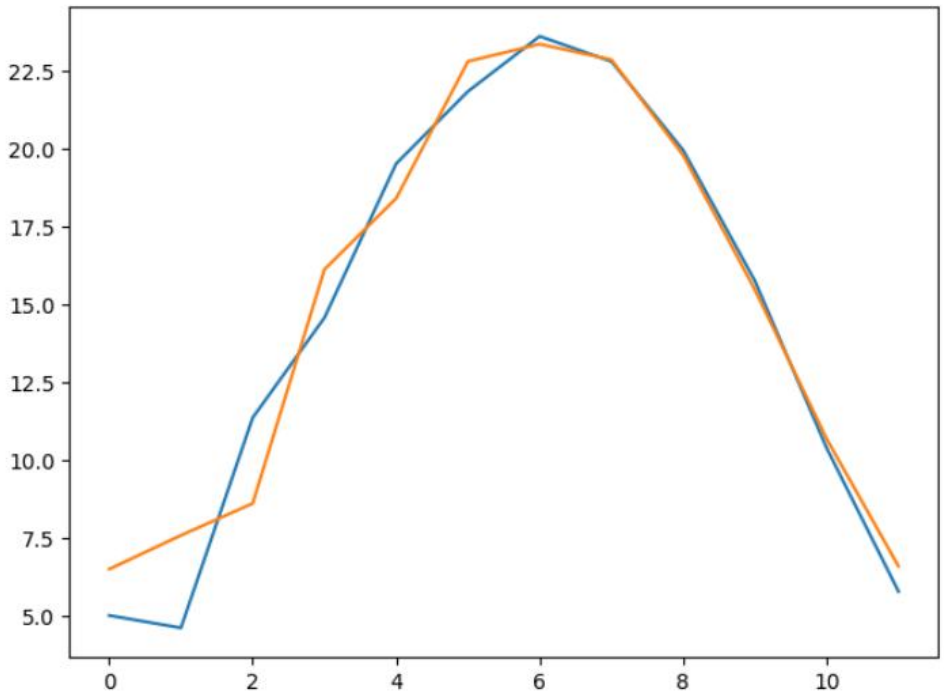

图 11 全球气温年度变化预测与实际值对比图

其中黄线为预测值，蓝线为实际值。

同时，得到如下各月气温的预测值与实际值的具体数据：

表 8 全球气温年度 12 月变化预测与实际值对比

Month	Predicted	Expected
1	6.515437	5.03
2	7.60165	4.63
3	8.627146	11.38
4	16.141693	14.59
5	18.426664	19.54
6	22.824744	21.86
7	23.377938	23.63
8	22.879059	22.82
9	19.805027	19.97
10	15.484918	15.77
11	10.697593	10.4
12	6.610937	5.8

Test RMSE=1.418

结合以上图标，我们可以看到，对于年度全球气温变化的实际值来说，基于 LSTM 模型进行拟合的预测曲线效果较好，基本与实际期望相贴合，两者总体都呈现出 1 月至 6 月上升，6 月达到气温峰值，6 月至 12 月下降的类二次函数曲线。

对于峰值（6 月）之前，其预测仍存在一定的抖动，但预测数据走势与实际期望值一致，均呈现上升趋势，从 1 月的谷值（预测值为 6.515437，实际期望值为 5.030000）上升到 6 月的峰值（预测值为 22.824744，实际期望值为 21.860000），且峰值点数据拟合差异较好；而在峰值点后，两者的变化曲线则基本重合，并呈现出平滑的下降趋势，从 6 月的峰值下降到 12 月的谷值（预测值为 6.610937，实际期望值为 5.800000）。

同时，我们将 LSTM 模型的预测值与实际期望值之间的差异通过均方根误差 Test RMSE 进行量化，得到的误差为 1.418，即表明 LSTM 模型计算出的预测效果相比与真实值平均相差 1.418，均方根误差较小，进一步说明基于 LSTM 模型对于全球气温单年各月的气温预测效果良好。

四、结论

宏观上，对全球气温以年为基本单位进行考察，根据 SESP 模型与 GF 模型的分析结果可以得出，其中 SESP 模型的二次指数模型因预测不符合实际，我们忽略其预测结果，从 SESP 三次指数模型与 GF 得出的预测不难看出，虽然这两个模型预测的数值有一些差异，但全球年平均气温的增长率差异不大，证明全球年平均气温总体呈增长趋势，且增长速率较快。

微观上，将全球气温每年数据拆解成逐月进行考察：首先以数据集涉及的全时间段（2001-2012 年）作为参考范围，并基于 ARIMA 算法进行预测，观察到全球气温除宏观上的增长趋势，全球年平均气温在各个月均在零度以上外，每一年的各个月份的气温都呈现出相似的周期性变化趋势。

因此我们由进一步使用 LSTM 模型，将一年的各个月份的气温进行细化预测，并排除了宏观上气温增长的干扰，最终，得出全球平均气温在每年 1 月与 12 月相对较低，而 6 月份则达到峰值，总体呈现先上升，后下降的趋势。

五、模型现实意义与未来工作

本次使用的模型都可以根据时间序列进行对未来全球气温进行预测，根据不同的模型，可以预测出年平均气温的走向趋势以及具体未来一年的月平均气温，且整体模型是一步步精细的，形成了合理的预测体系。在全球变暖的今天，预测未来全球气温走向是十分重要的，可以让人们直观的看到全球变暖的数据趋势，引起人们对环保、节能减排的重视。

不仅在全球气温方面，按照年、月份的其他时间序列也可以利用本模型对未来进行预测，例如降雨量、二氧化碳排放量等等，但根据不同问题，具体的参数与数据需要进行调整，才能更精准的预测，这也是我们未来工作的主要部分。

不仅在此方面，在未来，我们也可以将文中使用的几个模型进行整合、汇集，将模型通过代码书写汇成网页或者软件，可以更便利的让人们对自己的数据进行预测，也可以省去来回改代码的麻烦。

六、参考文献

[1]李志超,刘升.基于ARIMA模型、灰色模型和回归模型的预测比较[J].统计与决策,2019,35(23):38-41.DOI:10.13546/j.cnki.tjyjc.2019.23.007.
[2]沈露露,梁嘉乐,周雯.基于ARIMA-LSTM的能量预测算法[J/OL].无线电通信技术:1-8[2023-01-07].http://kns.cnki.net/kcms/detail/13.1099.TN.20230106.0919.004.html
[3]韩金磊,熊萍萍,孙继红.基于LSTM和灰色模型的股价时间序列预测研究[J/OL].南京信息工程大学学报(自然科学版):1-22[2023-01-07].http://kns.cnki.net/kcms/detail/32.1801.N.20230105.1635.003.html
[4]李颖若,韩婷婷,汪君霞,权维俊,何迪,焦热光,吴进,郭恒,马志强.ARIMA时间序列分析模型在臭氧浓度中长期预报中的应用[J].环境科学,2021,42(07):3118-3126.DOI:10.13227/j.hjkx.202011237.
[5]赵欢. 基于时间序列模型与灰色模型的广东省旅游人数预测研究[D].华南理工大学,2019.DOI:10.27151/d.cnki.ghnlu.2019.001478.