中国研究生创新实践系列大赛

“华为杯”第十六届中国研究生数学建模竞赛

数学建模竞赛

学　　校	南京大学
参赛队号	**19102840016**
队员姓名	**1.** 孙东
	2. 张纪康
	3. 王兆君

中国研究生创新实践系列大赛

“华为杯”第十六届中国研究生数学建模竞赛

数学建模竞赛

题　目　　全球变暖气候预测分析

摘　　要：

近年来，全球极端天气频发引起了人们对全球变暖现象的新思考。在此背景下，本文主要研究了全球变暖的诸多问题，广泛收集了各类相关数据，并进行数据处理，综合运用了 Mann-Kendall 突变检验、小波分析、时间序列预测模型、随机森林分类模型等统计与算法知识建立了相关问题的数学模型，并利用 Python、ArcGIS、Tableau 等软件得出了较为合理的结论。

针对问题一中（1）加拿大温度时空变化趋势：为保证时间序列的连续性以及观测数据的空间分布合理性，最终提取了 299 个观测站点数据，数据预处理后对加拿大温度进行分区域、分季节的时空趋势分析。时间变化分别从趋势、突变和周期三个维度，运用气候倾向率、Mann-Kendall 突变法和小波分析建模分析，最终结论是加拿大地区温度呈现上升趋势，温度上升过程中发生了由低到高的突变，且存在 5 年的主周期变化；空间变化采用反距离权重法插值分析，**结果表明加拿大地区平均温度在空间分布上由东北向西南和东南逐渐升高，总体出现升温趋势，其中升温最显著的地区发生在魁北克省的西南部和安大略省东南部。**（2）海洋表面温度的规律探究：文章构建了分布函数对海洋表面温度数据进行年代际特征分析。分析结果显示**在 19 世纪 60 年代初至 20 世纪 10 年代末，海洋表面温度略有降低，自 20 世纪 20 年代开始，海洋表面温度开始升高，总体呈现为升温趋势。**

针对问题二，首先，从温度、辐射、温室气体、大气动力、震荡特征五个层次出发，以与全球温度特征显著相关为准则，选取用于气候预测的 9 个因素；其次，根据不同因素在时间序列维度的对应情况，选取 1948 年到 2018 年的因素数据进行统计；然后，对数据进行缺失值和异常值预处理，并分析因素之间的相关性，运用主成分分析法，在保证 92% 的累计方差贡献率的前提下将 9 个因素降至 4 维；紧接着，运用随机森林回归预测模型对降维后的因素进行建模，测试集结果表明模型的调整R^2值达到了 0.966，预测效果良好，模

型同时求出了不同因素对气候变化的影响情况，其中**海洋表面温度对气候的影响较大，此外二氧化碳仍然是造成全球变暖的重要因素**；接下来，运用 ARIMA 自回归时间序列预测模型及基于 Prophet 框架的可加时间序列预测模型对未来 25 年的 9 个因素值进行了预测，有效的弥补了单一时间序列预测模型的不足；最后，利用训练好的随机森林回归预测模型实现了对未来 25 年全球气候的预测，预测结果显示 **2019-2021 年维持了前几年的下降趋势，2021 年之后温度开始升高，总体保持“变暖”的趋势，并于 2041 年达到该预测区间的最大值 15.11 摄氏度**。

针对问题三，首先，极寒天气定义为零下 40 度以下的天气，但考虑到温度的高低的相对性，故将研究区域选定为有一定概率发生极寒天气的中低纬度地区。本文选取的纬度范围为 40N～60N；然后，统计该区域 1948 年到 2018 年的极寒天气出现频次；紧接着，为了描述极寒天气和气候变化的内在关系，以每年的极寒天气出现频次为标签，以问题二中的因素数据为特征，建立随机森林分类模型并进行训练，模型在测试集上针对极寒天气的召回率值达到了 90%，说明模型很好的识别了极寒天气，同时模型也求出了不同因素对极寒天气的影响情况，其中**海洋表面温度及北极涛动指数对极寒天气有较大影响，同时也说明了极寒天气和气候变化存在内在关联**；接下来，运用 Pearson 相关模型刻画海洋温度和全球温度的相关关系，运用 Spearman 相关模型分别刻画海洋温度和北极涛动指数、北极涛动指数和极寒天气的相关关系；最终，得到问题三结论：**全球变暖的同时，海洋由于吸收大量热量温度也在上升，进而造成了北极涛动转向负位相，使得极寒天气出现的概率增大，因此，全球变暖与局部极寒天气并不矛盾。**

针对问题四，首先根据问题三的结论，用通俗易懂的文字解释了“全球变暖了，某地今年的冬天特别冷”之间的关系；然后，从趋势性和复杂性角度，提出了**多维度全球“变暖”**的新概念。

关键词：小波分析；ARIMA 自回归；Prophet 框架；随机森林

目 录

一、问题重述

1.1 问题背景

全球气候变暖的解释是由于温室效应不断积累所致。事实上，由于人们焚烧化石燃料，如石油、煤炭等，或砍伐森林并将其焚烧时会产生大量的二氧化碳，即温室气体，这些温室气体对来自太阳辐射的可见光具有高度透过性，而对地球发射出来的长波辐射具有高度吸收性，能强烈吸收地面辐射中的红外线，使得地球温度上升，即温室效应。由于存在温室效应，影响地气系统吸收与发射的能量平衡，能量不断在地气系统累积，从而导致温度上升，造成全球气候变暖。许多科学家认为，全球变暖可能导致更多的极端气象的产生，导致全球降水量重新分配、冰川和冻土消融、海平面上升等威胁人类生存的因素。不过，虽然温室气体的浓度在不断上升，但自从进入 21 世纪以来，10 年间全球全年平均气温上升率仅为 0.03℃，几乎未变化，这种现象叫作 Hiatus（全球变暖停滞状态）。正因为出现全球变暖停滞现象，使公众对全球变暖产生了怀疑。 2019 年 1 月美国 2/3 的地区变成了一个大冰窖，出现了“几十年一遇”的极度寒冷天气，成为有人怀疑全球变暖的依据之一。

导致分歧的原因在于观察问题的角度和范围。今年的夏天特别热或今年的冬天特别冷是地球上局地人们的直接感受，是一种天气现象。天气是一定区域短时段内的大气状态（如冷暖、风雨、干湿、阴晴等）及其变化的总称。而气候则是长时间内气象要素和天气现象的平均或统计状态，时间尺度为月、季、年、数年到数十年。气候是长时间的平均状态，在短时间内变化不大，所以人们一般感受不到。全球变暖是在气候尺度上看全球问题。从气候角度研究全球温度变化需要全球范围长时间的观测积累，但过去这方面的时空数据并不完整，给统计计算带来极大困难。不仅如此，海洋吸收热量对全球气候变化的影响很大。观测发现海洋表面温度的变化具有某种震荡特征，如年代际太平洋震荡、厄尔尼诺现象、拉尼娜现象等。这些因素使得研究全球温度变化更加困难。

命题宗旨是：利用现有的统计数据建立简化的气候模型和极端天气模型。所建立的模型区别于复杂的专业气候模型，有利于非专业人士理解和认识全球气候变化的态势，解释极端天气现象的发生，寻找、求证影响气候变化的因素，从而增强人们气候变化的意识，从现在做起、从自我做起的同时，督促决策者迅速制定应对气候变化的政策。

1.2 本文拟解决的问题

（1）能否从加拿大各地天气变化的历史数据中挖掘出该地区温度的时空变化趋势？海洋表面温度历史数据中蕴含着什么样的规律？

（2）建立一个刻画气候变化的模型对未来 25 年的气候变化进行预测，该模型至少需要考虑地球的吸热、散热以及海洋的温度变化等要素？

（3）“极寒天气”是某地的天气现象，这种极端气象的出现，与气候变化有无关系？请建立相应的模型，并利用题目所提供的数据以及你能收集的数据说明：全球变暖和局地极寒现象的出现之间是否矛盾？

（4）请用通俗易懂的文字解释：“全球变暖了，某地今年的冬天特别冷”之间的关系。请用一个新概念替代“全球变暖”，来反映气候变化的趋势和复杂性？并给予解释。

二、模型假设与符号说明

2.1 模型的基本假设

根据全球气候情况和文中所给条件，本文做出如下假设：

（1）假设 2019 年开始的未来 25 年内影响气候的人为因素不会出现过大的变异。

（2）假设 2019 年开始的未来 25 年内影响气候的非人为因素不会出现过大的变异。

（3）假设查找所得数据具有一定的可信性与合理性。

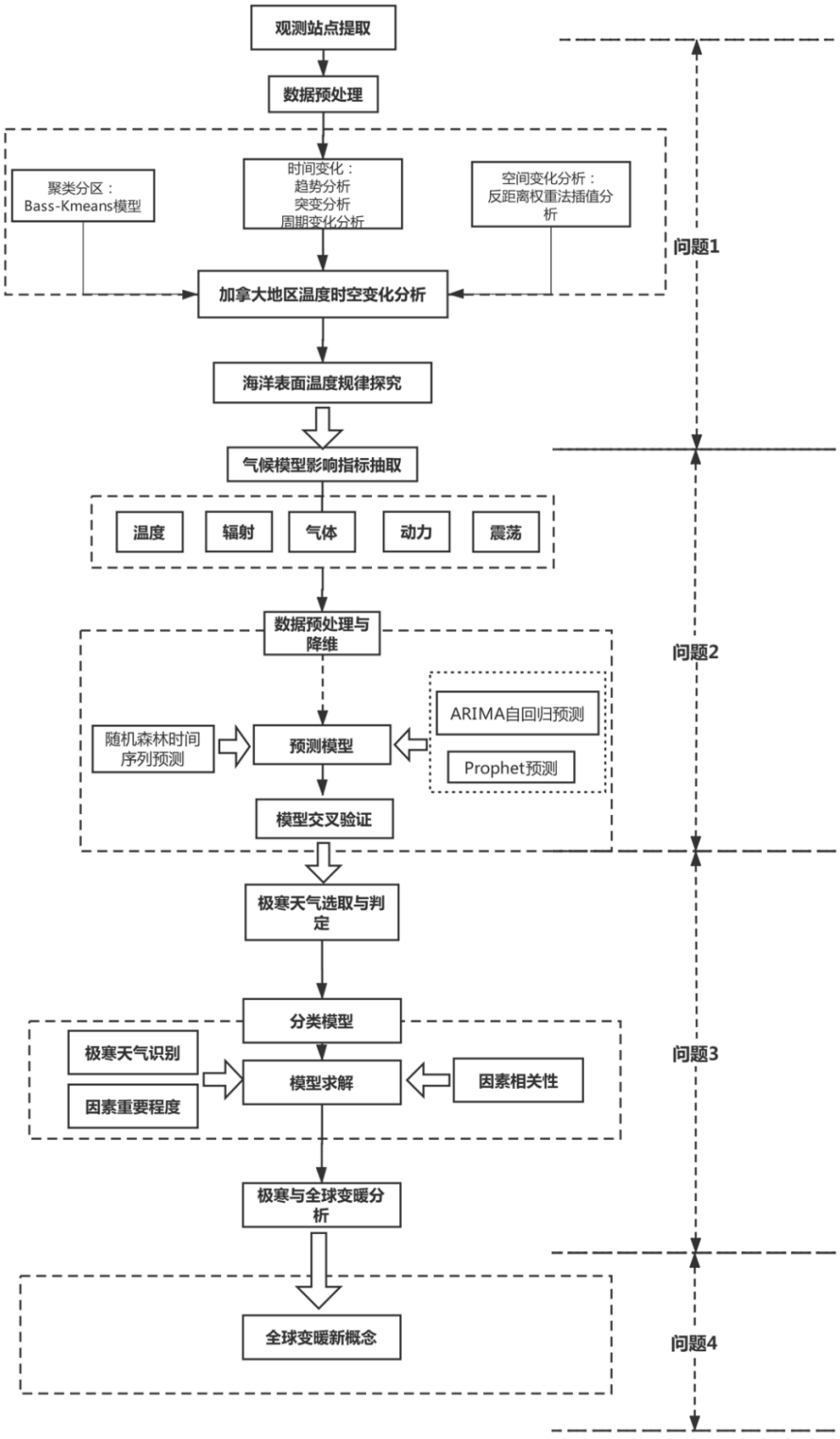
2.2 模型符号说明

表 2-1 符号说明

序号	符号	说明
1	a	气候倾向率
2	t	时间序列
3	α	显著性水平
4	$E(S_k)$	秩序数列S_k的均值
5	$\sqrt{Var(S_k)}$	秩序数列S_k的方差
6	$W_y(a,c)$	小波变换系数
7	c	平移参数
8	γ_i	自回归模型AR的自相关系数
9	ϵ_t	自回归模型的误差
10	g(t)	prophet 的趋势项
11	k	趋势模型的增长率
12	p_k	随即森林模型选中的样本属于 k 类别的概率

三、技术路线图

本文技术路线图如下：



四、问题一：模型的建立与求解

4.1 问题分析

问题一分为两小问，第一问需要从加拿大各地天气变化的历史数据中挖掘出整个地区温度的时空变化趋势，第二问需要从海洋表面温度历史数据中探究蕴含的规律。在加拿大政府气象网站[1]搜集了大量的温度观测数据，以及提取赛题附件中所给的海洋表面温度数据，对数据中存在个别数值异常、缺失等值进行预处理，进而对加拿大地区温度进行时空趋势分析，以及海洋表面温度历史规律的探究。

4.2 加拿大温度时空变化分析

本文在加拿大政府气象网站[1]中搜集了大量的温度观测数据，为保证时间序列的连续性以及观测数据的空间分布合理性，最终提取了 299 个观测站点数据，对加拿大温度进行分区域、分季节的时空趋势分析。

4.2.1 数据处理

1)　观测数据的选取

根据加拿大政府气象网站[1]数据显示，加拿大共有 3000 多个观测站点，分别分布在加拿大 13 个省及地区。由于部分观测站点在时间序列上存在较多数据的缺失，因此，为保证时间序列的连续性，本文对 3000 多个观测站点的数据进行交集处理，最终选取了 299 个站点 2000 年-2018 年的月度温度数据，进行时空趋势的挖掘。

最终提取的 299 个观测站点均匀地分布在加拿大 13 个省及地区中，其数据具有代表性，具体分布位置如图 4-1 所示。

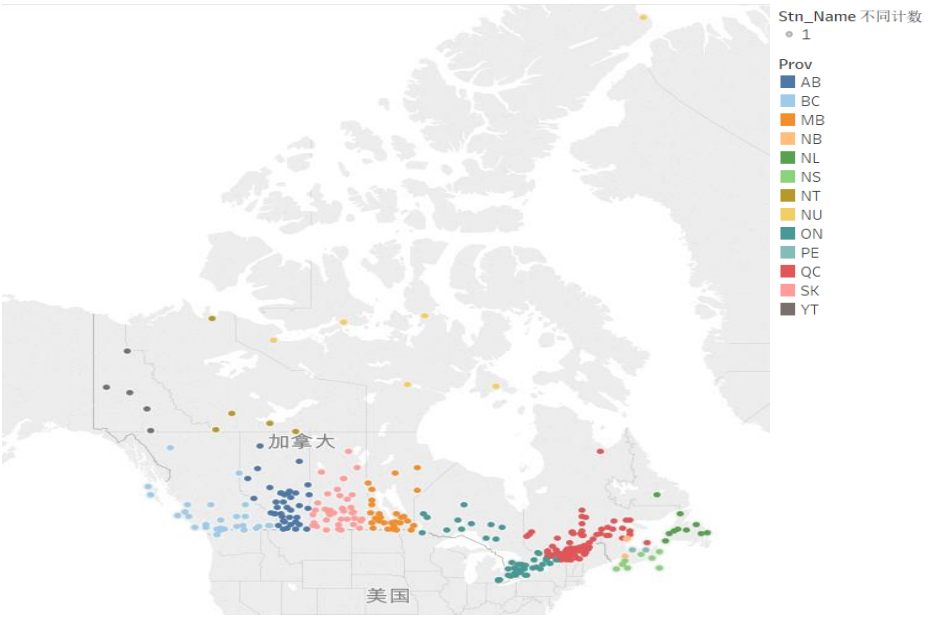


图 4-1　299 个观测站点的地理位置分布

2) 异常值识别

从所提取的数据来看，个别数据存在异常和缺失，为了保证分析结果的科学合理，必须对异常数据进行识别剔除，以及缺失数据的插值处理。

本文采用 Smoothed z-score 方法对异常数据进行识别，该方法的主要思想是：在一段历史时间序列中，利用平均值、方差信息对下个节点值进行预测，根据其位于可接受域的上方还是下方分别标记为 1、-1，在可接受域内标记为 0，若是其超过了一定的阈值，则被识别为异常点。

Smoothed z-score 方法的主要优点是在识别过程中对异常点的数值进行了平滑修正，降低了当前异常值对平均值、方差的影响，以便准确评估下个节点值是否为异常点。

以 BARWICK 观测站 2000-2018 年的 1 月份时间序列异常值识别为例，识别结果见图 4-2。

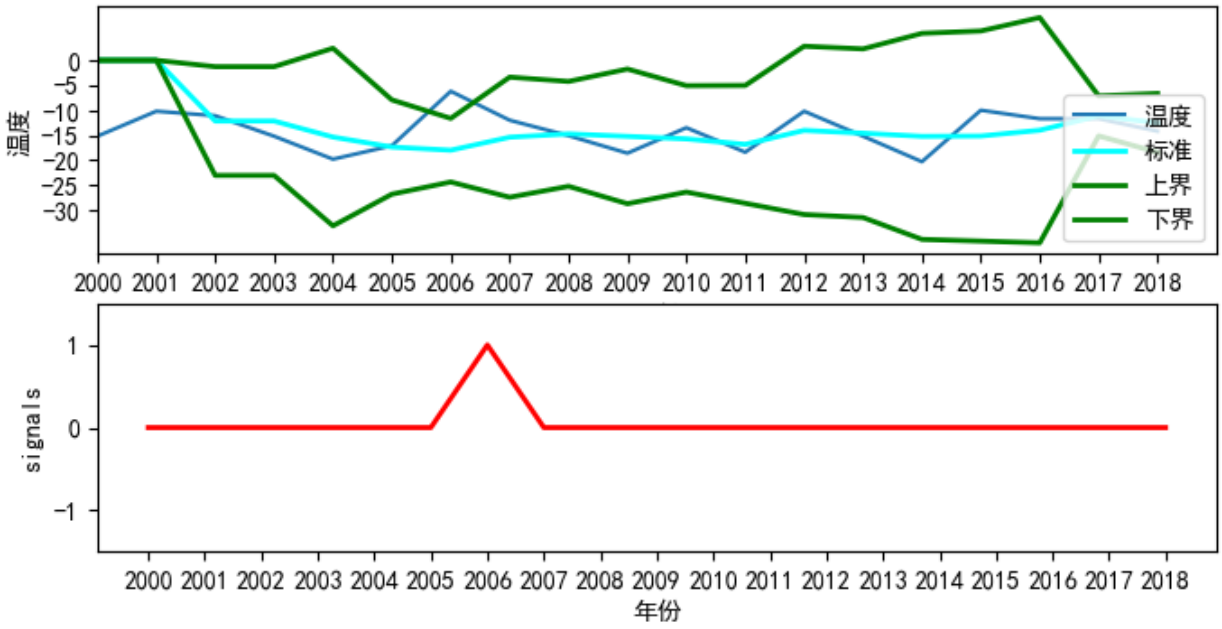


图 4-2 BARWICK 观测站 2000-2018 年 1 月份时间序列异常值识别结果

由图 4-2 可知，2006 年 1 月的温度观测值超过了可接受的范围，被识别为异常值。识别出的异常值先剔除，后作为缺失值一并插值处理。

3) 缺失值处理

由于选取的站点数据存在着一定的缺失值，在建模前先对缺失值进行处理。常见的处理方法有直接删除法、移动平均法和指数平滑法等。在时间序列数据中，如果直接删除缺失值，则会对后续的处理带来麻烦，而指数平滑的方法适用于变化较为明显的数据。移动

平均法是用一组实际数据值来预测未来一期或几期的数据，适用于既不会快速增长也不会快速下降的数据。温度数据恰好符合此特征，

因此，本文选取的数据缺失值处理方法为“移动平均法”。在存在季节因素时，移动平均法的预测值会受到一定影响。为了避免季节因素的影响，本文的处理方法为利用同一时期不同年份的数据做缺失值的填充。比如某年份某月的温度数据有缺失，则利用邻近几年相同月份的数据来预测该缺失年份的数据。

4.2.2 模型建立与求解

1) 观测区域聚类划分

为了更好地分析加拿大不同区域之间的温度时空趋势，从而科学合理的反映加拿大整体的变化趋势，本文对 297 个观测站点的温度变化趋势进行聚类，将加拿大整体划分为 5 个区域。

聚类分析的步骤如下：

Step 1：设定聚类类别数为 5；

Step 2：基于 Bass-Kmeans 模型,利用最小二乘法对原始数据拟合趋势

Step 3：利用 kmeans 算法对拟合的趋势线进行聚类，得到 5 类地区的内部特征。

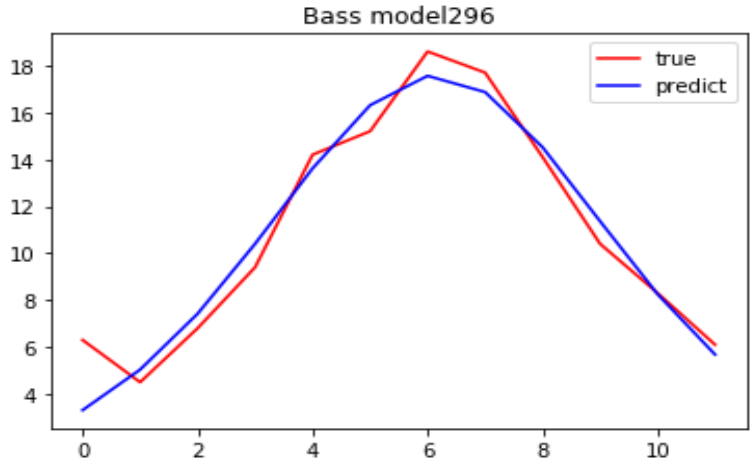


图 4-3 “SALMON ARM CS”观测站点的最小二乘趋势拟合

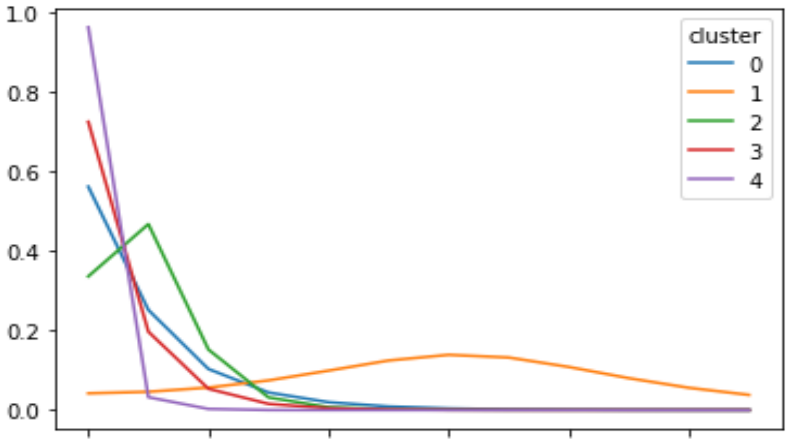


图 4-4　5 个地区的内部特征刻画

观测区域聚类结果如图 4-5 所示，结合加拿大气候分布，本文发现聚类结果与气候分布结构保持一致。加拿大的南部区域 1 为图中蓝色标示，主要为高山气候与温带大陆性气候，紫色标示的中部区域 2 分布较广，为温带大陆性气候，太平洋海岸区域 3 属于温带海洋性气候，而分布在大西洋的区域 4 属于亚热带季风性气候，区域 5 分布在北部地区，包含北极群岛，主要为极地气候。

因此，该聚类结果科学合理，具有较高的可信度。

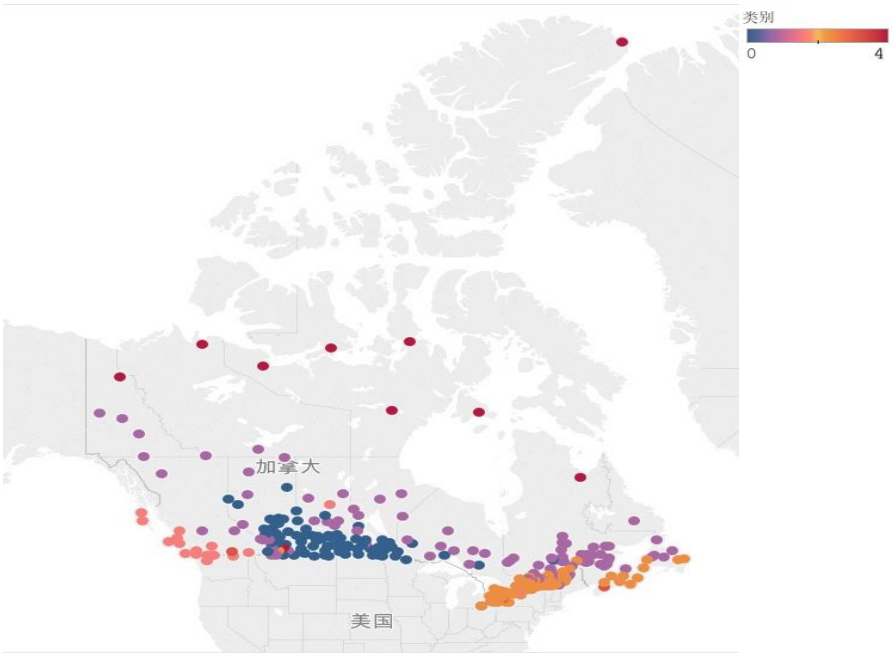


图 4-5 观测区域聚类结果

区域的观测站点聚类结果如表 4-1 所示，完整的聚类结果详见附件。

表 4-1 观测站聚类结果

区域	观测站点
区域 1	MOOSE JAW CS、ESTHER 1、CHAPLEAU A、BROOKS、ELKHORN 2、EAST、GRANDE PRAIRIE A、EDMONTON INTERNATIONAL CS、SCOTT CDA、MORDEN CDA CS、CORONATION CLIMATE、EAR FALLS (AUT)等 82 个观测站
区域 2	ST THOMAS WPCP、ST-ANICET 1、DRUMMOND CENTRE、WHITECOURT A、FUNDY PARK (ALMA) CS、TORONTO BUTTONVILLE A、BONAVISTA、CARIBOU POINT (AUT)、ANGERS、LEMIEUX 等 87 个观测站
区域 3	AMHERSTBURG、PEMBERTON AIRPORT CS、CATHEDRAL POINT (AUT)、PITT MEADOWS CS、GREENWOOD A、WHITE ROCK CAMPBELL SCIENTIFIC、KEJIMKUJIK 1、SARTINE ISLAND (AUT)等 29 个观测站
区域 4	THE PAS A、CAP-MADELEINE、TWILLINGATE (AUT)、CAP-CHAT、HAY RIVER A、CAUSAPSCAL、TESLIN (AUT)、GERALDTON A、PRINCE ALBERT A、BOW VALLEY、VAUXHALL CDA CS、NICOLET、KAPUSKASING A 等 92 个观测站
区域 5	PAULATUK、KUGLUKTUK A、TALOYOAK A、CAMBRIDGE BAY A、ALERT UA、BAKER LAKE A、CORAL HARBOUR A 等 9 个观测站

2) **时间变化分析**

完成聚类分区后，本文对 5 大观测区域 299 个站点全年平均温度和四季平均温度进行时间层面的变化分析，主要包括趋势分析、突变分析和周期变化分析，分析结果用 Python 实现。季节划分为：春季（3-5 月）、夏季（6-8 月）、秋季（9-11 月）、冬季（9-翌年 2 月）。

a) 趋势分析

利用线性倾向估计法分析温度时间序列的变化趋势，建立时间序列t与温度y之间的一元线性回归方程：

$$y = at + b$$

其中t为时间序列，a为气候倾向率，表示气候要素温度每年增加或减少的量。

首先对加拿大整体进行趋势分析，结果如下表所示。

表 4-2 加拿大温度趋势变化

区域	温度因素	多年均值（℃）	气候倾向率（℃/a）	模型 R^2
加拿大	春季平均温度	-0.063	0.020	0.012
	夏季平均温度	14.879	0.038	0.272
	秋季平均温度	3.665	0.017	0.026
	冬季平均温度	-11.133	0.019	0.010
	年平均温度	1.837	0.023	0.052

整体来看，加拿大年平均温度气候倾向率为 0.023℃/a，整体全年呈现上升趋势。四季的平均温度倾向率为夏季（0.038℃/a）>春季（0.020℃/a）、冬季（0.019℃/a）>秋季（0.017℃/a），全年气温上升主要是由于夏季升温趋势更为明显。

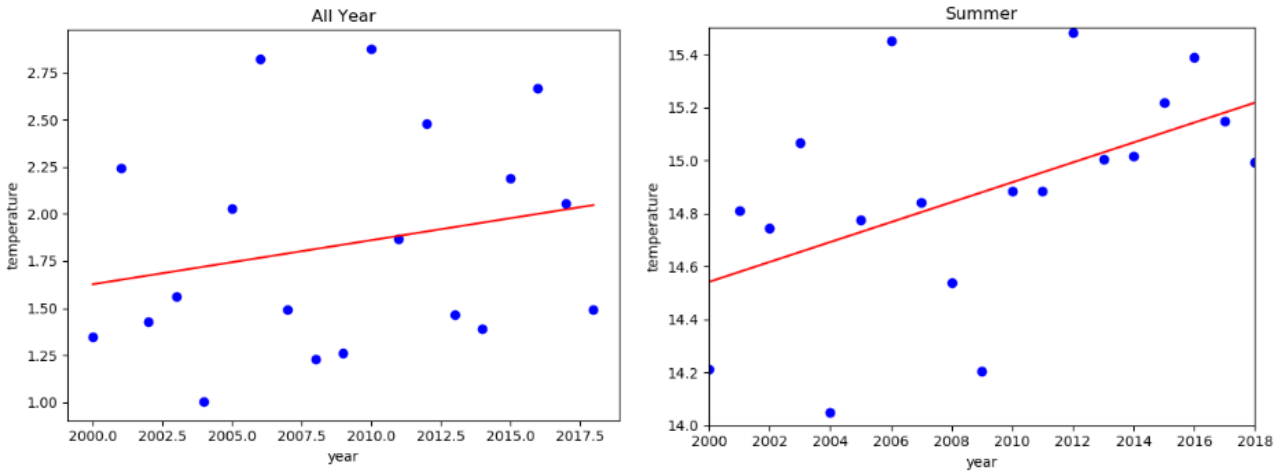


图 4-6 加拿大全年和夏季平均温度回归拟合图

其次对 5 大区域进行趋势分析，结果如表 4-2 所示。

表 4-3 加拿大各区域温度趋势变化

区域	温度要素	多年均值（℃）	气候倾向率（℃/a）	模型 R^2
区域 1（南部）	春季平均温度	2.726	0.02	0.004
	夏季平均温度	16.904	0.038	0.079

续表 4-3

	秋季平均温度	4.192	-0.005	0.001
	冬季平均温度	-12.145	0.003	0.009
	年平均温度	2.919	0.014	0.009
区域 2（中部）	春季平均温度	5.117	0.011	0.003
	夏季平均温度	18.565	0.028	0.073
	秋季平均温度	8.668	0.025	0.001
	冬季平均温度	-6.449	0.048	0.038
	年平均温度	6.475	0.028	0.060
区域 3（太平洋海岸）	春季平均温度	8.431	0.045	0.104
	夏季平均温度	16.405	0.046	0.154
	秋季平均温度	10.156	0.04	0.262
	冬季平均温度	3.117	-0.021	0.019
	年平均温度	9.527	0.027	0.097
区域 4（大西洋海岸）	春季平均温度	1.087	0.003	0.015
	夏季平均温度	15.616	0.036	0.212
	秋季平均温度	4.101	0.004	0.001
	冬季平均温度	-12.230	0.016	0.005
	年平均温度	2.144	0.015	0.016
区域 5（北部）	春季平均温度	-17.677	0.018	0.005
	夏季平均温度	6.907	0.04	0.091
	秋季平均温度	-8.791	0.022	0.011
	冬季平均温度	-27.958	0.051	0.030
	年平均温度	-11.880	0.033	0.031

分区域看，加拿大 5 个区域内年平均温度气候倾向率均为正值，说明年平均温度呈现上升趋势。其中区域 5 即北部地区年均温度上升幅度较大，每年上升 0.033℃，主要原因是北部区域冬季平均温度气候倾向率较高为 0.051℃/a，提升了全年升温幅度；加拿大中部的区域 2 和太平洋海岸区域 3 年均温度气候倾向率较为接近，分别为 0.028℃/a和 0.027℃/a；区域 1 和区域 4 年均温度分别上升 0.014℃和 0.015℃，夏季气温升高是导致两个区域全年平均气温增加的主要原因。

同时，上述模型拟合的 R^2 均较小，达到显著性水平。

b)　突变分析

气候突变是气候从一种稳定态跳跃式地转变到另一种稳定态的现象，主要表现为气候要素统计特性在时间上的不连续性[2]。本文中采用 Mann-Kendall 突变法（M-K 检验）对加拿大的温度进行突变研究。

模型建立的具体步骤如下：

对含有n个样本的温度时间序列t，构造一秩序数列：

$$S_k = \sum_{i=1}^{k} p_{ij} \quad (k = 2,3,\dots,n)$$

其中，

$$p_{ij} = \begin{cases} 1, & t_i > t_j \\ 0, & t_i \le t_j \end{cases} \quad (j = 1,2,\dots,i)$$

定义统计量：

$$UF_k = \frac{S_k - E(S_k)}{\sqrt{Var(S_k)}} \quad (k = 2,3,\dots,n)$$

其中，$E(S_k)$为S_k的均值，$\sqrt{Var(S_k)}$为S_k的方差，UF_k为标准化后的S_k统计量。再构造温度时间序列t的逆序列，重复以上过程得到统计量UB_k。给定显著性水平$\alpha = 0.05$时，$\mu_{0.05} = 1.96$。

若$|UF_k| > \mu_{0.05}$，则表明样本序列存在明显的趋势变化；若UF_k和UB_k统计量曲线出现交点，且交点位于置信区间内，那么交点对应的时刻就是突变开始的时刻。

首先对加拿大整体进行 M-K 突变检验，结果如下图所示。

加拿大年均温度整体呈先降低后升高的变化趋势，2000-2004 年间平均温度特征曲线UF呈波动下降，随后开始波动上升，2017-2018 年开始出现下降趋势。在 2016 年、2017 年和 2018 年发生了三次突变。四季平均温度变化趋势与全年保持一致，均呈先降低后升高的

变化趋势，夏季变化趋势最为显著，其中冬春两季无突变，夏秋两季均发生三次突变，且秋季突变时间早于夏季。

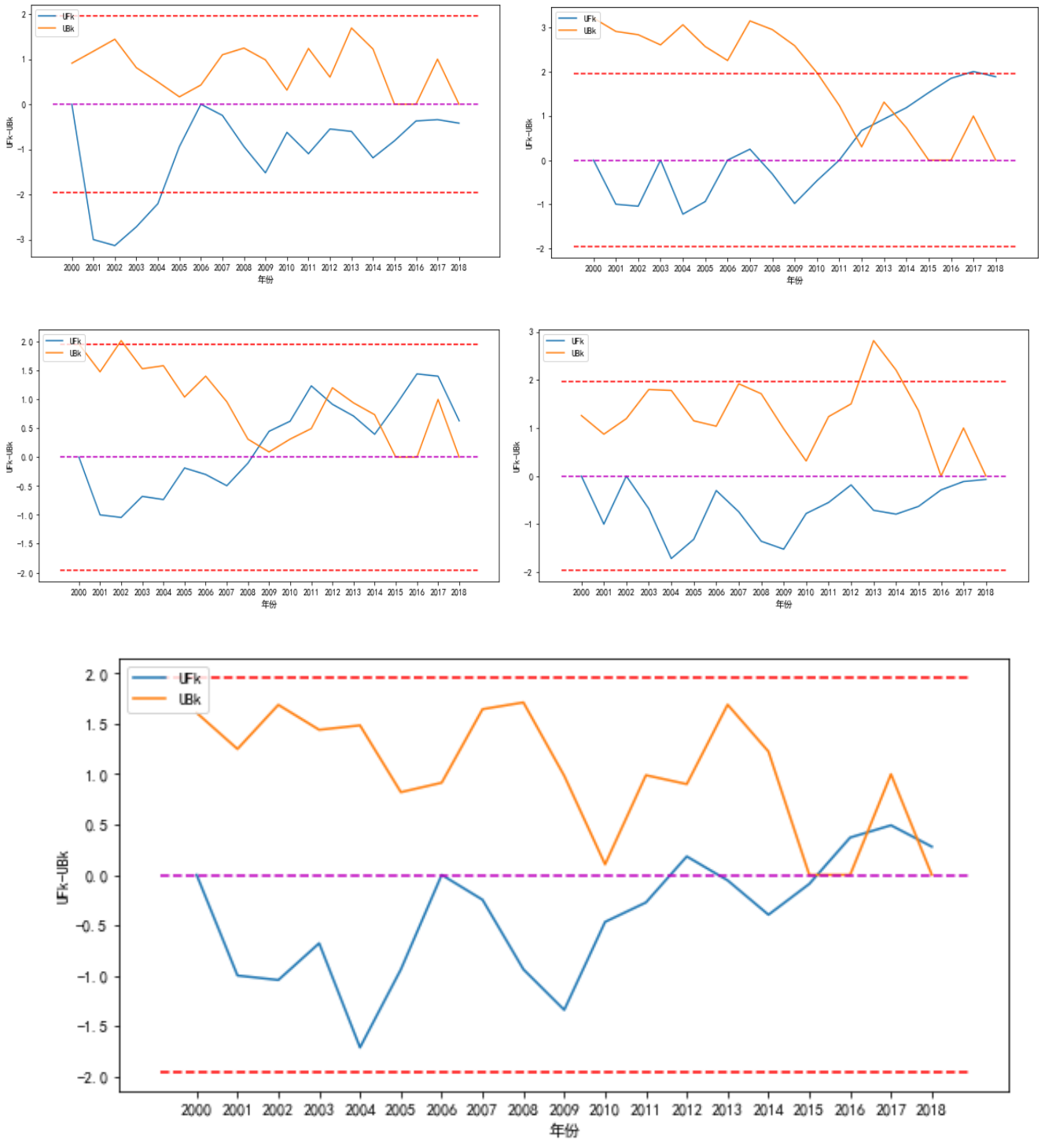


图 4-7 加拿大全年和四季平均温度 M-K 突变检验结果

其次对加拿大 5 大区域分别进行 M-K 突变检验，结果汇总如下表所示。

表 4-4 加拿大各区域 M-K 突变检验结果

区域	温度要素	趋势变化	突变年份
区域 1（南部）	春季平均温度	-+-	/
	夏季平均温度	+	2018
	秋季平均温度	-+-	2009、2010、2015、2017
	冬季平均温度	-+	/
区域 2（中部）	春季平均温度	-+	2010、2011、2012、2013
	夏季平均温度	-+	2012、2013、2018
	秋季平均温度	-+	2016、2017
	冬季平均温度	-+	2016、2018
区域 3（太平洋海岸）	春季平均温度	-+	2018
	夏季平均温度	-+	2018
	秋季平均温度	-+	2016
	冬季平均温度	+-	2001、2002、2004
区域 4（大西洋海岸）	春季平均温度	-+	/
	夏季平均温度	+	2012、2014、2018
	秋季平均温度	-+-	2011、2015、2017
	冬季平均温度	-+	/
区域 5（北部）	春季平均温度	-+	2014
	夏季平均温度	-+	2012、2013、2016

	秋季平均温度	-+	2003、2006、2013、2016
	冬季平均温度	-+	2018

注：+表示升高趋势，-表示降低趋势，/表示无时间突变。

分区域看，5 个区域的整体和季节性平均温度变化趋势与加拿大大体上保持一致，虽然个别区域春秋两季均温在 2017 年后出现了降低趋势，但总体仍呈现出先降低后升高的变化态势。在突变检测中，除区域 1 和区域 4 的春冬两季未发生突变外，其余区域各季节均出现了一到四次不同程度的温度突变。

综合来看，加拿大整体和分区域的平均温度发生了由低到高的突变。

c) 周期变化分析

地区温度由于受到自然因素的影响，往往呈现出复杂的、非线性且多时间尺度的变化特征。小波分析是将温度时间序列分解到时间频率域内，得到时间序列的显著波动模式，即周期动态变化，目前小波分析方法已广泛用于气候要素周期变化研究[3]。

本文采用 Morlet 小波分析法对温度进行周期变化研究，首先对温度时间序列$y(t)$进行去趋势化和距平标准化，再进行小波变换。小波变换的步骤如下：

Morlet 小波的一般形式为：

$$\psi_0(t) = e^{ibt}e^{-t^2/2}$$

其小波变换系数为：

$$W_y(a,c) = a^{-0.5}\int_R y(t)\, e^{ib(\frac{t-c}{a})}e^{-0.5(\frac{t-c}{a})^2}dt$$

小波方差为：

$$var(a) = \int_R \left|W_y(a,c)\right|^2 dc$$

其中：$W_y(a,c)$为小波变换系数，a为伸缩尺度，c为平移参数，b为常数。小波方差图即显示了小波方差随尺度a的变化过程，它可以体现出观测数据波动的能量随尺度a的分布。

因此，小波方差图可检验出观测数据中主要时间尺度，即主周期。

首先，对加拿大整体进行小波分析，得到全年与四季的小波方差结果如下图所示，可以看出加拿大全年平均温度的小波方差在第 1、5、16、32 年处出现峰值，其中第 1 年和第 5 年通过了显著性检验，说明加拿大平均气温年际变化对应两个周期变化，第一个是 5a 尺度的主周期，第二个是 1a 尺度的次周期。其中小波方差图显示，春季平均温度具有 0-3a 和 4-7a 两种尺度的周期震荡，以及 6a 的主周期和 1a 的次周期；秋季平均温度的小波方差在

第 7、16、32 年处出现峰值，其中第 7 年通过了显著性检验，具有 7a 的周期变化；夏季与冬季周期变化情况较为相似，都具有 2-7a 尺度的周期变化,周期为 5a。

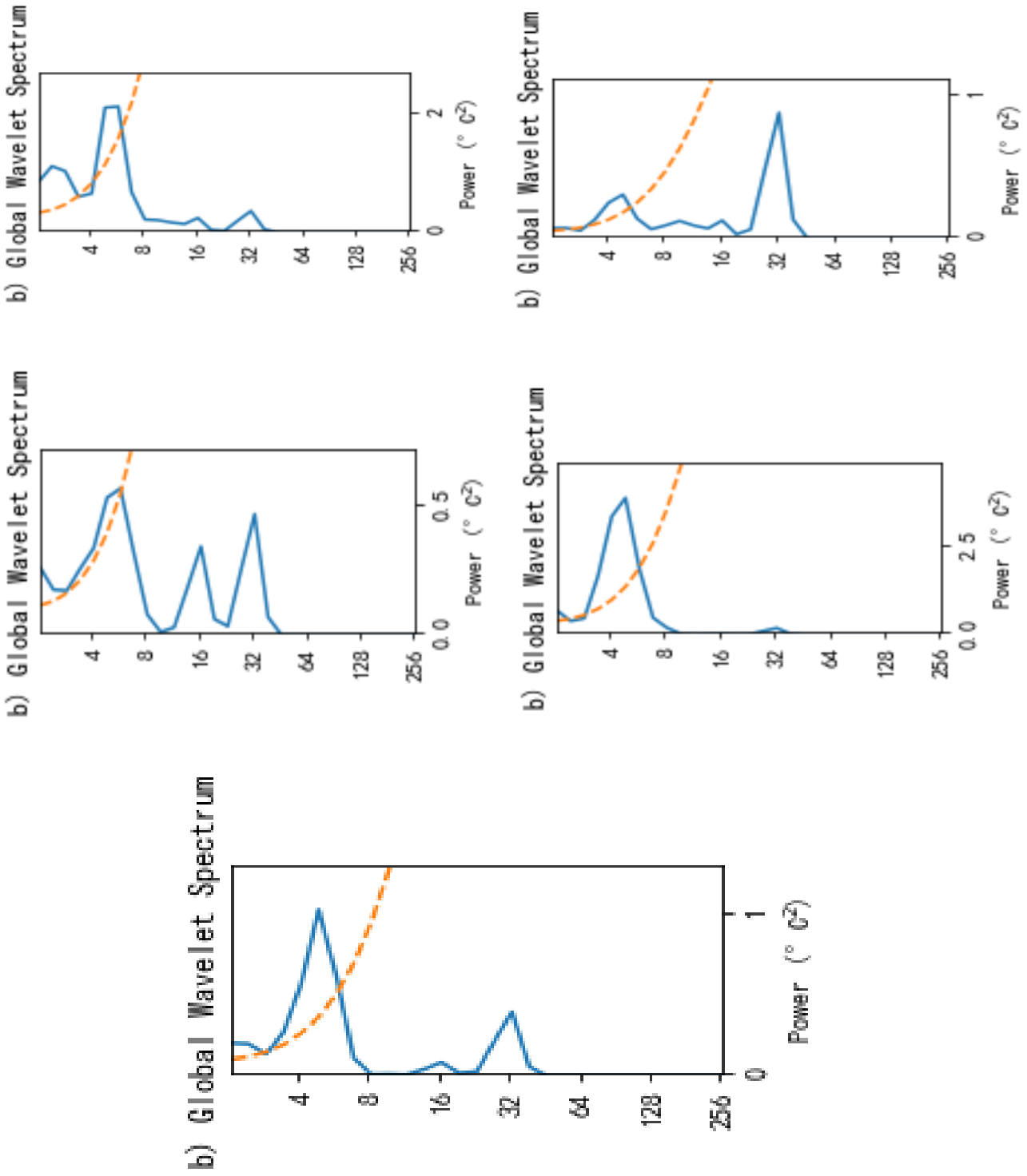


图 4-8 加拿大全年和四季平均温度小波方差图

其次对加拿大 5 大区域的周期变化进行小波分析，结果汇总如下表所示。

表 4-5 加拿大各区域小波分析结果

区域	温度要素	周期尺度	主/次周期
区域 1（南部）	春季平均温度	1-3a、5-7a	2a/6a
	夏季平均温度	2-8a	5a
	秋季平均温度	2-5a、5-10a	3a/7a
	冬季平均温度	2-7a	5a
区域 2（中部）	春季平均温度	1-4a	2a
	夏季平均温度	2-10a	6a
	秋季平均温度	3-7a	6a
	冬季平均温度	2-7a	6a
区域 3（太平洋海岸）	春季平均温度	/	/
	夏季平均温度	/	/
	秋季平均温度	/	/
	冬季平均温度	0-3a、3-6a	1a/4a
区域 4（大西洋海岸）	春季平均温度	1-3a、4-8a	2a/7a
	夏季平均温度	2-7a	5a
	秋季平均温度	2-10a	7a
	冬季平均温度	2-7a	5a
区域 5（北部）	春季平均温度	0-2a、2-7a	1a/4a
	夏季平均温度	2-7a	5a
	秋季平均温度	2-6a	3a
	冬季平均温度	2-7a	4a

3) 空间变化分析

本文采用反距离权重法（IDW）对加拿大 299 个观测站温度数据在整个地区进行空间插值，运用 ArcGIS 软件实现，分别输出了加拿大 2000 年平均温度空间分布图和 2018 年平均温度空间分布图，如下图所示。

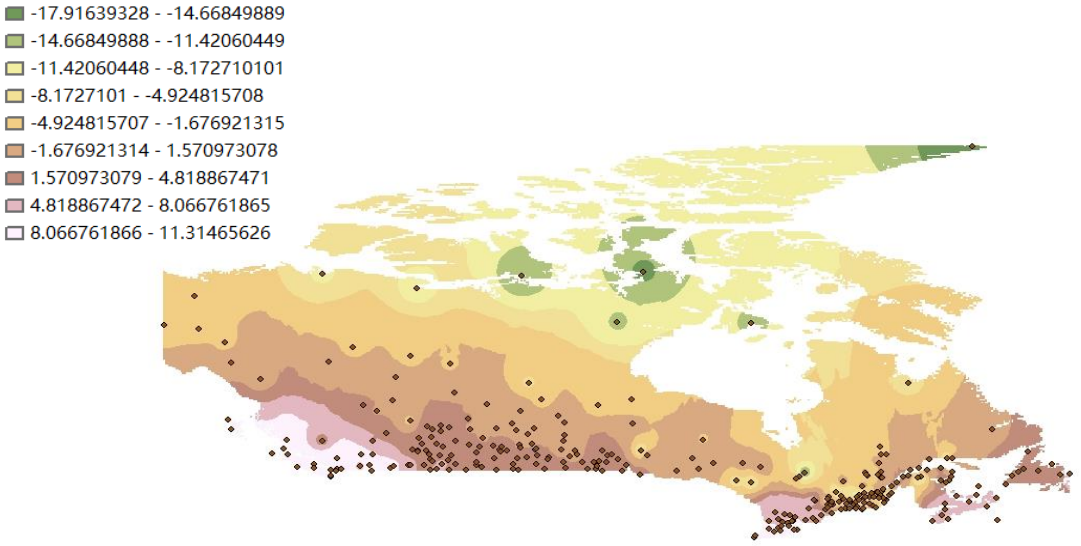


图 4-9 2000 年温度空间分布

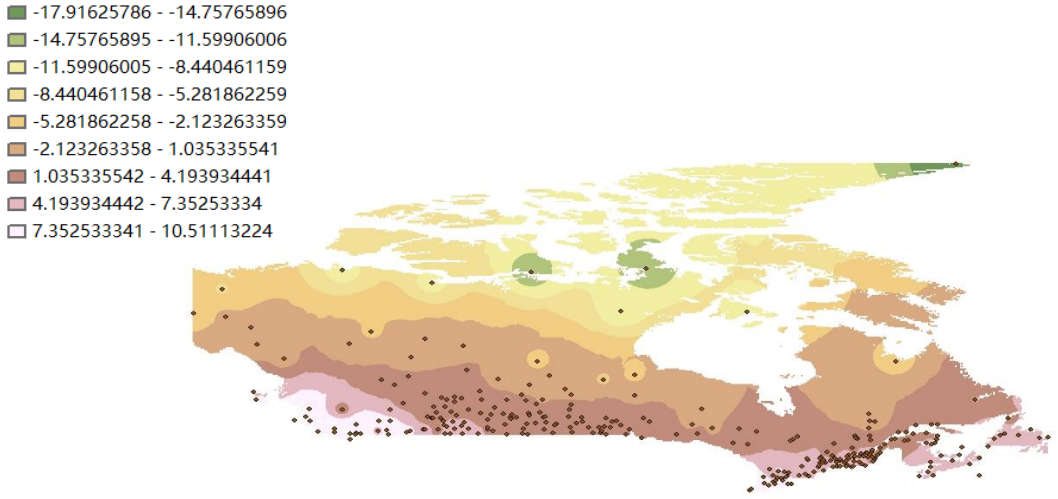


图 4-10 2018 年温度空间分布

从两张空间分布图可以看出，加拿大地区平均温度由东北向西南和东南逐渐升高，且太平洋和大西洋沿岸地区平均温度高于内陆地区，内陆地区平均温度显著高于极地地区。整个地区的年平均温度最大值出现在太平洋沿岸的维多利亚地区，年平均温度达 10.51℃，最小值出现在努勒维特省最北端的城镇 Alert，平均温度为-17.92℃，地区温度差值为 28.43℃。

在变化趋势上，可以看出近 20 年加拿大总体出现温度升高的趋势。其中增温趋势最显著的地区发生在魁北克省的西南部和安大略省东南部，升温达 6.48℃；加拿大南部地区如爱德华王子岛省、安大略省、马尼托巴省和萨斯克彻温，以及加拿大中北部地区如魁北克省中部及北部和努勒维特省中部均出现升温趋势，温度升高约 3.15℃-3.2℃；太平洋沿岸地区如不列颠哥伦比亚省的温哥华岛未出现明显温度变化趋势，近 20 年平均温度均保持在 7.35℃至 10.51℃之间。

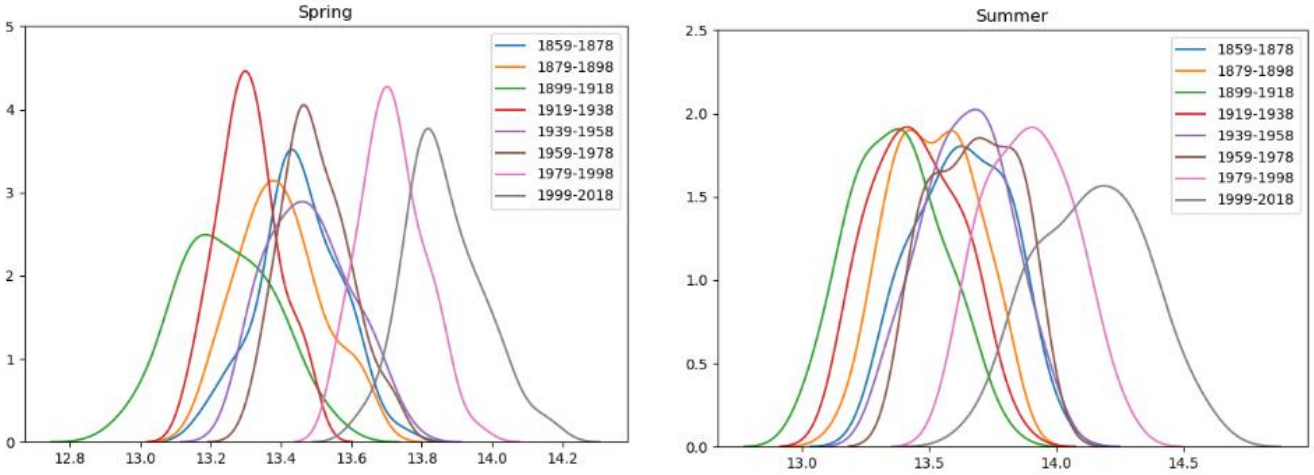
4.3 海洋表面温度规律探究

根据题目以及查阅相关文献发现，海洋表面温度的变化具有某种震荡特征，如年代际太平洋震荡，指的是北太平洋所表现出的一种年代际时间尺度上的变化现象[4]。为了进一步分析探究海洋表面温度的年代际变化特征，应用分布函数对年代际海洋表面温度进行研究，分布函数原理如下：

假设X为一组随机变量，x为任意实数，那么函数$F(x)=P\{X\le x\},-\infty<x<+\infty$称为$X$的分布函数，记为$X-F(x)$。对任意实数$x_1$、$x_2(x_1<x_2)$，$P\{x_1<X\le x_2\}=P\{X\le x_2\}-P\{X\le x_1\}=F(x_2)-F(x_1)$。

已知X的分布函数，那么就知道X落在区间$(x_1,x_2]$上的概率，因此可以说分布函数完整的展现出随机变量的规律性。如果将X看成是数轴上随机点的坐标，那么分布函数 $F(x)$在x处的函数值就表示X落在区间$-\infty,x]$上的概率。

根据题目附件所给的海洋表面温度（SST）数据，用分布函数法对其进行年代际特征分析的结果如下图所示。



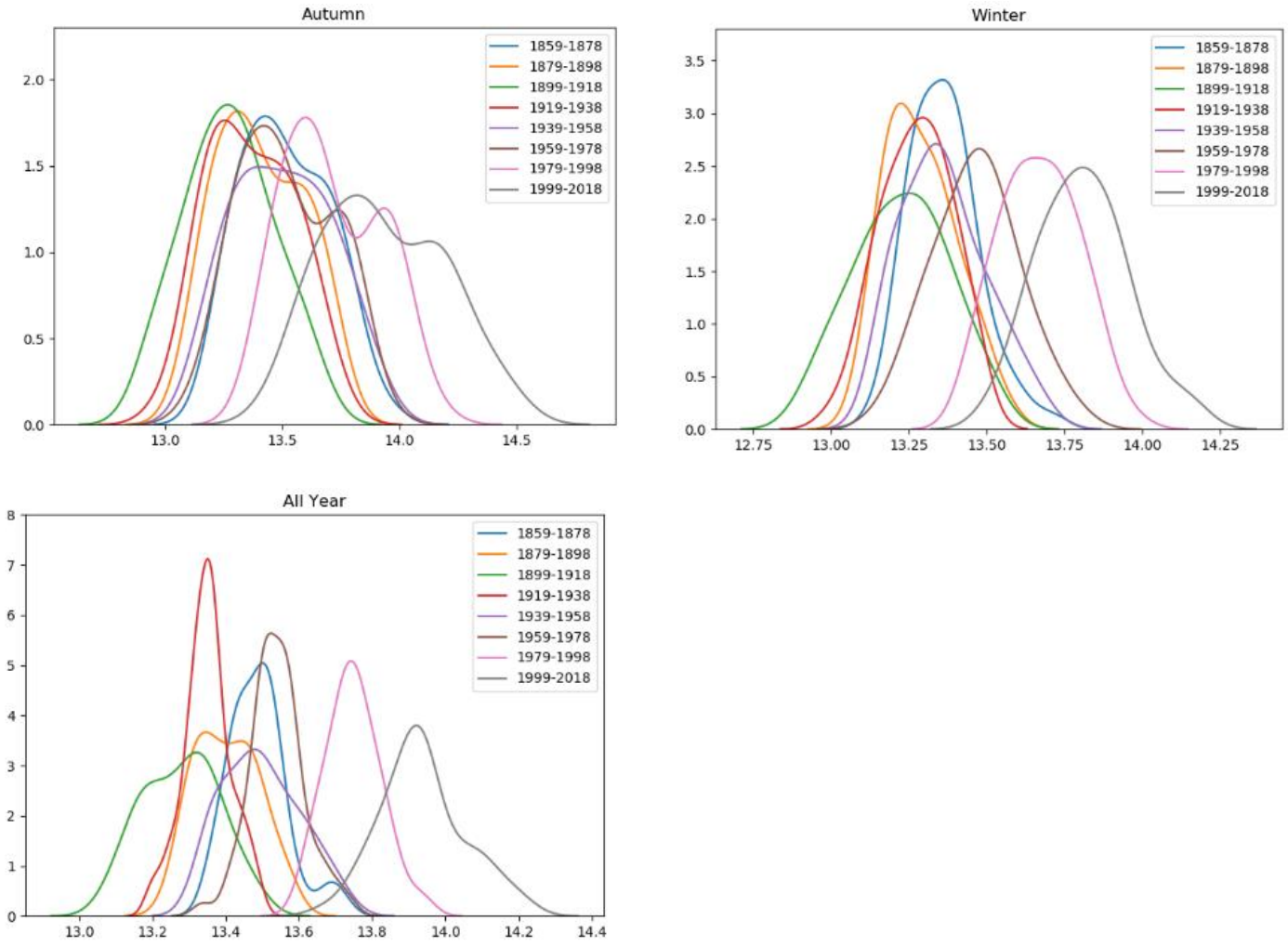


图 4-11 1859-2018 年海洋表面温度年和季节分布函数图

海洋表面温度总体表现为升高的趋势，但在 19 世纪 60 年代初至 20 世纪 10 年代末，占比最大的海洋表面温度值（以下称“海洋表面温度最大值”）向负方向移动，温度略有降低，由 13.5℃降至 13.3℃。自 20 世纪 20 年代开始，海洋表面温度最大值向正方向移动，最大值由 13.35℃增加到 13.95℃。

春季海洋表面温度变动明显，最大值变化幅度较大，随时间先减后增，总体呈现正方向移动趋势，春季海洋表面温度呈不断升高的趋势；夏季和秋季海洋表面温度变化趋势相似，从 1859-1878 年，海洋表面温度最大值像负方向移动，随后在 19 世纪 80 年代到 20 世纪 30 年代末，以及 40 年代到 70 年代末变化幅度不大，从 1979 年开始，海洋表情温度逐渐升高；冬季海洋表面温度从 19 世纪 50 年代末到 19 世纪 70 年代末，海洋表面温度最大值像负方向移动，随后在 19 世纪 80 年代到 20 世纪 30 年代末变化幅度不大，从 20 世纪 40 年代初开始升高，总体趋势为温度不断升高。

4.4 模型小结

针对加拿大温度时空变化趋势问题，本文在加拿大政府气象网站[1]中搜集了大量的温度观测数据，为保证时间序列的连续性以及观测数据的空间分布合理性，最终提取了 299 个观测站点数据，对加拿大温度进行分区域、分季节的时空趋势分析。

时间变化分别从趋势、突变和周期三个维度建模分析，最终结论是加拿大地区温度呈现上升趋势，温度上升过程中发生了由低到高的突变，且存在 5 年的主周期变化。空间变化采用反距离权重法插值分析，结果表明加拿大地区平均温度在空间分布上由东北向西南和东南逐渐升高，总体出现升温趋势，其中升温最显著的地区发生在魁北克省的西南部和安大略省东南部。

针对海洋表面温度规律探究问题，构建分布函数对海洋表面温度数据进行年代际特征分析。分析结果显示在 19 世纪 60 年代初至 20 世纪 10 年代末，海洋表面温度略有降低，自 20 世纪 20 年代开始，海洋表面温度开始升高，总体呈现为升温趋势。

五、问题二：模型的建立与求解

5.1 问题分析

问题二需要建立一个刻画气候变化的模型对未来 25 年的气候变化进行预测，题目要求该模型至少需要考虑地球的吸热、散热以及海洋的温度变化等要素。本文从 NOAA 等各气象网站[5][6]下载气象数据，抽取用于预测的要素指标及全球温度数据，建立**基于要素维度的随机森林回归预测模型**，用于对未来 25 年气候变化进行预测。同时在对未来 25 年的要素数据进行预测时，为了解决单个时间序列模型无法实现对全部要素数据预测的情况，本文建立 **ARIMA 自回归时间序列预测模型和基于 prophet 框架的时间序列预测模型**用于对未来 25 年的要素数据进行预测，为对未来 25 年的气候变化的预测提供数据支持。

5.2 数据获取及处理

5.2.1 数据获取

本文从 NOAA 等各气象网站下载气象数据，抽取了用于预测的要素指标及全球温度数据，并用 spss 简单计算该因素与全球温度的相关系数，选择相关系数较大的前几个因素用于建模，以海洋温度 spss 的相关性检验为例，结果如下表所示（海洋温度与全球温度显著线性相关）。

表 5-1 相关性检验表

因素		海洋温度	全球温度
海洋温度	Pearson 相关性	1	0.972
	显著性	-	0.000
	N	71	71
全球温度	Pearson 相关性	0.972	1
	显著性	0.000	-
	N	71	71

从而，综合相关性分析和文献资料，最终选定的因素考虑如下：

针对吸热和散热数据无法直接获取的情况，本文从净长波辐射和净短波辐射两个维度进行间接衡量。

然后，对大气环流考虑北极涛动指数和南方涛动指数。

此外，除了考虑海洋表面温度等温度数据外，本文同时考虑了二氧化碳等温室气体对气候的影响。

最后，引入了拉尼娜、厄尔尼诺及太平洋年际带震荡因素。

最终，本文确定了如下几个因素，包含温度、辐射、气体排放（人类活动）、大气环流（动力）和震荡特性 5 个层次，用于对气候变化进行建模。

表 5-2 用于气候变化建模的因素

编号	层次	因素
1	温度	海洋温度
2		陆地温度
3	辐射	净长波辐射
4		净短波辐射
5	气体排放（人类活动）	二氧化碳排放量
6	大气环流（动力）	北极涛动指数
7		南方涛动指数
8	震荡特征	太平洋年际带震荡
9		多元 MEI 指数（厄尔尼诺、拉尼娜）

5.2.2 数据预处理

对于抽取数据的缺失值和异常值均按问题一的方式处理，这里不再赘述。

对抽出的因素数据进行预处理，首先对因素进行内在相关性分析结果如下图所示：

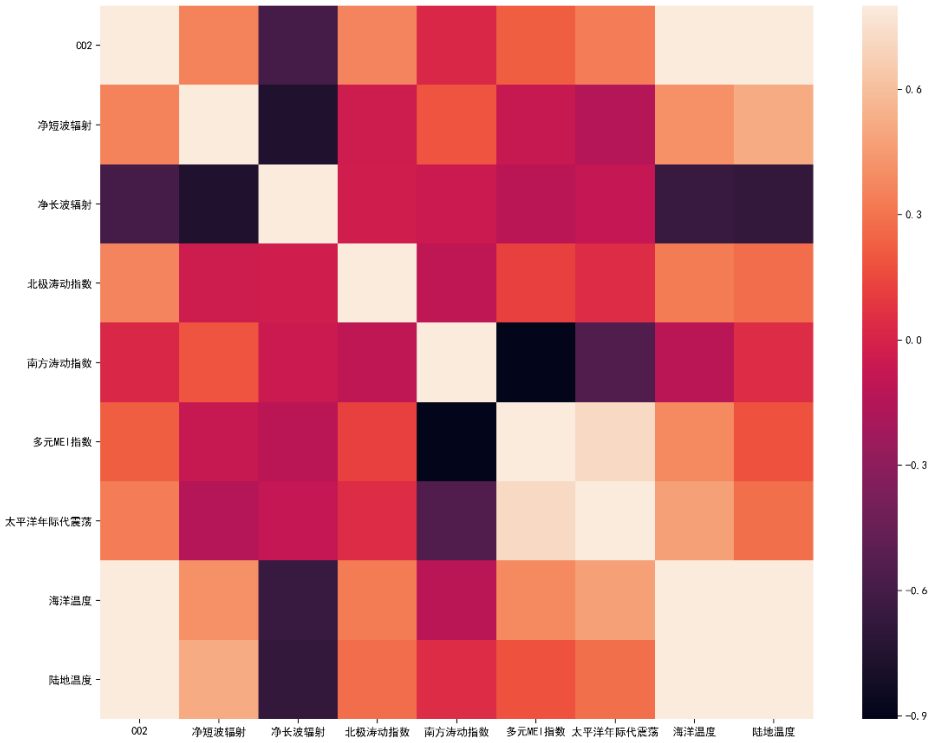


图 5-1 因素内在相关性分析

从图 5-1 右侧可以得知，相关系数越接近黑色（值越接近-1），则表示因素之间负相关；相关系数越接近白色（值越接近 1），则表示因素之间越正相关。相关性分析结果表明部分指标之间存在相关性，故利用主成分分析法（PCA）对原始的因素进行降维。结果如下图所示。

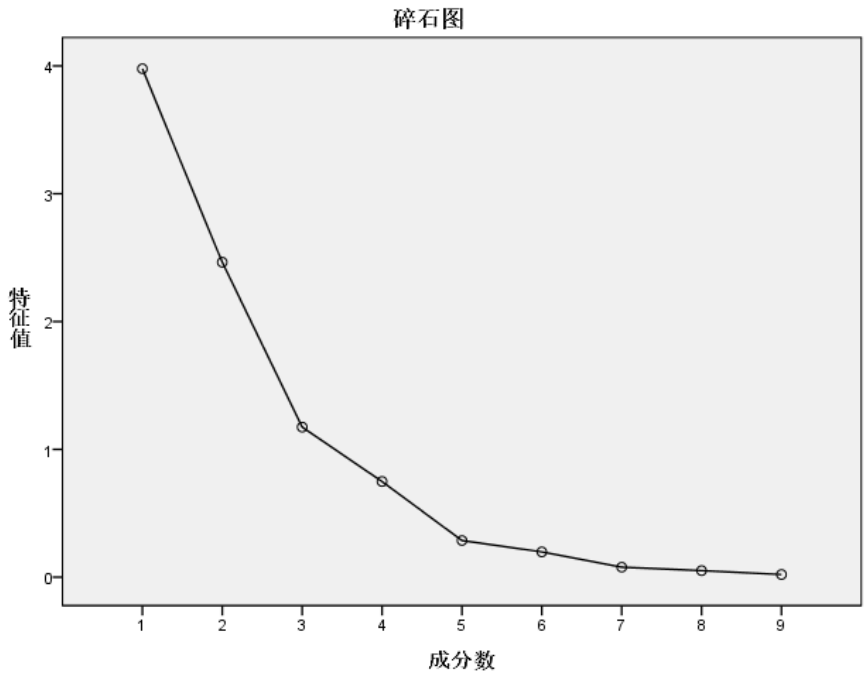


图 5-2 降维维度累计贡献率

由上图可知，最优维度数为 3-4 维，4 维的累计方差贡献率到达了 92%，能够很好的代表所有因素的信息，故最终确定将原始因素降至四维。累计方差贡献图如下图所示。

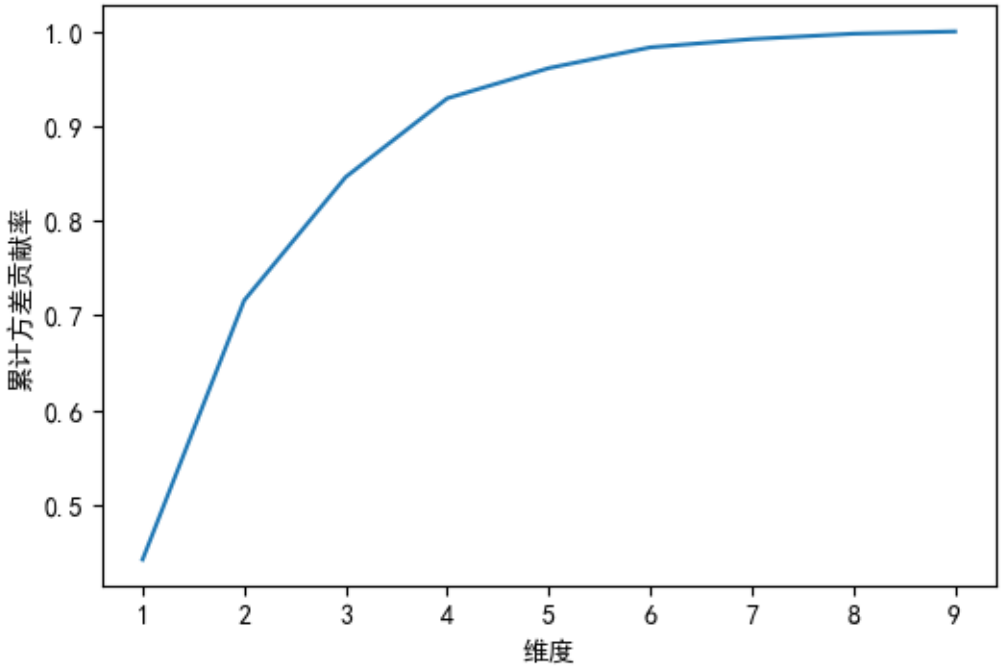


图 5-3 维度累计方差贡献率

5.3 模型建立

5.3.1 随机森林回归预测模型

1) 随机森林回归模型建立

随机森林回归是通过集成学习的思想将多棵 CART 树进行集成进行回归预测的一种模型。

本文利用 CART 回归树进行模型建立，以最小均方误差原则为原则，针对各气候影响因素，对每个因素叶子节点（划分点）s 进行划分，完成建模。具体建模过程如下：

对于任意划分因素 A，对应的任意划分点 s，父节点分裂出两个子节点（假设），两个节点的数据集为 D_1 和 D_2，求出使 D_1 和 D_2 各自集合的均方差最小，同时 D_1 和 D_2 的均方差之和最小所对应的因素和因素值作为划分点。表达式为：

$$\min_{A,s}\left[\min_{c_1}\sum_{x_i\in D_1(A,s)}\left(y_i-c_1\right)^2+\min_{c_2}\sum_{x_i\in D_2(A,s)}\left(y_i-c_2\right)^2\right]$$

式中，c_1 表示第一个节点的均值，c_2 表示第二个节点的均值。

以上式为原则进行因素划分点确定，最终建立起基于各因素的气候预测回归模型。

2) 随机森林模型参数调优

在训练集上进行参数选择，结果如下图所示。

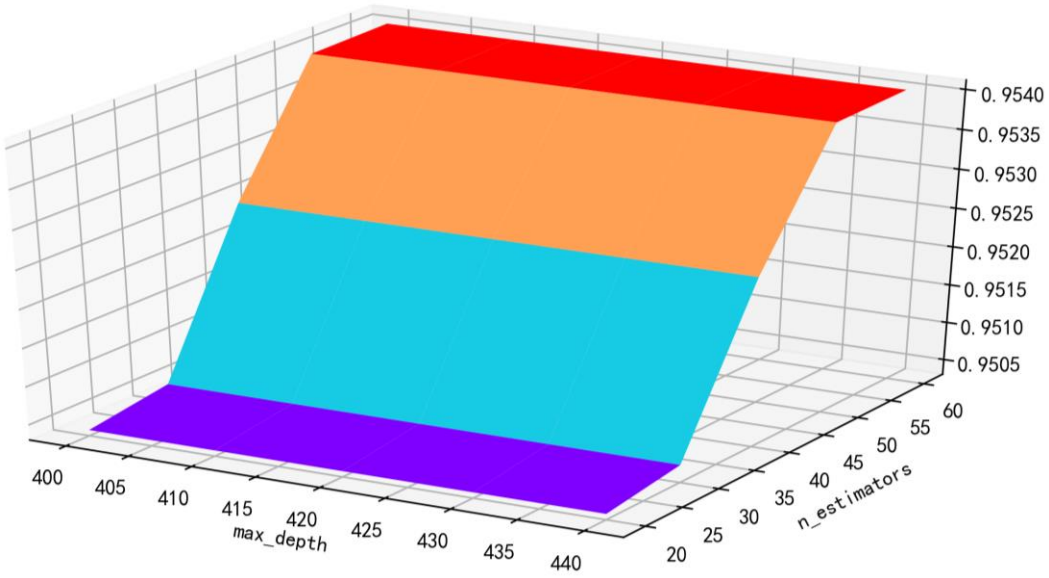


图 5-4 随机森林最佳参数选择

从上图可以看出，最佳参数选择如下表所示时，模型收敛，准确率收敛到 0.954，达到最佳建模效果。

表 5-2 模型最佳参数

参数	最优值
n_estimators	400
max_depth	60

3) 因素贡献率评估

随机森林模型同时可以输出气候影响因素的重要程度，结果如下图所示。

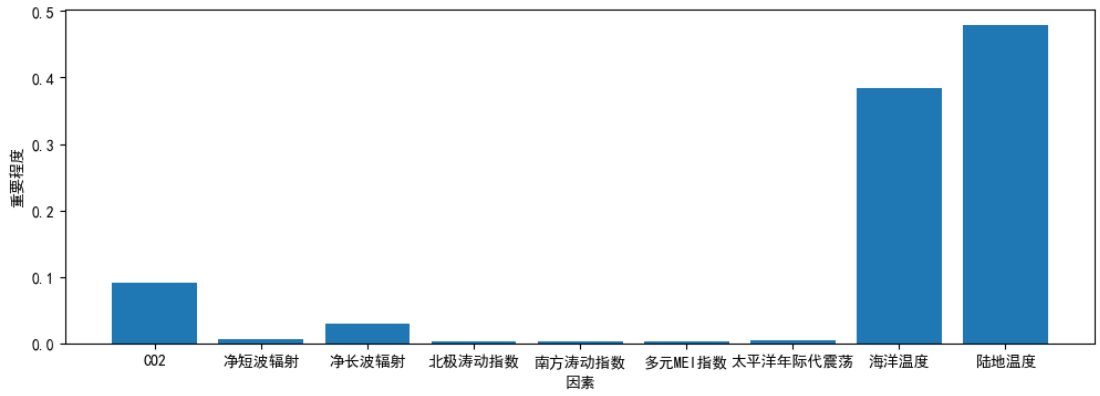


图 5-5 气候影响因素重要性排序

从上图可以看出，海洋在全球气候变化中起着重要作用。同时通过数据趋势观察可以的得出一点：当全球气温降低时，海洋温度存在一定程度的上升，这也说明近年来存在的“全球变暖中断”现象，有很大一部分原因在于海洋对热量的吸收。同时二氧化碳温室气体的排放，对全球气候的变化产生了一定影响，而二氧化碳的时间序列一直是上升趋势，这与全球温度总体升高的趋势吻合，因此减少二氧化碳等温室气体排放仍是减缓全球变暖的重要手段。此外，净长波辐射较净短波辐射更能影响全球气候。

为了更直观的体现海洋的吸热对全球变化的影响，选取“全球变暖中断”的一个区间2002 年-2005 年，绘制海洋温度和地球温度变化图。

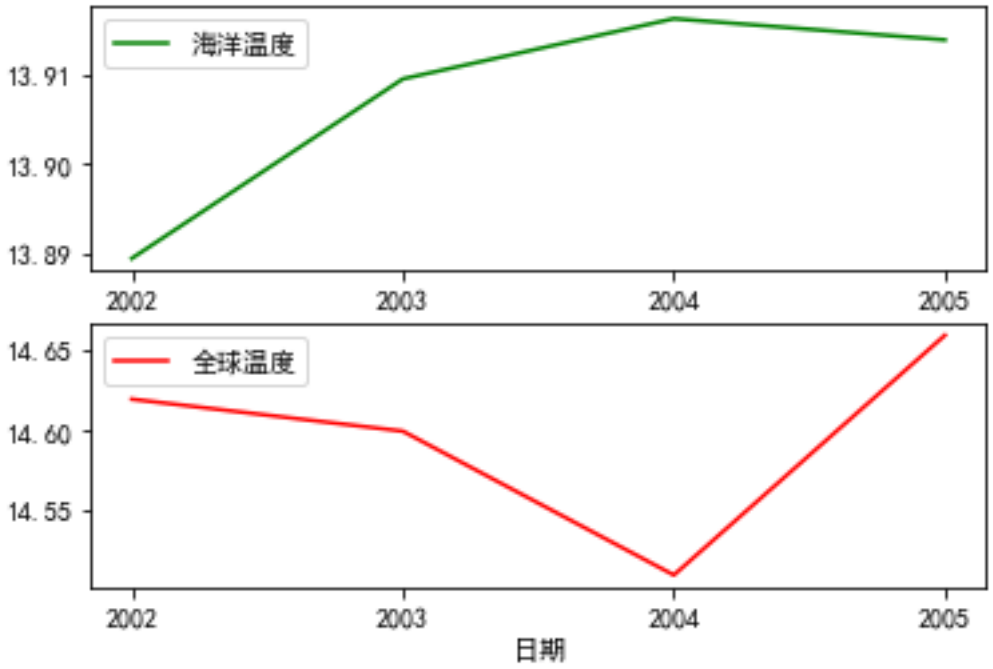


图 5-6 2002-2005 年海洋温度和地球温度

从上图可以看出，全球温度在 2002-2004 年急剧下降，并于 2004 年达到最低点，而海洋温度则在这短时间持续上升，并与 2004 年达到最高点，说明了海洋的吸热能力在一定程度上造成了“全球变暖中断”的现象。

4) 随机森林预测效果评估

在测试集上对随机森林的预测效果进行测试，结果表明，预测的调整R^2值达到了 0.966，预测效果良好，可以用于未来 25 年的气候变化预测。

5.3.2 ARIMA 自回归时间序列预测模型

ARIMA 模型是差分模型、自回归模型和移动平均模型的结合，符号为ARIMA(p, d, q)。差分模型是为了使数据更加平稳，差分阶数用d来表示。

1) 自回归模型AR

自回归模型是利用因素自身的历史数据对自身进行预测，预测只能在满足平稳性要求的情况下才能进行，平稳性检测采取 ADF 指标进行检验。

自回归模型表达式如下：

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

式中，y_t表示当前值，μ为常数，p为阶数，γ_i为自相关系数，ϵ_t为误差

2) 移动平均模型MA

移动平均模型是为了消除自回归模型的随机波动即误差ϵ_t ，其表达式如下：

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

3) 自回归移动平均模型ARMA

将自回归模型和移动平均模型结合得到自回归移动平均模型，其表达式如下：

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

5.3.3 基于 prophet 的时间序列预测模型

本文利用 FaceBook 的开源框架 prophet 建立基于时间序列趋势和周期的 prophet 可加预测模型，其表达式如下：

$$y(t) = g(t) + s(t) + \epsilon_t$$

式中，g(t)为趋势项，s(t)为周期项，ϵ_t为误差项。

1) 趋势模型g(t)

利用 prophet 的分段线性模型预测趋势变化，其表达式如下：

$$g(t) = (k + \alpha(t)^T \delta)t + (b + \alpha(t)^T \gamma)$$

式中，k表示增长率，b表示偏移量，δ表示增长率 k 发生变化的点的斜率调整值构成的向量，γ表示增长率 k 发生变化的点的偏移量调整值过程的向量，$\alpha(t)^T$表示是否进行了斜率或偏移量调整（是为 1，否则为 0）。

2) 周期模型s(t)

利用傅里叶级数近似表示周期变化，其表达式如下：

$$s(t) = \sum_{n=1}^{N}(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \sin\left(\frac{2\pi nt}{P}\right))$$

式中，P 表示某个固定的周期，2n 表示希望在模型中使用的这种周期的个数，较大的 N 值可以拟合出更复杂的周期性函数，但也伴随着过拟合问题。

5.4 模型求解

5.4.1 基于时间序列的因素预测求解

由于 ARIMA 对数据的平稳性和可预测性要求较高，根据预测可行性结果，对要素的预测采取以下策略：

用 ARIMA 模型对净长波辐射和净短波辐射进行预测；

用基于 prophet 可加预测模型对海洋温度、二氧化碳浓度等剩余因素进行预测。

下面以为净短波辐射和陆地温度为例，分别说明 ARIMA 模型和 prophet 可加预测模型的求解过程。

1) ARIMA 模型

Step1 绘制 1948 年—2018 年的时间波动趋势，并对原始数据平稳性进行评估。

首先，1948 年—2018 年净短波辐射的时间波动趋势如下图所示。

图 5-7 1948 年—2018 年的净短波辐射时间波动趋势图

对原始数据进行平稳性检验，检验结果如下。

表 5-3 原始数据 ADF 检验表

ADF 检验结果	P-value
-1.8934007651485905	0.3351719393380691

注：参考显著性水平为：'1%': -3.5289, '5%': -2.9044, '10%': -2.5897

在 ARIMA 这样的自回归模型中，模型对时间序列数据的平稳是有要求的，从原始数据结果来看，原始序列的 P 值大于 0.05，未能通过显著性检验，因此原始数据是不平稳的，需要对原始数据进行差分处理。

Step2 对原始数据进行一阶差分

对原始数据进行一阶差分，结果如下图所示。

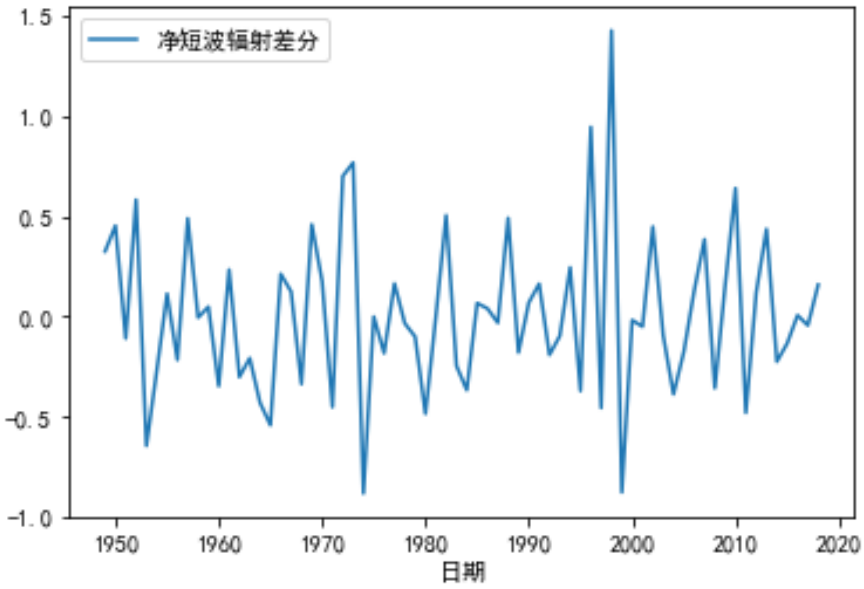


图 5-8 原始数据的一阶差分结果

Step3 对一阶差分后的结果进行平稳性检验及白噪声检验

对一阶差分后的数据进行平稳性及白噪声检验，结果如下表所示。

表 5-4 差分后的数据 ADF 检验表

ADF 检验结果	P-value
-12.044261048136999	2.6876780670431496e-22

注：参考显著性水平为：'1%': -3.5289, '5%': -2.9044, '10%': -2.5897

表 5-5 差分后的数据白噪声检验表

白噪声检验结果	P-value
9.68029189	0.00186255

从平稳性及白噪声检验结果可以看出，一阶差分后的数据通过了平稳性及白噪声检验，说明一阶差分后的数据表现平稳，且具有良好的被预测性。

Step4 利用赤池信息准则 AIC 对 ARIMA 模型进行参数选择，并利用最优参数建立模型，实现对未来 25 年的预测。预测结果如下表所示。

表 5-6 净短波辐射预测数据

年份	净短波辐射	年份	净短波辐射
2019	-131.249	2032	-131.473
2020	-131.227	2033	-131.496
2021	-131.108	2034	-131.452
2022	-131.153	2035	-131.421
2023	-131.334	2036	-131.457
2024	-131.423	2037	-131.42
2025	-131.537	2038	-131.441
2026	-131.513	2039	-131.44
2027	-131.431	2040	-131.393
2028	-131.449	2041	-131.417
2029	-131.414	2042	-131.382
2030	-131.463	2043	-131.372
2031	-131.514		

可视化结果如下图所示。

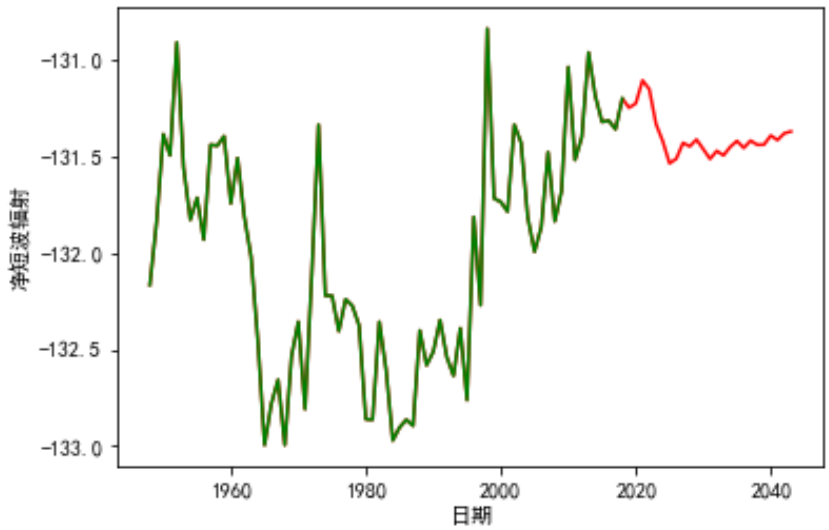


图 5-9　25 年间的净短波辐射预测值可视化

2)　基于 prophet 的时间序列预测模型

预测结果如下表所示。

表 5-7 陆地温度预测数据

年份	温度	年份	温度
2019	9.405119	2032	9.834414
2020	9.357852	2033	9.980724
2021	9.504162	2034	10.03631
2022	9.559744	2035	10.04053
2023	9.563973	2036	9.993268
2024	9.516706	2037	10.13958
2025	9.663016	2038	10.19516
2026	9.718598	2039	10.19939
2027	9.722827	2040	10.15212
2028	9.67556	2041	10.29843
2029	9.82187	2042	10.35401
2030	9.877452	2043	10.35824
2031	9.881681		

陆地温度的预测情况如下图所示。

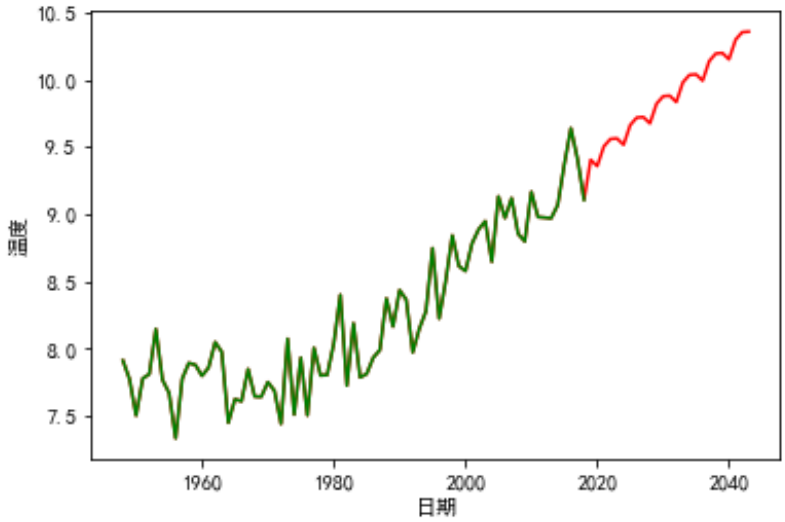


图 5-10 陆地温度的预测情况

剩余气候影响因素预测完成之后，将用于未来 25 年气候变化的预测。

5.4.2 基于因素的随机森林气候预测模型求解

基于时间序列预测的未来 25 年因素数据，利用建立的随机森林模型，实现对未来 25 年的预测。预测结果如下表所示。

表 5-8 未来 25 年温度预测结果

年份	温度	年份	温度
2019	14.84967	2032	14.96013
2020	14.81917	2033	15.01711
2021	14.77243	2034	15.02824
2022	14.81041	2035	14.99462
2023	14.8577	2036	15.02426
2024	14.83554	2037	15.06949
2025	14.84122	2038	15.04539
2026	14.90898	2039	15.03273
2027	14.92288	2040	15.08419
2028	14.89722	2041	15.10607
2029	14.9378	2042	15.0714

续表 5-8

2030	14.988	2043	15.09391
2031	14.9645		

可视化结果如下图所示。

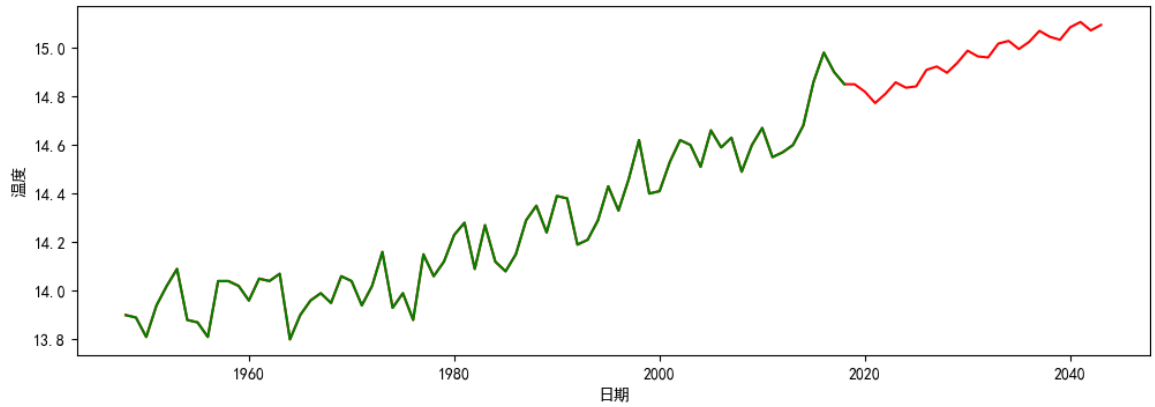


图 5-11　未来 25 年的预测值可视化

从预测结果可以看出，2019-2021 年维持了前几年的下降趋势，2021 年之后温度开始升高，总体保持“变暖”的趋势，并于 2041 年达到该预测区间的最大值 15.11 摄氏度。

5.5 模型小结

针对问题二，首先确定了因素选取的温度、辐射、气体排放（人类活动）、大气环流（动力）和震荡特性 5 个层次，再根据统计学上的相关性进行一定的筛选，确定最终用于预测气候的因素；然后建立基于因素的随机森林回归预测模型，通过模型参数的选优保证了模型拥有 96%左右的准确性；为了对未来 25 年的气候变化进行预测，建立了 ARIMA 自回归预测模型和基于 prophet 的时间序列预测模型共同完成对未来 25 年气候影响因素的预测，为未来 25 年气候的预测提供支持；最后利用随机森林回归模型实现了对未来 25 年气候的预测。预测结果表明，2019-2021 年维持了前几年的下降趋势，2021 年之后温度开始升高，总体保持“变暖”的趋势，并于 2041 年达到该预测区间的最大值 15.11 摄氏度。

此外，通过影响因素对模型预测的重要性结果可知：①海洋在全球气候变化中起着重要作用。同时通过数据趋势观察可以的得出一点：当全球气温降低时，海洋温度存在一定程度的上升，这也说明近年来存在的“全球变暖中断”现象，有很大一部分原因在于海洋对热量的吸收。②二氧化碳温室气体的排放，对全球气候的变化产生了一定影响，而二氧化碳的时间序列一直是上升趋势，这与全球温度总体升高的趋势吻合，因此减少二氧化碳等温室气体排放仍是减缓全球变暖的重要手段。③对于辐射层面，净长波辐射较净短波辐射更能影响全球气候。

六、问题三：模型的建立与求解

6.1 问题分析

针对问题三，要分析极寒天气和气候变化的内在关系，并判断全球变暖和极寒天气是否矛盾。为解决这个问题，本文认为首先应该建立起极寒天气和气候因素之间的联系，分析极寒天气和气候因素之间的关联，确定极寒天气和气候变化存在的内在相关性。其次，针对影响极寒天气的气候因素，引入全球温度数据表征全球变暖趋势，建立他们之间的相关分析，从而确定全球变暖与极寒天气之间的关系。

6.2 数据预处理

6.2.1 数据选取

根据气象专业制定的寒冷程度等级表，气温从 9.9℃到零下 40℃以下。极寒程度一共分为八级，其中极寒天气是指零下 40℃以下的天气。近年来，在欧洲与亚洲的许多地区，暴雪的袭击使许多地区出现了百年来的最低气温，给当地居民的生活造成了严重不便。在中国的某些地区，比如内蒙古和黑龙江，最低气温也跌破了零下 40℃。

对于极寒天气产生的原因，有专家认为是因为拉尼娜事件的持续发展，导致大部分冬季气温偏低。同时西伯利亚高压持续增强，导致东亚冬季风偏强，带来冷空气。此外，北极涛动也会带来降温和寒潮天气。

通过文献和相关分析可知，目前对极寒天气的讨论主要集中在温带和亚热带，即为中低纬度地区。对于高纬度地区，出现极寒天气是大概率事件，因此分析高纬度地区是没有意义的。所以本文选取纬度在 40N～60N 的地区。首先，本文统计了从 1948 年到 2018 年每个月同一纬度下出现的最低气温，数据来源为 NOAA 官网[5]。如果该月的最低气温值小于或等于零下 40℃，即出现了极寒天气，则将此月的数值标记为 1；如果该月的最低气温值大于零下 40℃，即没有出现极寒天气，则将此月的数值标记为 0。最后，根据此数值统计全年极寒天气出现频次，并将频次数据作为分类标签。

6.2.2 数据生成

针对极寒天气数据量较正常天气少的情况，本文在模型训练时，采用 SMOTE 算法，对极端数据进行上采样来解决数据不平衡的问题。

6.2.3 因素降维

针对本问题降维方式与问题二一致，这里不再赘述。

6.3 模型建立

6.3.1 随机森林分类模型

1) 随机森林分类模型建立

与问题二中的随机森林回归模型一样，随机森林分类也是通过集成学习的思想将多棵 CART 树进行集成进行分类的一种模型。不过，这里的 CART 树为分类树，采用的评估标准也和回归不同，采用的是基尼系数，其表达式如下：

$$\text{gini(p)} = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

式中，p_k表示选中的样本属于 k 类别的概率（$1-p_k$表示被分错的概率）。

这样，基尼系数越小，表明样本被分错的概率越小,表明划分越合理。

2) 随机森林分类模型选参

将问题二的因素数据作为随机森林分类的特征数据，同时本题的出发点重点关注随机森林分类模型对极寒天气的识别情况，因此以对极寒天气的召回率（recall 值）为依据，进行参数选择，结果如下图所示。

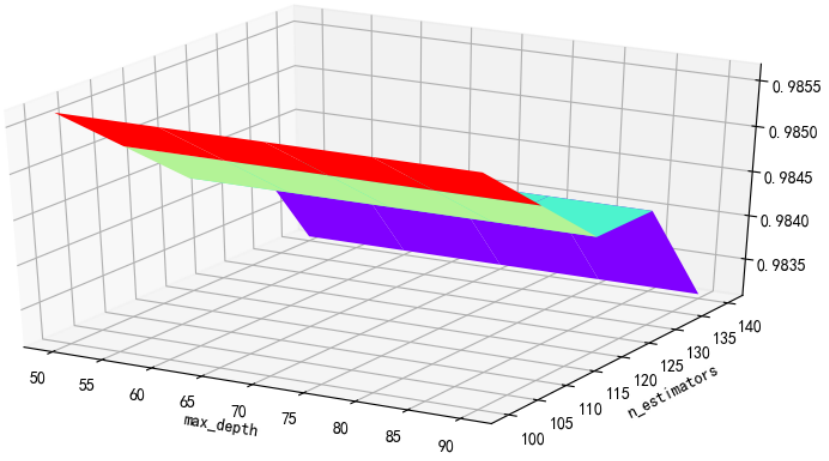


图 6-1 随机森林参数选优过程

6.3.2 Pearson 相关模型

Pearson 相关模型是基于正态连续性分布数据进行相关性分析的模型，其模型表达式如下式所示。

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

式中，X,Y 为两个待分析因素，E表示的是数学期望。

6.3.3 Spearman 相关模型

Spearman 相关又称等级变量之间的 Pearson 相关，可以处理满足任何分布的任何形式的变量，其“表达式”如下：

$$\rho_{x,y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

式中，x,y 为两个待分析因素，$\bar{x}$，$\bar{y}$表示的是两个因素的均值。

6.4 模型求解

6.4.1 随机森林分类模型结果

1) 极寒天气识别结果

极寒天气的识别 recall 值在不同训练集上能达到 90%左右，说明随机森林分类模型能够很好的识别极寒天气。

2) 因素重要程度结果

通过对重要因素分析可以洞悉极寒天气的出现和什么因素密切相关。随机森林回归模型输出因素重要程度结果如下图所示。

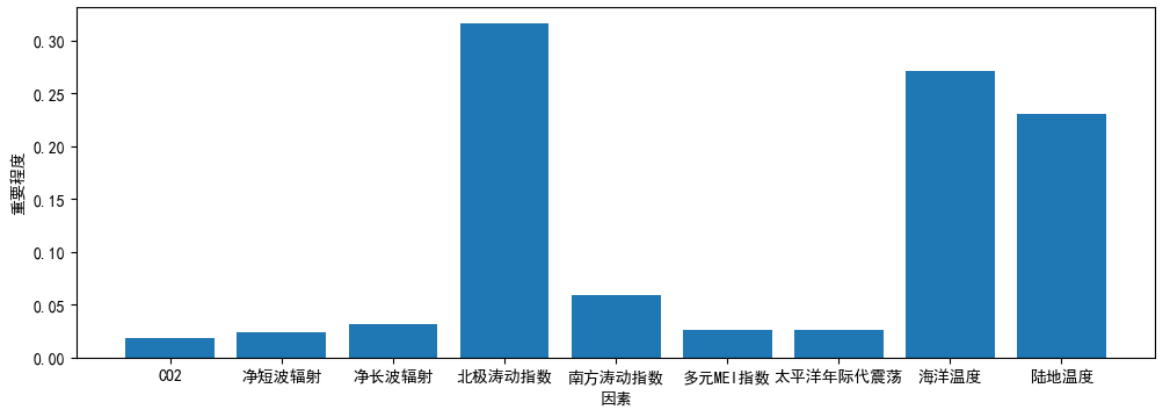


图 6-2 因素重要程度图

由上图可知，极端天气的出现与北极涛动之术和海洋温度变化密切相关。

6.4.2 相关性模型结果

首先，利用 spss 对海洋温度数据、北极涛动指数数据和全球温度数据进行正态性检验，结果表明海洋温度数据和全球温度数据均服从正态分布，而北极涛动指数数据不服从正态分布，以全球温度数据为例，其正态性检验如下表所示。

表 6-1 全球温度数据的正态分布检验结果

	Kolmogorov-Smirnova			Shapiro-Wilk		
全球温度	统计量	df	Sig.	统计量	df	Sig.
	0.134	71	0.003	0.943	71	0.003

故对海洋温度和全球温度的相关性分析利用 Pearson 相关模型进行建模，对北极涛动和海洋温度、北极涛动指数和极寒天气的相关分析利用 Spearman 相关模型进行建模。

1) 海洋温度和全球温度的 Pearson 相关性结果

根据海洋温度和全球温度，利用 spss 对海洋温度与全球温度的关系进行检验，检验结果如下：

表 6-2 海洋温度和全球温度的 Pearson 相关性结果

因素		海洋温度	全球温度
海洋温度	Pearson 相关性	1	0.972
	显著性（双侧）	-	0.000
	N	71	71
全球温度	Pearson 相关性	0.972	1
	显著性（双侧）	0.000	-
	N	71	71

检验结果表明，全球变暖和海洋温度增加整体上为显著的线性正相关。全球变暖的时候，往往海洋吸收的热量也多，导致海洋表面温度升高。

2) 海洋温度和北极涛动指数 Spearman 相关性结果

根据海洋温度和北极涛动指数数据，利用spss对北极涛动指数与海洋温度的关系进行spearman检验，检验结果如下：

表6-3 海洋温度和北极涛动指数Spearman相关性结果

Spearman的rho		海洋温度	北极涛动指数
海洋温度	相关系数	1.000	-0.277
	Sig.（双侧）	-	0.019
	N	71	71
北极涛动指数	相关系数	-0.277	1.000
	Sig.（双侧）	0.019	-
	N	71	71

检验结果表明，北极涛动指数和海洋温度增加整体上为负相关。海洋温度的升高会导致北极涛动向负相位变动。

3) 北极涛动指数和极寒天气的 Spearman 相关性结果

根据极寒天气和北极涛动指数数据，利用spss对北极涛动指数与极寒天气的关系进行spearman检验，检验结果如下：

表6-4 北极涛动指数和极寒天气的Spearman相关性结果

Spearman的rho		极寒天气	北极涛动指数
极寒天气	相关系数	1.000	-0.241
	Sig.（双侧）	-	0.043
	N	71	71
北极涛动指数	相关系数	-0.241	1.000
	Sig.（双侧）	0.043	-
	N	71	71

检验结果表明，北极涛动指数和极寒天气出现次数整体上为负相关，北极涛动处于负位相会导致极寒天气出现的概率增大。

6.4.3 全球变暖与局地极寒分析

从随机森林分类模型可知，极寒天气和海洋温度变化及北极涛动指数相关，尤其与北极涛动联系密切。说明极寒天气和气候变化存在内在的关联性。

通过海洋温度和全球温度相关性分析，可知，全球变暖与海洋温度升高呈明显正相关趋势，也就是说全球变暖的同时，海洋也在吸收大量热量，温度也在上升。

通过海洋温度和北极涛动指数关性分析，可知，北极涛动与海洋温度呈一定负相关趋势，也就是说海洋也温度在上升时，北极涛动向负相位变化。而北极涛动和极寒天气联系最为密切，相关性检验结果也表明，北极涛动向负相位变动会导致极寒天气出现的概率增大。

因此，不难得出以下结论：

全球变暖的同时，海洋由于吸收大量热量温度也在上升，这就造成了北极涛动指数向负相位移动，进而导致极寒天气出现的概率增大，因此，全球变暖与局部极寒天气并不矛盾。

6.5 模型小结

针对本问题，首先通过建立随机森林分类模型，确定了和极寒天气相关的气候因素，得到了极寒天气实际上和气候变化存在内在关联性的结论；然后建立起海洋温度和全球温度的Pearson相关性模型、海洋温度和北极涛动指数Spearman相关性模型，并结合极寒天气和北极涛动的Spearman相关性规律，得到最终结论：全球变暖的同时，海洋由于吸收大量热量温度也在上升，这就造成了北极涛动指数向负相位移动，进而导致极寒天气出现的概率增大，因此，全球变暖与局部极寒天气并不矛盾。

七、问题四：解释分析

本文问题三分析了全球变暖与局地极寒现象的关系，根据问题三分析得到的结论：全球变暖时，海洋吸收的热量增加，导致海洋温度的升高，使得海水在冬季时很难结冰，继而影响了北极地区的气压和环流变化，增加了冷空气南袭的可能性，因此出现了“某地今年冬天特别冷”的现象。

代替“全球变暖”的新概念：**多维度全球“变暖”**，解释如下：

①趋势性。加引号表明“变暖”仍是全球气候变化的大趋势，虽然某些年份某些地区仍会出现气温下降的现象，但这种局地气温下降归根结底仍是由全球变暖造成的。

②复杂性。复杂性主要表现在全球变暖会出现所谓的“中断现象”，使得全球是否变暖这个问题显得扑朔迷离。根据文章研究结论，这种“中断”现象很大程度上与海洋的吸热有关，同时根据问题三的结论，海洋吸热的效力极有可能通过地球的动力系统进行传播，诱发全球各种极端天气。从该层面来说全球变暖所谓的“中断”现象其实是转移到其他层面上，如造成局地极寒现象，产生全球变暖“中断”假象，实则继续对整个地球造成影响。因此，复杂性可以用多维度的影响来概括。

八、模型评价与推广

8.1 模型的优点

（1）本文在考虑问题时使用了多个模型。比如针对问题二建立气候变化模型对未来 25 年的气候变化进行预测时，本文建立了随机森林回归预测模型、ARIMA 自回归时间序列预测模型，以及基于 Prophet 框架的可加时间序列预测模型等三种不同的预测模型，并且用不同的模型进行交叉验证，极大地保证了预测结果的合理性和完善性。

（2）本文在研究问题时考虑的数据广度较高。比如问题一研究加拿大地区温度变化时，从全国 3000 多个观测站进行筛选，考虑时间序列连续性最终选取了 299 个观测站点，而不是笼统地选取站点对加拿大整体进行研究。文章针对问题二和问题三在研究气候变化模型时，各个气候要素的数据选取均超过 70 年。数据的较高广度保证了文章的严谨，消除了因时间区间较短可能造成预测或分类不准的可能。

8.2 模型的缺点

（1）由于计算机算力的限制，本文在考虑数据广度的同时无法兼具数据的深度，因此文章研究所选取的气候要素等数据都是以月或年为最小单位。本文中运用的时间序列预测模型与机器学习相关，数据量深度越大，预测的结果会更加准确。

（2）由于时间的限制，本文在考虑极寒天气时，只是对每个月是否出现极寒天气做出了判断，采取的是二元变量，对具体的极寒天气相关数值以及数值变化趋势并未进行探讨。

8.3 模型改进与推广

在后续研究中，为保证同时兼顾数据的深度与广度，首先需要提升硬件性能，对数据集进行多次训练，不断优化模型，提升模型预测精度。其次改进之处在于，对极寒天气的数值以及变化趋势进行分地区研究，提高判断某一地区出现极寒天气的概率。

本文在研究全球气候变化建模过程中，并未使用传统的气象学模型，但是模型的准确率较高，因此在气象学领域的从业人员可以考虑使用本文建立的模型与传统气象模型相互验证。

本文主要是研究气温在一定影响因素下的变化，但是建立的模型也可以研究其他气候要素的变化，比如降水量等，模型具有较高的适用性。

九、参考文献

[1] 加拿大政府气象网，加拿大温度数据查询，http://climate.weather.gc.ca/historical_data/search_historic_data_e.html，2019.9.20

[2] 刘雅各,袁凤辉,王安志,吴家兵,郑兴波,尹航,关德新.长白山生态功能区气候变化特征[J].应用生态学报,2019,30(05):1503-1512.

[3] Torrence C，Compo GP. A practical guide to wavelet analysis. Bulletin of the American Meteorological Society，1998,79: 61-78

[4] G. A. Meehl, A. Hu, B. D. Santer & S. Xie, Contribution of the interdecadal Pacific Oscillation to twentieth-century global surface temperature tends, Nature Climate Change, 6, 1005-1008 (2016)

[5] 美国国家海洋和大气管理局，气象数据查询，http://www.noaa.gov/web.html，2019.9.22

[6] 地球政策研究所，二氧化碳浓度数据查询，http://www.earthpolicy.org/press_room/C68/au_revoir_thank_you，2019.9.22

附 录：Python 代码

程序 1	加拿大观测站的交集和并集

```
import pandas as pd
import os

Folder_Path = r'C:\\Anaconda\\Project\\stat'
file_name = os.listdir(Folder_Path)
del(file_name[-1])
print(file_name)
print(len(file_name))

# 切换到当前路径
os.chdir(Folder_Path)
print(os.getcwd())

df = pd.read_csv("C:\\Anaconda\\Project\\stat\\" + file_name[0],error_bad_lines=False)
same_stn = df['Stn_Name']
print(len(same_stn))

# 提取每个月均出现的观测站
for i in range(1,228):
    df = pd.read_csv("C:\\Anaconda\\Project\\stat\\" + file_name[i],error_bad_lines=False)
    # print(df.head())
    list1 = df['Stn_Name']
    # same_stn = list(set(list1).union(set(same_stn))) 取并集
    same_stn = list(set(list1) & set(same_stn)) #取交集
    # print(i)
```

```
        # print(file_name[i])
        print(len(same_stn))
        # print(same_stn)


# 将数据抽取出来
for i in range(0,228):
        df= pd.read_csv("C:\\Anaconda\\Project\\stat\\" + file_name[i], error_bad_lines=False)
        # 根据索引取得值
        index_values = []
        for stn in same_stn:
                a = df[(df['Stn_Name']==stn)].index.tolist()
        #       for i in a:
        #               if not i in index_values:
                index_values.append(a[0])
        print(len(set(index_values)))
        df2 = pd.DataFrame(df,index=index_values)
        df2.sort_values(by='Stn_Name')
        df2 = df2[['Stn_Name','Lat','Long','Prov','Tm']]
        # print(df2)
        df2.to_csv("C:\\Anaconda\\Project\\stat2\\" + file_name[i].split('-')[-1],index=False)


# 数据合并
for i in range(0,228):
    df= pd.read_csv("C:\\Anaconda\\Project\\stat2\\" + file_name[i], error_bad_lines=False)
```

程序 2	缺失值处理
import pandas as pd	

```
for i in range(1,13):
    df = pd.read_csv("C:\\Anaconda\\Project\\stat4\\" + str(i) +".csv")
    # 检验缺失值的个数
    # print(df[df.isnull().values==True])
    # 删除空行
    a = df[(df['Stn_Name']== 'RIVERS PETTAPIECE')].index.tolist()
    b = df[(df['Stn_Name']== 'UNCAS')].index.tolist()
    df = df.drop(index=a[0])
    df = df.drop(index=b[0])
    #转置
    df = df.T
    # 删除表头
    df = df.drop(['Stn_Name'])
    for column in list(df.columns[df.isnull().sum() > 0]):
        mean_val = df[column].mean()
        print(mean_val)
        df[column].fillna(mean_val,inplace = True)
    filled_df = df.T
    filled_df.to_csv("C:\\Anaconda\\Project\\stat5\\filled" + str(i) + ".csv" ,index=False)
```

程序 3	长短波辐射数据下载

```
import urllib
import time


start_url='ftp://ftp.cdc.noaa.gov/Datasets/ncep.reanalysis.dailyavgs/surface_gauss/nswrs.sfc.gauss.'
```

```
for i in range(1948,2020):
    url = start_url + str(i) + '.nc'
    file_name = "C:\\Anaconda\\Project\\stat7\\nswrs.sfc.gauss\\" + url.split('/')[-1]  # 文件保存位置+文件名
    urllib.request.urlretrieve(url,file_name)
    time.sleep(0.5)
    print('下载完成的年份是%s'%i)
```

程序 4	陆地温度处理与绘图

```
from netCDF4 import Dataset
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.basemap import Basemap


nc_obj = Dataset('C:\\Anaconda\\Project\\stat7\\air.mon.mean.nc')


#查看 nc 文件
print(nc_obj)
print('-------------------------------------')


# 查看具体的信息
for var in nc_obj.variables:
    print(var,end=':\n')
    for attr in nc_obj[var].ncattrs():
        print('%s: %s' % (attr,nc_obj[var].getncattr(attr)))
    print()
```

```
lons = nc_obj.variables['lon'][:]
lats = nc_obj.variables['lat'][:]
time = nc_obj.variables['time'][:]
air = nc_obj.variables['air'][:]
air1 = nc_obj.variables['air'][:][0]

lt = []
df = pd.DataFrame()
for i in range(0,860):
    air2 = nc_obj.variables['air'][i]
    air_0 = air2.min()-273.5
    lt.append(air_0)

# lt2 = range(0,1985)
lt2 = pd.date_range('19480101',periods=860,freq='1M')
print('---------------------------------------')
print(lt)
df['mon'] = lt2
df['mair'] = lt
df.to_csv("C:\\Anaconda\\Project\\stat7\\air.mon.max.csv",index=False)
plt.scatter(lt2,lt)
plt.title('air.mon.mean')
plt.show()
print('---------------------------------------')

# 绘图
lon_0 = lons.mean()
```

```
lat_0 = lats.mean()


m = Basemap(lat_0=lat_0, lon_0=lon_0)
lon, lat = np.meshgrid(lons, lats)
xi, yi = m(lon, lat)


cs = m.pcolor(xi,yi,air1-273.5,cmap='PuBu_r')


m.drawparallels(np.arange(89.5,-89.5,30), labels=[1,0,0,0], fontsize=10)
m.drawmeridians(np.arange(0.5,359.5,30), labels=[0,0,0,0], fontsize=10)


# Add Coastlines, States, and Country Boundaries
m.drawcoastlines()
m.drawstates()
m.drawcountries()


# Add Colorbar
cbar = m.colorbar(cs, location='bottom', pad="10%")
# cbar.set_label(sst2_units)


# Add Title
plt.title('land Surface Temperature 1948.1')
plt.show()
```

程序 5	线性回归
import matplotlib.pyplot as plt	

```
import pandas as pd
from sklearn import linear_model

df = pd.read_excel("C:\\Anaconda\\Project\\mcm2019\\fenbu\\canada.xlsx")
X =df[['year']]
y = df['winter'].values

model = linear_model.LinearRegression()
model.fit(X, y)
print(model.intercept_)
print(model.coef_)
print(model.score(X,y))

plt.scatter(X,y,color='blue')
plt.plot(X,model.predict(X),color = 'red')
plt.title('All Year')
plt.xlabel('year')
plt.ylabel('temperature')
plt.show()
```

程序 6	极寒地区判定与数据获取

```
from netCDF4 import Dataset
import pandas as pd

nc_obj = Dataset('C:\\Anaconda\\Project\\stat7\\air.mon.mean.nc')

#查看 nc 文件
```

```
print(nc_obj)
print('---------------------------------------')

# 查看具体的信息
for var in nc_obj.variables:
    print(var,end=':\n')
    for attr in nc_obj[var].ncattrs():
        print('%s: %s' % (attr,nc_obj[var].getncattr(attr)))
    print()

lons = nc_obj.variables['lon'][:]
lats = nc_obj.variables['lat'][:]
time = nc_obj.variables['time'][:]
air = nc_obj.variables['air'][:]
lt = []
df = pd.DataFrame()
for i in range(0,860):
    air2 = nc_obj.variables['air'][:][i][60:80].min()-273.5
    lt.append(air2)
    print('当前是%s 月'%i)
# print(lt)
df['min_tem'] = lt
df.to_csv("C:\\Anaconda\\Project\\mcm2019\\lon_lan_cold\\lon_lan_cold2.csv",index=False)
```

程序 7	ARIMA 时间序列预测模型&基于 Prophet 的时间序列模型

```
import pandas as pd
import matplotlib.pyplot as plt
```

```
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.graphics.tsaplots import plot_pacf
from statsmodels.tsa.stattools import adfuller as ADF
import datetime as dt
import matplotlib as mpl
import numpy as np
import warnings
warnings.filterwarnings("ignore")

filename='C:\\Users\\lenovo\\Desktop\\ARIM\\ARIMA-master\\历年海洋温度.xlsx'
forrecastnum=5
data=pd.read_excel(filename,index_col=u'日期')
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.rcParams['axes.unicode_minus'] = False
data.plot()
plt.title('Time Series')
plt.show()
plot_acf(data)
plt.show()
print(u'原始序列的 ADF 检验结果为：',ADF(data[u'温度']))
D_data=data.diff(periods=2).dropna()
D_data.columns=[u'温度差分']
D_data.plot()
plt.show()
plot_acf(D_data).show()
plot_pacf(D_data).show()
print(u'1 阶差分序列的 ADF 检验结果为：',ADF(D_data[u'温度差分']))
from statsmodels.stats.diagnostic import acorr_ljungbox
```

```
print(u'差分序列的白噪声检验结果为：',acorr_ljungbox(D_data,lags=1))
from statsmodels.tsa.arima_model import ARIMA
data[u'温度'] = data[u'温度'].astype(float)
pmax=int(len(D_data)/10)
qmax=int(len(D_data)/10)
bic_matrix=[]
for p in range(pmax+1):
    tmp=[]
    for q in range(qmax+1):
        try:
            tmp.append(ARIMA(data,(p,1,q)).fit().bic)
        except:
            tmp.append(None)
    bic_matrix.append(tmp)
bic_matrix=pd.DataFrame(bic_matrix)
print(bic_matrix)
p,q=bic_matrix.stack().idxmin()
model=ARIMA(data,(p,1q)).fit()
model.summary2()
forecast=model.forecast(25)[0]
print('未来 25 年的预测值:',forecast)


######################基于 Prophet 的时间序列模型############
from fbprophet import Prophet
import numpy as np
import pandas as pd
sales_df = pd.read_excel('C:\\Users\\lenovo\\Desktop\\ARIM\\ARIMA-master\\要素.xlsx')
model = Prophet()
```

```
model.fit(sales_df)
future_data = pd.read_excel('C:\\Users\\lenovo\\Desktop\\ARIM\\ARIMA-master\\25 年.xlsx')
forecast_data = model.predict(future_data)
model.plot(forecast_data)
forecast_data.to_excel('C:\\Users\\lenovo\\Desktop\\ARIM\\ARIMA-master\\25 年-结果.xlsx')
```

程序 8	随机森林回归预测模型

```
mport pandas as pd
from sklearn.model_selection import train_test_split,cross_val_score,GridSearchCV
from sklearn.ensemble import RandomForestRegressor
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.decomposition import    PCA
import numpy as np
from sklearn.metrics import r2_score,mean_squared_error,explained_variance_score
model_data=pd.read_excel('C:\\Users\\lenovo\\Desktop\\gather.xlsx')
#取出车辙属性作为因变量
y=model_data['温度（因变量）']
#确定自变量
X= model_data[model_data.columns.difference(['温度（因变量）','年份'])]
#划分训练集和测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=33)
#进行特征相关性分析
corrmat = X_train.corr()
f, ax = plt.subplots(figsize=(16, 12))
sns.heatmap(corrmat, vmax=.8, square=True)#,annot=True
```

```
#相关性结果表明特征相关性高，需要降维
estimator = PCA(n_components=8)
X_train = estimator.fit_transform(X_train)
X_test = estimator.transform(X_test)
cf=estimator.components_
efc=estimator.explained_variance_
zb=estimator.explained_variance_ratio_
print(estimator.components_)# 输出主成分，即行数为降维后的维数，列数为原始特征向量转换为新特征的系数
print(estimator.explained_variance_)# 新特征 每维所能解释的方差大小
print(estimator.explained_variance_ratio_) #新特征 每维所能解释的方差大小在全方差中所占比例
#选参数
x =[]
y=[]
z=[]
for n_estimators in range(20,70,10):
    for max_depth in range(400,450,10):
        #参数 scoring 设置为 roc_auc 返回的是 AUC，cv=5 采用的是 5 折交叉验证
        auc = cross_val_score(RandomForestRegressor(n_estimators=n_estimators,max_depth=max_depth,random_state=0),X_train, y_train,cv=2,scoring='r2').mean()
        print('ok')
        x.append(n_estimators)
        y.append(max_depth)
        z.append(auc)
x = np.array(x).reshape(5,5)
y = np.array(y).reshape(5,5)
z = np.array(z).reshape(5,5)
```

```
fig = plt.figure()
ax = Axes3D(fig)
ax.plot_surface(y, x, z, rstride=1, cstride=1, cmap=plt.get_cmap('rainbow'))#'RdBu'
plt.xlabel('max_depth')
plt.ylabel('n_estimators')


rfr = RandomForestRegressor(n_estimators=100,max_depth=430,random_state=0)
rfr.fit(X_train, y_train)
print("Variable importance:\n", rfr.feature_importances_)#输出特征重要性
print("train r2:%.3f"%r2_score(y_train,rfr.predict(X_train)))#输出训练集决定系数
print("test r2:%.3f"%r2_score(y_test,rfr.predict(X_test)))#输出测试集决定系数
predata=pd.read_excel('C:\\Users\\lenovo\\Desktop\\25-gather.xlsx')
preresult=rfr.predict(predata)
```

程序 9	随机森林分类模型

```
import pandas as pd
from sklearn.model_selection import train_test_split,cross_val_score,GridSearchCV
from sklearn.decomposition import    PCA
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import classification_report
import warnings
from imblearn.over_sampling import SMOTE
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import RandomOverSampler
```

```
first_data=pd.read_excel('C:\\Users\\lenovo\\Desktop\\p3_mon.xlsx')
#取因变量
y=first_data['极寒']
#确定自变量
X= first_data[first_data.columns.difference(['极寒'])]
################################PCA 降维############
#进行特征相关性分析
corrmat = X.corr()
f, ax = plt.subplots(figsize=(16, 12))
sns.heatmap(corrmat, vmax=.8, square=True)#,annot=True
##相关性结果表明特征相关性高，需要降维
estimator = PCA(n_components=1)
X = estimator.fit_transform(X)
cf=estimator.components_
efc=estimator.explained_variance_
zb=estimator.explained_variance_ratio_
print(estimator.components_)# 输出主成分，即行数为降维后的维数，列数为原始特征向量转换为新特征的系数
print(estimator.explained_variance_)# 新特征 每维所能解释的方差大小
print(estimator.explained_variance_ratio_) #新特征 每维所能解释的方差大小在全方差中所占比例
############################切分训练集和测试集##################


X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.35, random_state=10)
rus = RandomUnderSampler(random_state=0, replacement=True)
X_train,y_train=rus.fit_sample(X_train,y_train)
X_train=pd.DataFrame(X_train)
y_train=pd.DataFrame(y_train)
#ros = RandomOverSampler(random_state=0)
```

```
#X_train,y_train = ros.fit_sample(X_train,y_train)
#X_train=pd.DataFrame(X_train)
#y_train=pd.DataFrame(y_train)
###############随机森林#################
print('Random Forest')
x =[]
y=[]
z=[]
import numpy as np
from mpl_toolkits.mplot3d import Axes3D
for n_estimators in range(130,180,10):
    for max_depth in range(50,100,10):
        #参数 scoring 设置为 roc_auc 返回的是 AUC，cv=5 采用的是 5 折交叉验证
        auc = cross_val_score(RandomForestClassifier(n_estimators=n_estimators,max_depth=max_depth,random_state=0),X_train, y_train,cv=2,scoring='roc_auc').mean()
        print('ok')
        x.append(n_estimators)
        y.append(max_depth)
        z.append(auc)
x = np.array(x).reshape(5,5)
y = np.array(y).reshape(5,5)
z = np.array(z).reshape(5,5)
fig = plt.figure()
ax = Axes3D(fig)
ax.plot_surface(y, x, z, rstride=1, cstride=1, cmap=plt.get_cmap('rainbow'))#'RdBu'
plt.xlabel('max_depth')
plt.ylabel('n_estimators')
clf = RandomForestClassifier(n_estimators=170, max_depth=50,random_state=10)
```

```
#参数：criterion='gini'或'entropy'
clf.fit(X_train, y_train)
predictions = clf.predict(X_test)
print("Variable importance:\n", clf.feature_importances_)#输出特征重要性
print(classification_report(y_test, predictions))
print("训练集：",clf.score(X_train, y_train))
print("测试集：",clf.score(X_test,y_test))
```