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

学　　校		复旦大学
参赛队号		**19102460057**
队员姓名	**1.**	詹晨
	2.	范祥宇
	3.	郑子琳

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

题 目 全球变暖问题

摘 要：

本文主要从数据分析角度研究全球变暖问题。虽然全球变暖日趋严重，但由于全球变暖停滞以及局地极寒等原因，造成了公众对全球变暖的怀疑。建立可解释性强的预测模型将有助于增强大众对于全球气候变化的认识。本文结合现有的统计数据，运用多种统计方法，例如 XGBoost、厚尾随机波动率模型、因子分析和 K-Means 聚类等方法进行气候变化的规律探索；提出了（1）预测全球气候变化的 XCSM 模型，（2）用于分析局地极寒天气的因子分析-多元非线性回归模型，（3）具有完整性和统一性的地-海-气耦合场理论。

问题一是挖掘加拿大地区以及海洋表面温度的变化规律。对于加拿大地区温度数据的分析,本文提出了一种有别于方差的、可刻画序列数据波动程度的指标 Intensity。为分析气候与位置数据的相关性，本文采用气候数据对测量基站进行了 K-Means 聚类，总结出气候类型与位置信息的关联规律。对于海洋温度数据的分析，分别从“时间维度”、“厄尔尼诺”以及“与陆地温度之间的关系”三个方面展开。本文用移动平均结合“MA(1)模型”刻画海洋温度的变化；对于厄尔尼诺现象通过分析热带太平洋东部及中部 28 年间的温度数据，取得该地区海洋温度变化的趋势,并通过热带海洋表面温度图的变动趋势，印证了厄尔尼诺现象的海洋温度变化特征。

问题二是建立一个全球气候变化预测模型。为了兼顾模型的预测准确性与可解释性，本文提出了基于 XGBoost、协整模型以及厚尾随机波动率模型的改进模型 XGBoost-Cointegration-SV Model(简记为 XCSM)。模型采用 XGBoost 进行变量重要性分析，并在此基础上进行随机波动率模型的时间序列建模。在模型可解释性与稳健性方面，采用协整模型以利于分析各因素对气候变化的影响大小与方向；在模型预测方面，建立了 XGBoost-SV 模型与 Cointegration-SV 模型分别进行时间序列预测，并采用 Bagging 思想，结合了 XGBoost-SV 与 Cointegration-SV 模型的预测结果，以获得更稳健的预测。预测结果为未来 25 年全球气候变化平缓，同时具有一定下降趋势。

对于问题四，本文建立了一个通过“全球气候因素”分析“局地极寒”的宏观模型。本文在气候学理论基础上，建立了因子分析-多元非线性回归模型。为解释全球变暖与局地极寒之间的矛盾，模型基于热量体系与厄尔尼诺现象的交互效应，认为全球变暖在厄尔尼诺现象的影响下，导致了局地极寒现象的产生；在局地极寒天气建模方面，从统计学角度提出了与气候学理论相统一的宏观模型，并反映了全球气候变化与局地气候异常的连锁性与复杂性。

对于问题四，本文将气候预测模型与局地极寒模型相结合，提出了具有完整性和统一性的“地-海-气耦合场”理论，以解释全球变暖、局地极寒等全球气候异常。耦合场理论将二氧化碳浓度、日照通量和太平洋年代际振荡假定为模型外生变量。外生

变量的变动使得全球热量体系（包括全球海面平均温度、全球海洋热量总存储）以及“厄尔尼诺-南方涛动-北极涛动”等气候体系发生改变，从而进一步影响全球气候，并带来局地气候异常。本文提出“全球气候异常”以取代“全球变暖”的概念，并从气候趋势及复杂性两方面来说明新概念的优势。

关键词： XGBoost　随机波动率模型　因子分析　K-Means　气候预测

目录

1 问题重述

1.1 问题背景

全球变暖涉及到人类生存环境的问题，是人们所关注的焦点。全球变暖是由于温室效应的累积而导致的现象，由于化石燃料的大量焚化，以及在为开发新土地而砍伐并焚烧森林，所产生的温室气体因为自身能够吸收地面辐射中的红外线的特性，使地气系统中的能量不断累积，因此影响了地气系统的能量平衡，使地球温度上升。在全球变暖的环境下，科学家认为可能会导致了更多极端气候的发生，诸如干旱、洪涝等降水重新分配的问题，以及热浪、极寒现象等极端气温的发生，此外全球暖化导致冰川及冻土的融化，使海平面上升。海平面的逐步上升威胁到低洼沿海国家及地区的人类生存空间，也增大了风暴潮的威胁。自 21 世纪以来，全球每年平均气温上升非常缓慢，近乎没有增长，且美国地区出现大规模的极寒现象，因此便开始出现对全球变暖质疑的声音。而全球每年平均气温上升缓慢的现象，被称为“全球气候停滞现象”。

出现意见分歧的原因在于观察的范围和角度的不同。质疑全球变暖的人们观察角度是在于个人的直观感受，受限于个人感受的范围与缺乏长时间的观察记录，是一种基于天气现象的判断，而天气是在短时间区域内对大气状态和变化的总称。而全球变暖关注的是长时间、大规模的气候现象问题。但目前对全球气象暖化的研究有两大问题，一方面在于气候的时空数据不足，使对长时间的气候观察、计算和研究产生了很大的困难;另一方面在于海洋温度会对全世界的气候现象产生极大的影响，而海洋温度本身具有震荡的特征，如厄尔尼诺现象，以及拉尼娜现象，厄尔尼诺现象使太平洋的海洋温度增加，致使全球气候的变化，而拉尼娜现象是在出现厄尔尼诺现象后交替出现的太平洋海洋温度降低，也会对全球气候产生一定的影响。海洋温度的震荡性使气候研究更加困难。

1.2 问题提出

核心问题：利用现有数据建立简化的气候模型和极端天气模型，通过简化的气候模型和极端天气模型，让非专业人士能对全球气候变化的状态和形势有一定的认识和理解。同时需要去解释极端气候的发生，去发现、求证对气候变化会产生影响的因素。透过个人对气候变化意识的增强，进一步督促政府制定相应的对策。

问题一：从加拿大各地天气变化的历史数据中挖掘加拿大各地区的温度时空变化趋势，并发现海洋表面温度历史数据的规律。

问题二：建立一个需要考虑到地球的吸热、散热、海洋温度变化等要素的气候变化模型，并对未来 25 年气候变化进行预测。

问题三：发现极寒天气与气候变化的关联性，并建立相应的极端天气模型。并通过数据说明全球变暖与局地地区极寒现象是否存在矛盾之处。

问题四：用通俗文字解析:“全球变暖了，某地今年冬天特别冷”之间的关系，并用一个新概念去替代“全球变暖”，使概念可以反映出气候变化的复杂性和趋势。

2 模型假设

下面提出几个假设，以便使得模型的构建合理。

- 假设基站是均匀分布的，测量值可以代表所在州的气候情况。
- 假设气候的情况主要体现在气温和降水量两个方面。
- 全球海洋平均温度与陆地温度符合正态分布。

3 符号说明与名词解释

3.1 符号说明

这一部分主要介绍正文中出现的较为重要的数学符号，但正文部分针对较为重要的数学符号不免重复说明。

表 3-1 重要符号说明

符号	说明
f_t	第t步迭代时需要拟合的树
T	这棵树包含T个叶节点
w_j	第j个叶节点上的估计值
γ	调优参数
λ	调优参数
C_i	K-Means 的第i个簇向量
μ_i	K-Means 的第i个簇向量的均值
M_t	t时刻的移动平均后的值

3.2 名词解释

厄尔尼诺年的定义：NIÑO3 指数距平高于+0.7 标准差。

拉尼娜年的定义：NIÑO3 指数距平低于-0.7 标准差。

表 3-2 名词解释

名词	解释
NIÑO1+2	极东热带太平洋 SST
NIÑO3	中部东部热带太平洋 SST
NIÑO4	中央热带太平洋 SST

SOI	南方涛动指数
NIÑO SST	本文通过 SST 计算所得 NIÑO 指数
GMST/GMT	全球平均地表温度
ENSO	厄尔尼诺-南方涛动
SST	海洋表面温度
MEI V2	多元 ENSO 指数
PDO	太平洋年代际涛动
AO	北极涛动
Solar Flux	10.7cm 太阳光通经

4 问题一 加拿大地区及海洋表面温度变化探索

这一节将进行加拿大地区温度数据以及海洋表面温度的数据探索，目的是挖掘加拿大各地温度时空变化规律以及海洋温度的变化。对于加拿大地区的温度数据将探索以下四个方面：

- 近十年来是否存在全球变暖停止状态
- 加拿大各州的年平均气温的变化趋势与程度
- 各州的气温变动程度是否在逐年增大
- 各州的气温与降水情况与其经纬度及海拔高度有怎样的关系

对于海洋表面温度的数据探索，将分为两个方面进行分析：

- 海洋的温度变化存在怎样的规律，包括时间维度与空间维度
- 海洋温度与陆地温度是否存在着相关关系

4.1 加拿大地区温度数据的探索

数据来源于加拿大气象网站上各州基站的观测数据，首先由于部分基站的数据缺失严重，我们根据数据的缺失情况进行筛选，筛选出缺失数据小于百分之八十的数据。进一步，考虑到数据处理时的时间效率，对于每个州选择 10 个基站作为该州的数据样本。

4.1.1 全球变暖停滞状态的印证

全球变暖停滞是指进入 21 世纪以来，10 年间全球全年平均气温上升率仅为 0.03 摄氏度，几乎没有变化。本文利用 2009 年至 2018 年加拿大的气温数据，统计每年的平均气温，如图 4-1 所示，发现平均气温非但没有上升，还有轻微的下降，在 2018 年达到了 10 年间的平均气温最低值 1.36 摄氏度。值得思考的是，是否由于各州的情况有所不同，而导致平均来看全国的平均气温变化并不显著，所以下面对各州分别进行统计。

图 4-1 加拿大平均气温变化图

4.1.2 加拿大年平均气温变化情况

对各州的年平均气温分别进行统计，为了反应温度变化的大小和方向，分别计算 2009 至 2013 年、2013 至 2018 年以及 2009 至 2018 年的温度的变化，结果如图 4-2。其中，色块颜色越深代表温度上升的程度越大，接近白色的色块所在州为温度下降的州。

可以看出，2009 到 2013 年各州的温度变化都不明显，最大的上升幅度为 1.33 摄氏度，最大的降温幅度为下降 0.66 摄氏度；2013 至 2018 年处于东部的几个州，即努纳武特、安

大略、魁北克地区有较大幅度的降温，特别是魁北克地区年平均气温下降了 5.63 摄氏度。

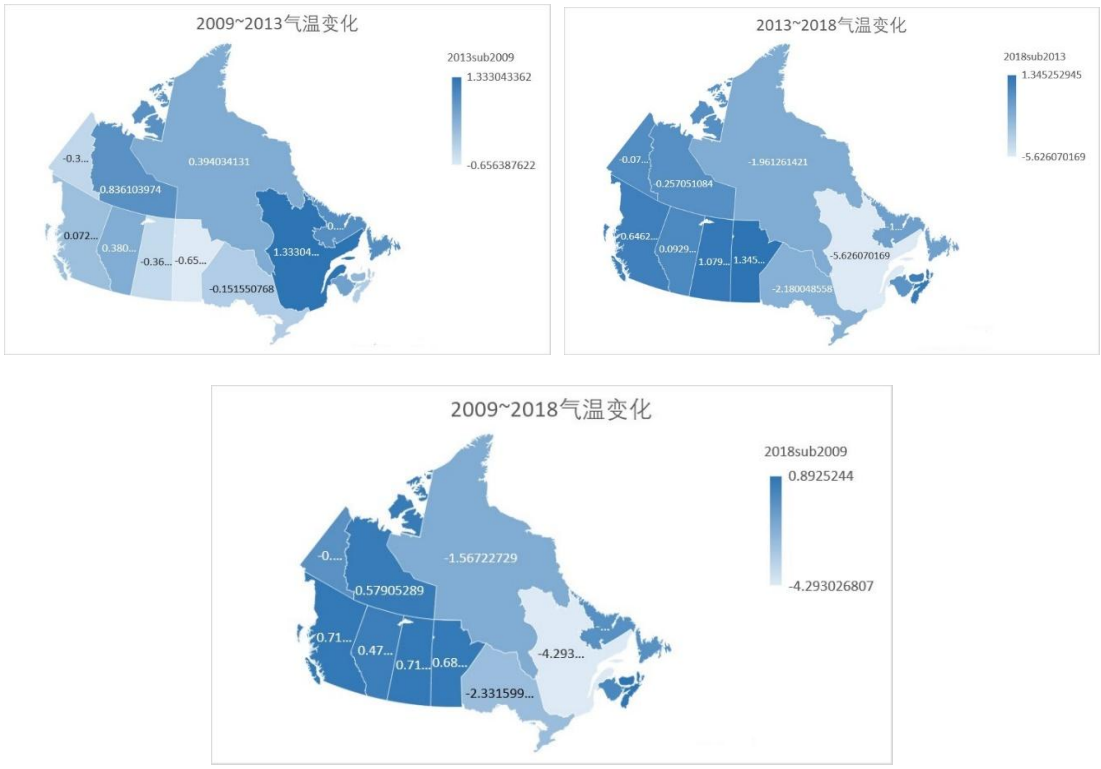


图 4-2 加拿大各州气温变化图

进一步考虑，如果温度的考察时间细致到月份，将易于看出是否存在每年温度的波动率的变化特点，如图 4-3 为各州每月的温度变化情况。从图中看出虽然每个州的温度波动率变化不同，比如努纳武特地区的每年温差大，而不列颠哥伦比亚省的温度变化小。但是不论对于哪个州，都看不出温度的波动程度会随着年份的增加而出现明显改变。

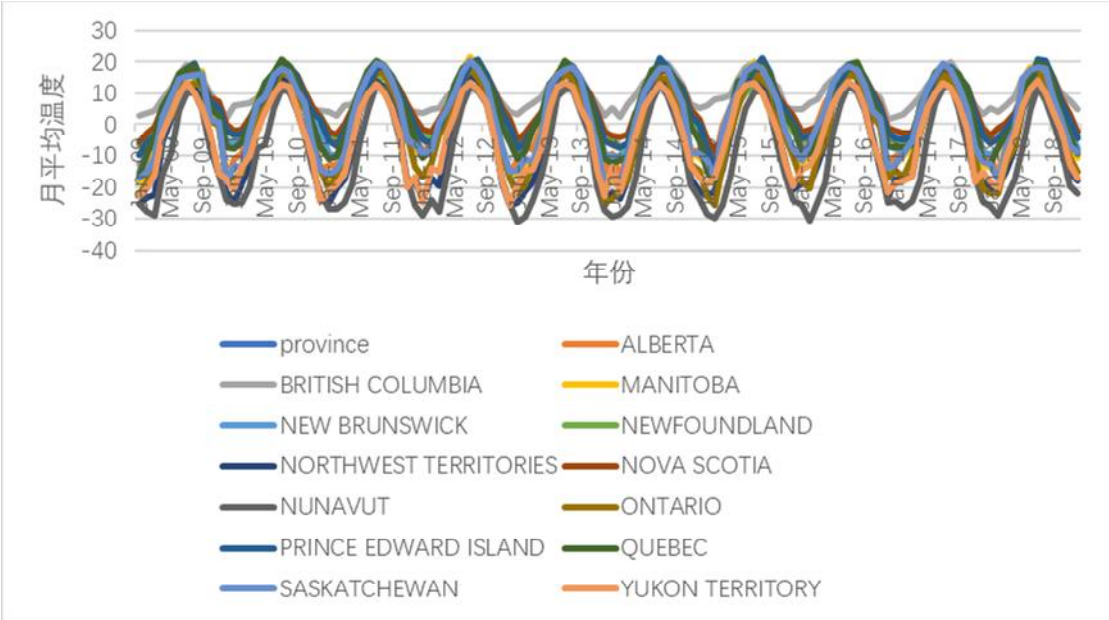


图 4-3 加拿大各州月度气温变化图

4.1.3 各州气温变化波动情况

在上一小节的可视化的过程中，我们看不出气温波动率如何变化，需要通过统计指标刻画温度波动率来定量考察。波动率的刻画可以分为两种，第一种是传统的标准差统计量，第二种是本文提出的一种结合序列差分设计的刻画时间序列波动程度的指标。

对于第一种方法，即用标准差刻画波动程度，也就是用将每个州的月度平均气温计算出来，再将属于同一年的月度数据计算标准差。这种方法的结果如图 4-4。发现气温的波动情况不存在明显变化。

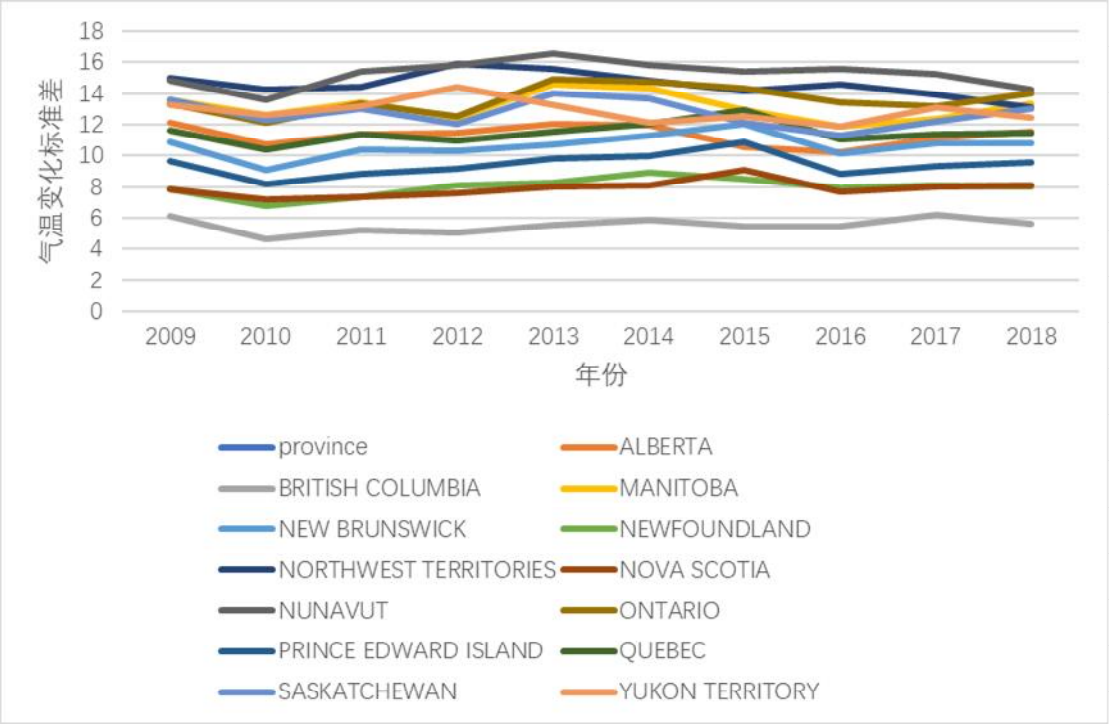
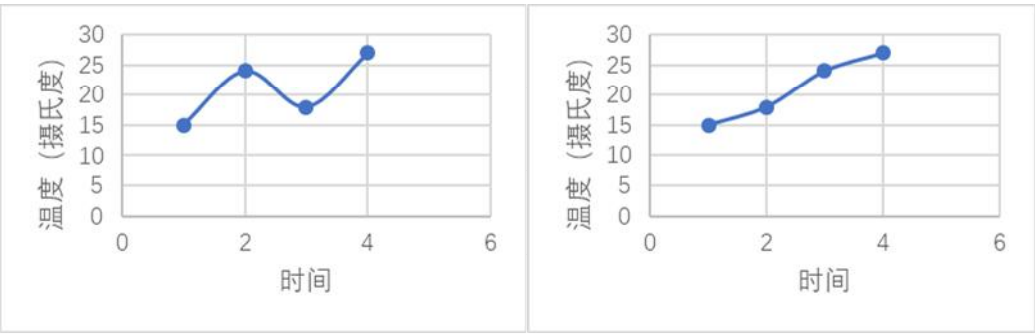
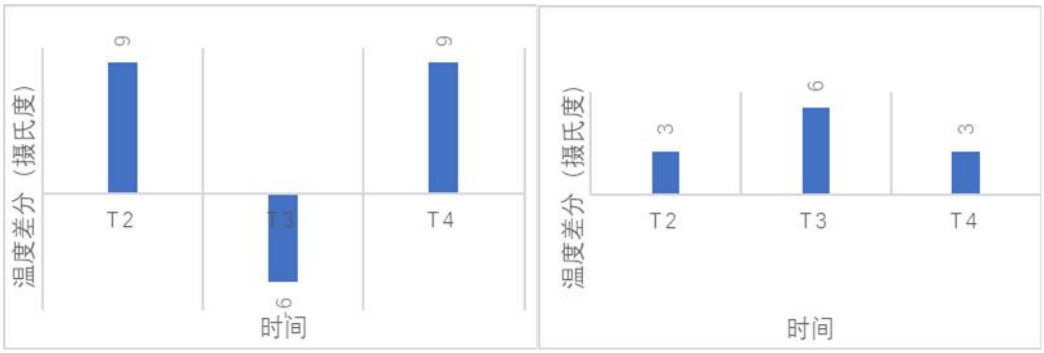


图 4-4 加拿大各州温度变化标准差

但是值得注意的是，第一种方法在刻画温度波动情况的时候存在局限性。可以通过图 4-5 展示标准差的弊端，同时引出更好的一种度量波动性的方法。



(a) (b)



(c) (d)

图 4-5 波动率度量改进方法说明示意图

其中，a 图与 b 图的四个点数值相同，所以标准差相同，但是 a 的变化曲线比 b 曲折，可以将 a 对应到不正常的气温变化，b 对应于平缓的温度上升过程，一个合理的波动情况度量指标应当为 a 的计算结果大于 b 的计算结果，本文提出用差分的绝对值相加来度量温度变化的剧烈程度，即公式(4-1)。如图 c 为 a 图差分后的结果，d 图为 b 图差分后的结果，Intensity 的公式 a 图对应结果大于 b 图。借助 Intensity 指标计算的各州温度剧烈程度变动指标如图 4-6。随着时间推移，波动程度有减弱趋势。

$$\text{Intensity} = \sum_{t=1}^{n} |T_t - T_{t-1}| \qquad (4\text{-}1)$$

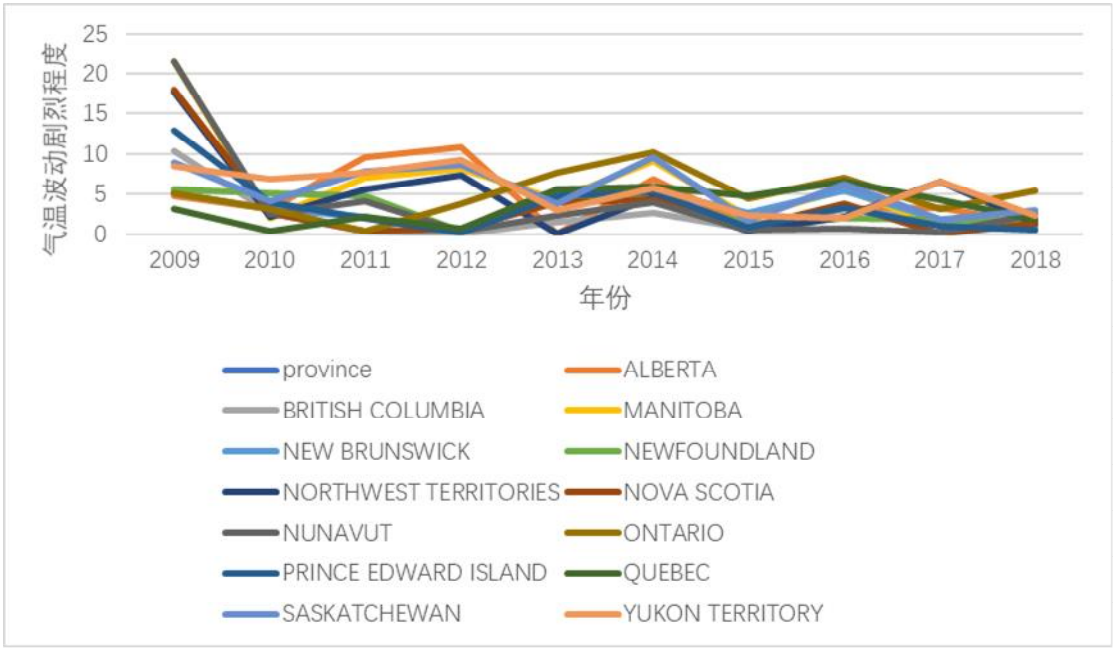


图 4-6 加拿大各州气温波动剧烈程度变化图

4.1.4 各州气候与位置数据的关系

对于每个基站的测量的气象数据进行统计，数据种类包括每日最高气温、每日最低气温、每日平均气温、总降雨量、总降雪量。先探索各个变量之间的相关关系，如图 4-7 所示，其中颜色越接近蓝色代表对应的所属行和所属列的变量之间的相关性越接近于 1，而越接近于深褐色则代表两个变量的相关系数越接近于-1。纬度与气温有正相关关系，经度与气温有负相关关系，海拔高度与其他变量的关系并不显著。纬度与气温正相关是由于太

阳光照的角度原因造成的北部气温较低，而经度与气温负相关按照加拿大的气候资料的说法，是受西风影响，西部气候温和湿润，东部较冷。

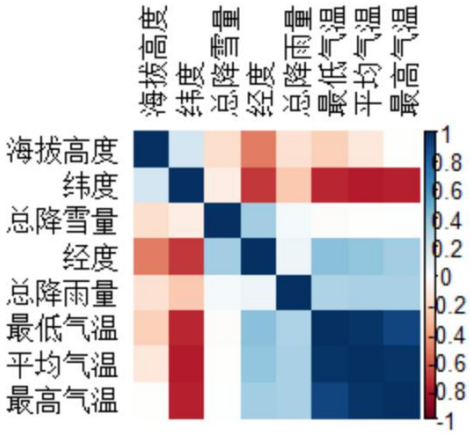


图 4-7 气候因素相关系数图

对基站的数据进行聚类，用无监督的方法探索不同基站的气候特点是否存在潜在类别。进一步再探究不同的类别与基站所处的经纬度以及海拔高度是否存在相关性。

首先根据不同的聚类个数的组内总方差的变化选择合适的聚类个数，见图 4-8，当聚类个数大于 4 时，组内方差的减小程度趋缓，本文选择聚成四类。采用 K-Means 聚类。K-Means 的方法可以描述为，假设组划分为$(C_1, C_2, \ldots, C_k)$，我们的目标是最小化平方误差$\mathbf{E}$，表达式为公式(4-2)：

$$\mathbf{E} = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|_2^2 \tag{4-2}$$

其中μ_i是簇C_i的均值向量，有时也称为质心，表达式为公式(4-3)：

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{4-3}$$

用加拿大各个基站的数据聚成四类，得到的每个类别的簇中心见表 4-1。类别 1 的特点是气温温和且降雪量多，类别 2 是天气较热且将于充沛，类别 3 为天气寒冷且降雪量多，而类别 4 是天气较冷且降水充沛。

最终聚类的结果结合 t-SNE 降到 2 维，如图 4-9 所示。不同的类别用不同的颜色标注，属于统一颜色的基站所属的类别相同。

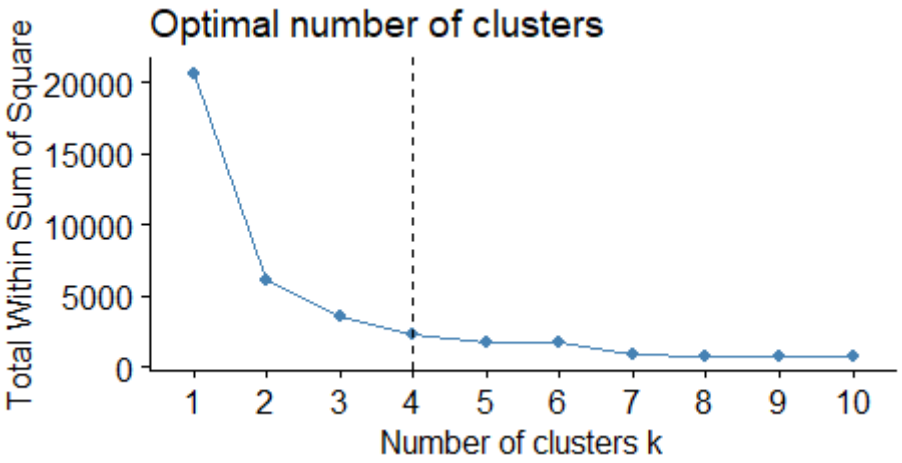


图 4-8 群组内方差随聚类个数变化图

表 4-1 类别的簇中心

类别	最大温度	最小温度	平均温度	总降雨量	总降雪量
类别 1	9.663	-0.923	4.357	0.678	0.467
类别 2	13.708	5.555	9.607	1.282	0.361
类别 3	-8.971	-16.470	-12.732	0.071	0.438
类别 4	3.588	-7.314	-1.828	0.242	0.457

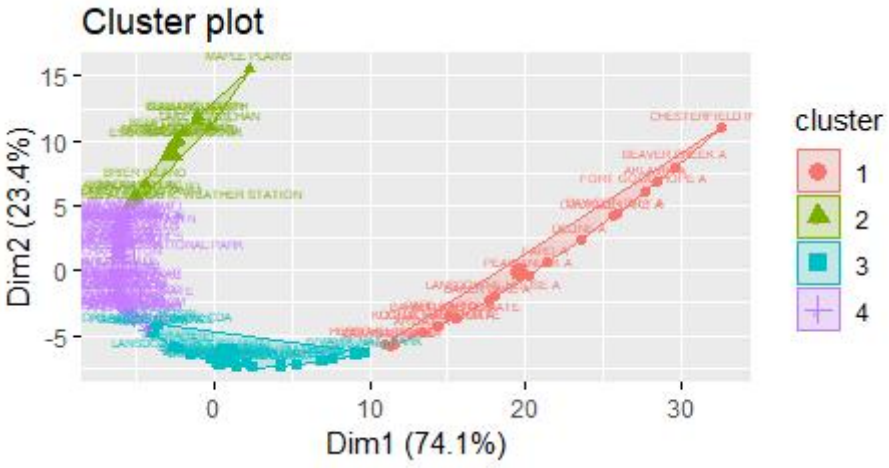


图 4-9 聚类的结果经 t-SNE 降维的展示图

进一步，我们探究气候的不同类别与地理上的分布的关系，如图 4-10，将抽样的基站位置标注在相应的经纬度，并用不同的颜色来表示其所属的不同类别，看到气候类型 2，即天气较热且将于充沛的气候在加拿大南部的沿海城市有分布，属于类别 3 天气寒冷且降雪量多的地点多在西北地区，而类型 1 多分布在东南方向。说明经纬度与气候类型存在联系。

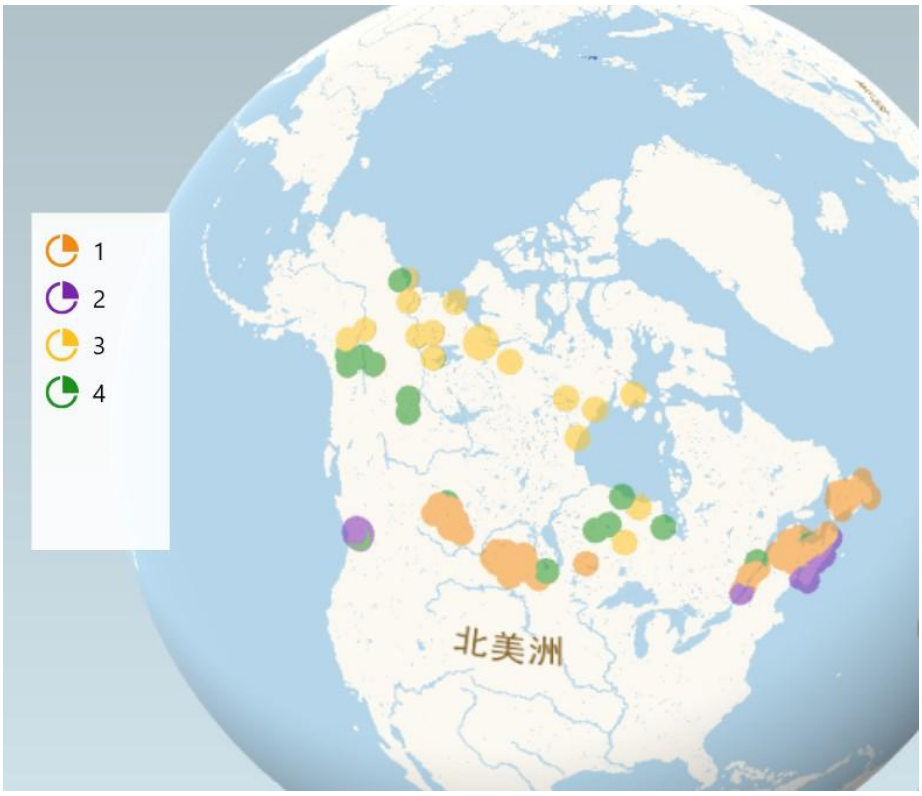


图 4-10 气候的不同类别与地理上的分布的关系

4.2 海洋温度数据的探索

海洋温度的探索将从两个方面展开，第一个角度是只考虑海洋为研究主体，分别从时间维度和厄尔尼诺现象进行探究。第二个角度是考虑海洋温度和陆地温度之间的关系。下面将分这两个部分介绍。

4.2.1 海洋温度变化规律

海洋温度的变化分为时间维度变化、厄尔尼诺现象两个方面。

4.2.1.1 时间维度上的变化规律

这一部分的数据来源于 NASA 网站上提供的全球海洋平均温度，时间跨度为 1880 至 2018 年，每年一个数据。可视化如图 4-11 其中实线为海洋温度的变化，虚线为 5 步移动平均的结果，其中移动平均的计算公式为公式(4-4)，其中 N 这里取为 5，从移动平均曲线可以看出从 1910 年开始海洋温度持续上升，初步判断此时间序列为非平稳时间序列。按照式(4-5)计算实际序列减去移动平均后的残差序列值，得到去除趋势的序列。对此序列进行平稳性检验，计算自相关系数与偏自相关系数。其中自相关系数与偏自相关系数公式分别为式(4-6)和式(4-7)。根据 AIC 值的大小选择模型为 MA（1），借助 R 语言估计参数得到结果，即（4-8）式。该模型可用于预测海洋温度的变化。

$$M_t = \frac{y_t + y_{t-1} + \cdots + y_{t-N+1}}{N} \tag{4-4}$$

$$x_t = y_t - M_t \tag{4-5}$$

$$\rho_x(h) = \frac{Cov(x_{t+h}, x_t)}{Cov(x_t, x_t)} \tag{4-6}$$

$$\varphi_{kk} = \mathrm{Corr}(x_t, x_{t+k} | x_{t+1}, \ldots, x_{t+k-1}) \tag{4-7}$$

$$\begin{cases} x_t = \mu + \varepsilon_t - 0.6452\varepsilon_{t-1} \\ \theta_q \neq 0 \\ E(\varepsilon_t) = 0,\ Var(\varepsilon_t) = 0.6452, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \end{cases} \tag{4-8}$$

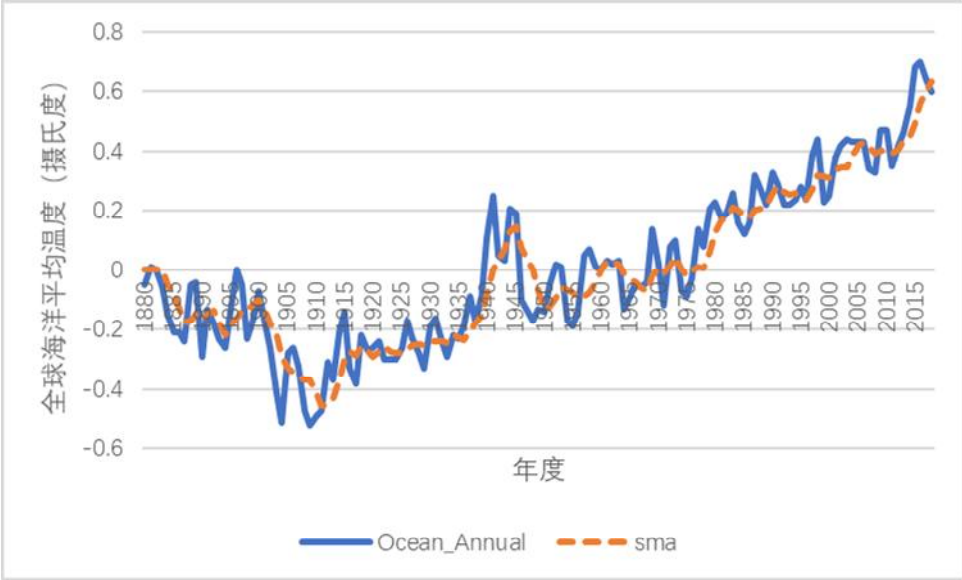


图 4-11 海洋温度变化及移动平均

4.2.1.2 热带海洋气候异常现象

厄尔尼诺现象是不规则的发生在热带海洋的异常气候现象，其显著的特征是海洋温度的上升以及气候模式的反常。而南方涛动与厄尔尼诺现象并称为“ENSO”，南方涛动是热带海洋异常气候现象在海洋温度和气压两方面的体现。在厄尔尼诺现象发生后，伴随而来的是拉尼娜现象，与厄尔尼诺现象相反，使太平洋东部及中部海洋温度持续降温。

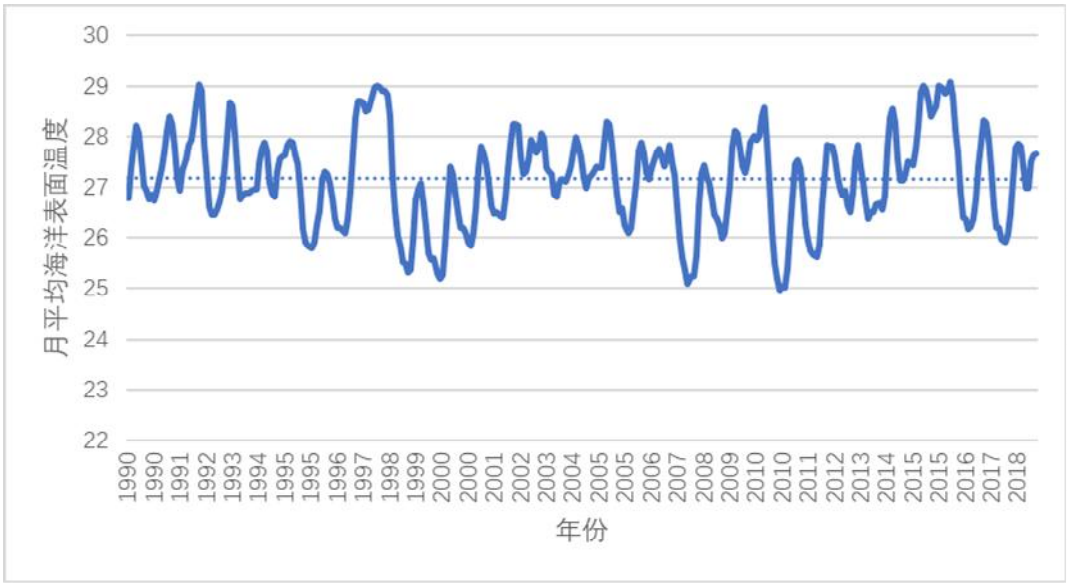


图 4-12 太平洋东部及中部的热带海洋月平均海洋表面温度图

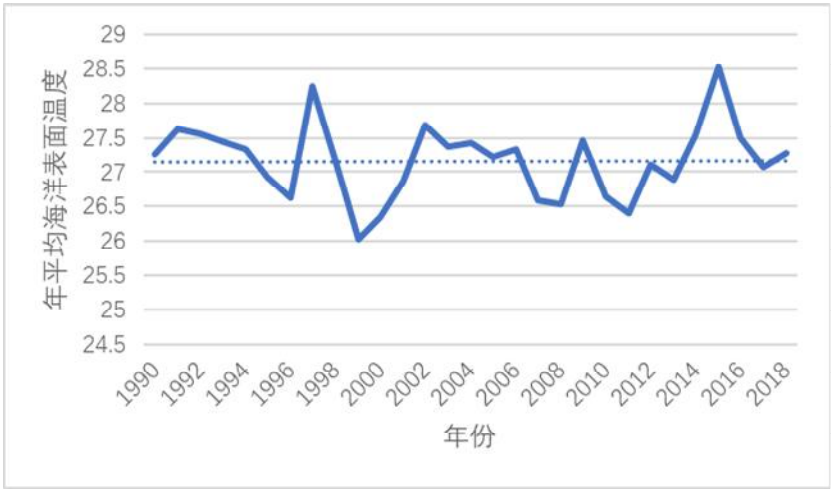
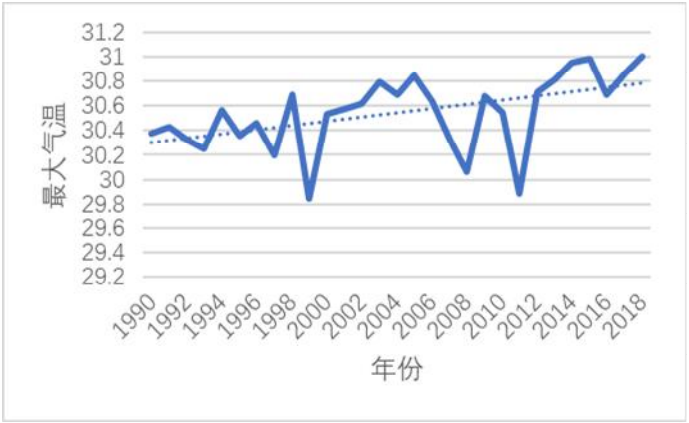


图 4-13 太平洋东部及中部的热带海洋年平均海洋表面温度图

从太平洋东部及中部的热带海洋 1990 至 2018 年间的月平均及年平均海洋表面温度数据，可以看出太平洋东部及中部的热带海洋在 1990 年到 2018 年间,有两个主要的峰值，一个是在 1997 年，另一个在 2015 年，分别对应上 1997/1998 及 2015/2016 的厄尔尼诺事件。



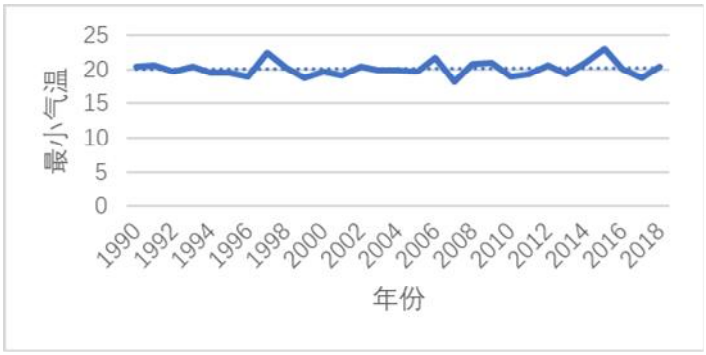


图 4-14 太平洋东部及中部的热带海洋最大温度及最小温度图

从太平洋东部及中部的热带海洋 1990 至 2018 年间最小温度图，并无发现有上升或下降的趋势，但从太平洋东部及中部的热带海洋 1990 至 2018 年间最大温度图中，可见太平洋东部及中部的热带海洋的每年海洋温度中的最大值有上升的趋势。

在近年间的厄尔尼诺现象中，在 1997/1998 ， 2015/2016 发生了两次超强厄尔尼诺事

件，下图为对 1996 至 1999 年间，以及 2014 至 2017 年间，每年 10 月到次年 3 月的全球海洋表面温度分布。

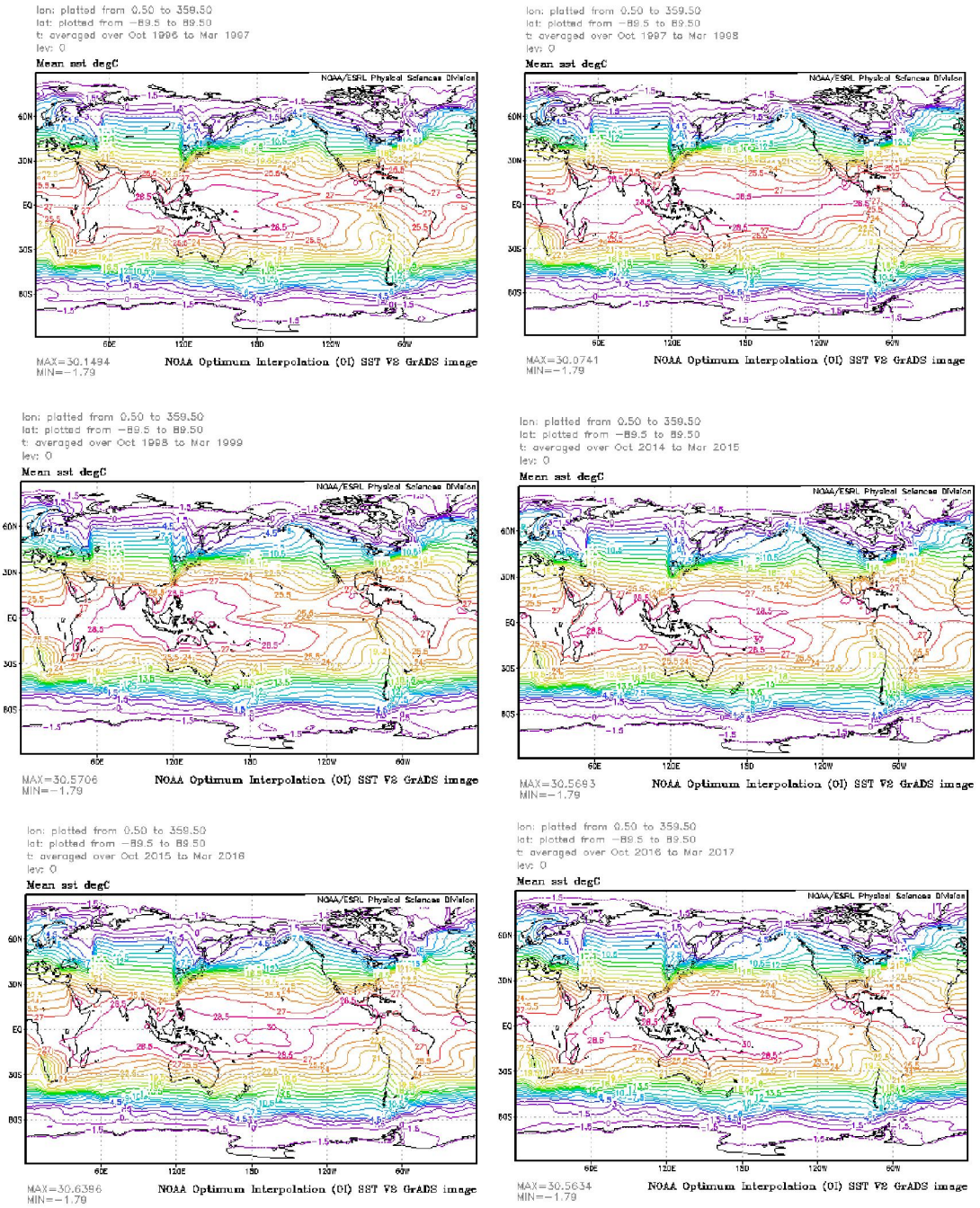


图 4-15 全球海洋表面温度图

由上图数据中可见在 1997 至 1998 年间，以及 2015 至 2016 年间赤道附近的平均海洋表面温度相较起邻近年份的对应时间，海洋表面温度高温面积范围的扩大，主要受影响的地区为太平洋中部及东部地区。而 2014 至 2017 年间对 1996 至 1999 年间 10 月至 3 月的相对平均海洋表面温度有所提升，太平洋中部区域的平均海洋表面温度达到了 30 度的高温，由此再次验证了海洋表面温度的上升趋势。

厄尔尼诺现象所伴随的气候反常现象，诸如涝灾、旱灾、台风等，各种不同类型反常气候的叠加增加了研究全球气温变化的复杂性，因此同时需要气候模型和极端天气模型来

研究全球气温变化，本文第五、六章将分别谈及气候模型和极端天气模型。

4.2.2 海洋温度与陆地温度的相关性

利用来源于 NASA 网站的陆地温度数据与海洋温度数据绘制散点图 4-16，粗略判断两者有正相关关系。进一步，进行皮尔森相关性检验。

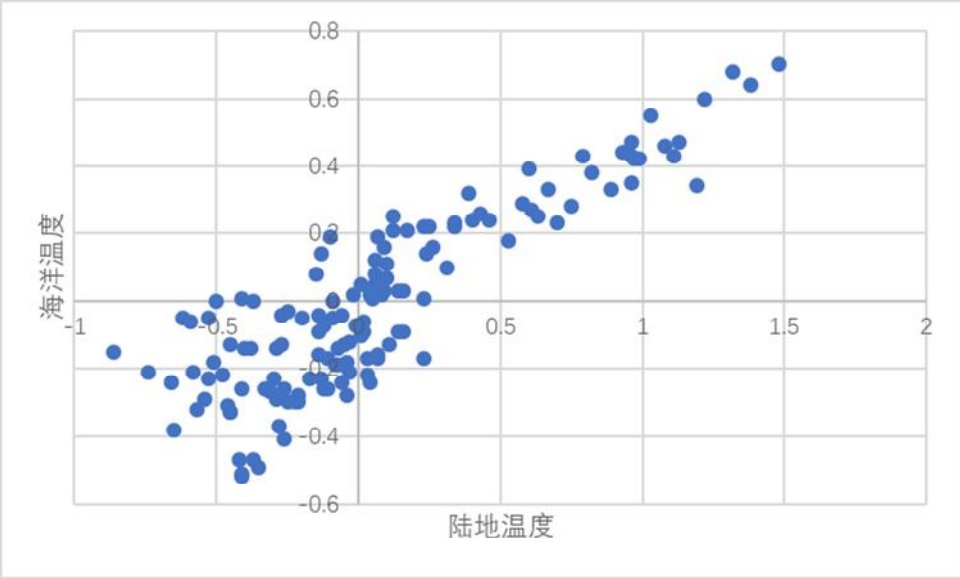


图 4-16 陆地与海洋温度散点图

皮尔森相关性检验的假设和检验统计量见(4-9)式。在保证 95%置信水平下进行单边检验，t 统计量计算结果为 20.984，p 值远远小于 0.05。拒绝原假设，接受备择假设，即结论为海洋温度与陆地温度的在 95%的置信水平下显著正相关。两者的样本相关系数为 0.873。

$$H_0:\rho = 0, H_1:\rho > 0$$

$$\mathbf{t} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$其中\boldsymbol{r} = \frac{\Sigma(\frac{x-\bar{x}}{s_x})(\frac{y-\bar{y}}{s_y})}{N} \qquad (4\text{-}9)$$

5 问题二 气候变化预测模型

XGBoost-Cointegration-SV Model

5.1 问题的描述与分析

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，并且模型至少要考虑地球吸热、散热以及海洋的温度变化等要素。模型的目的之一是保证预测的准确性，另一个目的是有利于非专业人士理解和人士全球气候变化的态势，有助于解释气候的产生原因。也就是说，我们的模型需要兼顾预测准确性与可解释性。

对于非线性模型，例如神经网络，能在数据量较大的情况下有更好的预测性能，但是可解释性差。而线性模型，例如线性回归、ARMA 等模型可解释性强，但是预测性能弱。本文为了综合模型的预测准确性与可解释性，提出了基于 XGBoost、协整模型以及随机波动率模型的改进模型 XGBoost-Cointegration-SV Model ，简记为 XCSM。XCSM 的模型示意图如图 5-1。

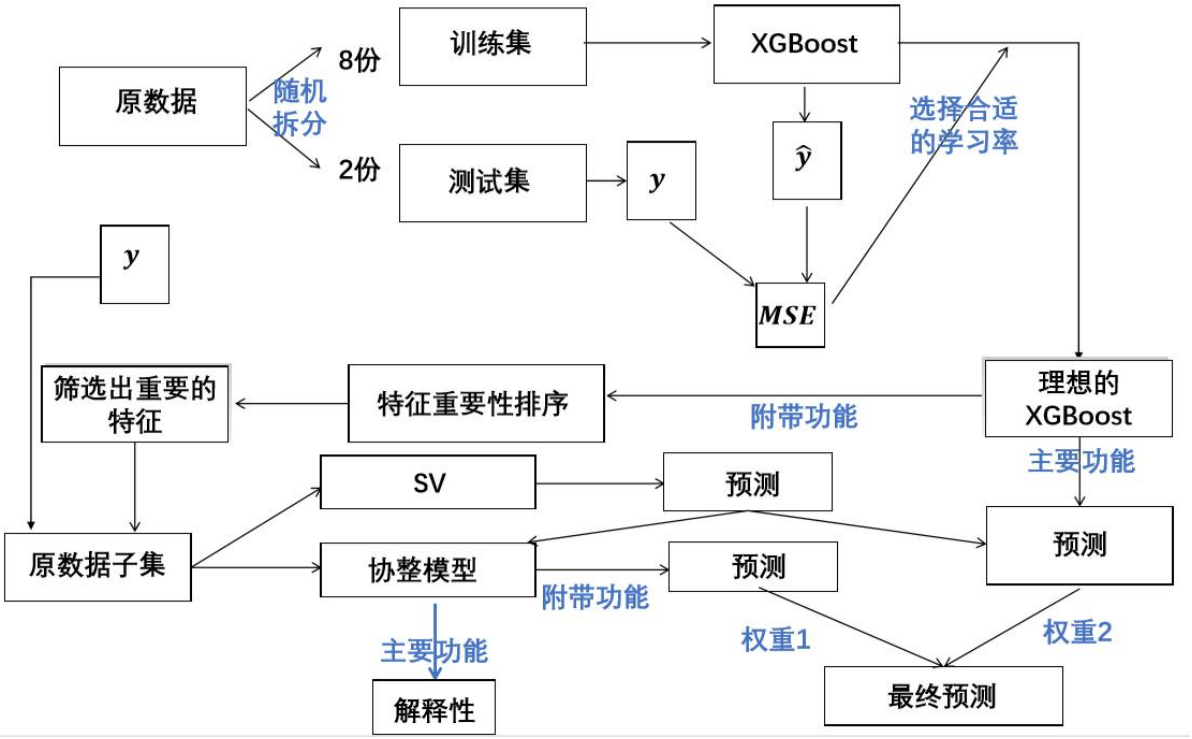


图 5-1 XCSM 示意图

5.2 数据的准备

5.2.1 变量选取与处理

气候变化预测模型的变量主要包括地球表面平均温度、二氧化碳浓度、太阳光通量以及厄尔尼诺海洋涛动指数。本文将地球表面平均温度作为衡量全球变暖所导致的气候变化的主要衡量指标，原因在于通常定义的全球变暖会导致全球气温升高。近地表气温和地表温度高度相关。使用二氧化碳浓度来衡量温室气体浓度是一种行之有效的办法，中外很多

温室效应相关论文都用二氧化碳浓度来替代温室气体浓度，原因在于二氧化碳占据了温室气体中绝大多数体积。虽然不少学者认为，温室气体除二氧化碳外，甲烷、氧化亚氮等化合物也有极强的造成温室效应能力，但是二氧化碳对于全球升温的贡献是最大的。同时，一般而言，二氧化碳排放的增加与其他温室气体排放的增加在比例上是保持一致的。选择太阳光通量作为模型的因素，原因在于不少研究温室效应停滞的论文中，都认为太阳光通量的变化是导致全球变暖停滞的原因之一，因此本文将该因素纳入考虑。厄尔尼诺现象以及海洋涛动对全球海气系统的影响是极其显著的。不少著作认为，全球变暖的停滞的原因，很大一部分归咎于海洋对于全球热量的吸收。研究认为，2010 年后，在多次厄尔尼诺现象出现后，大气中的热量经由热带海气耦合模式进入海洋中，导致海洋热量储存上升。进一步的，由于厄尔尼诺现象，海洋中的热量由表层储存逐渐沉浸为深层储存。热带海气耦合系统的存在和剧烈反应，可能使得全球变暖现象发生了延缓。

在海洋涛动变量的选择中，本文首先选取了 ENSO 相关指数，包括 NIÑO1,2,3,4, MEI V2, SOI。此外，本文采用了第一问中计算所得的 160° E~90° W, 5° S~5° N 的太平洋热带海洋表里面温度（SST）作为刻画厄尔尼诺现象的另一个指数 NIÑO SST。

在使用数据做建模和预测的过程中，事先不能确定哪些变量会对温度的预测有显著影响，所以首先，采用了 XGBoost 方法来进行变量重要性分析，从而确定对气候变化模型影响显著的自变量。

5.2.2 数据来源

地球表面平均温度、二氧化碳浓度、太阳光通量、海洋表面温度、海洋热量存储以及厄尔尼诺海洋涛动指数主要来自于 NOAA、NASA、NCAR、世界银行和 Berkeley 发布的公开数据集。所选取的数据中，除了海洋热量存储数据是年度数据外，其他数据都是月度数据，时间跨度从 1982 年 1 月至 2018 年。

5.2.3 数据预处理

由于本文所使用的海洋热量存储数据是年度数据，在时间跨度上较大，在数据精度上不够精细，因此需要通过预处理，将数据近似处理为精细的月度数据，在建模方面能进一步提高模型的准确度，避免数据误差带来的模型整体误差。

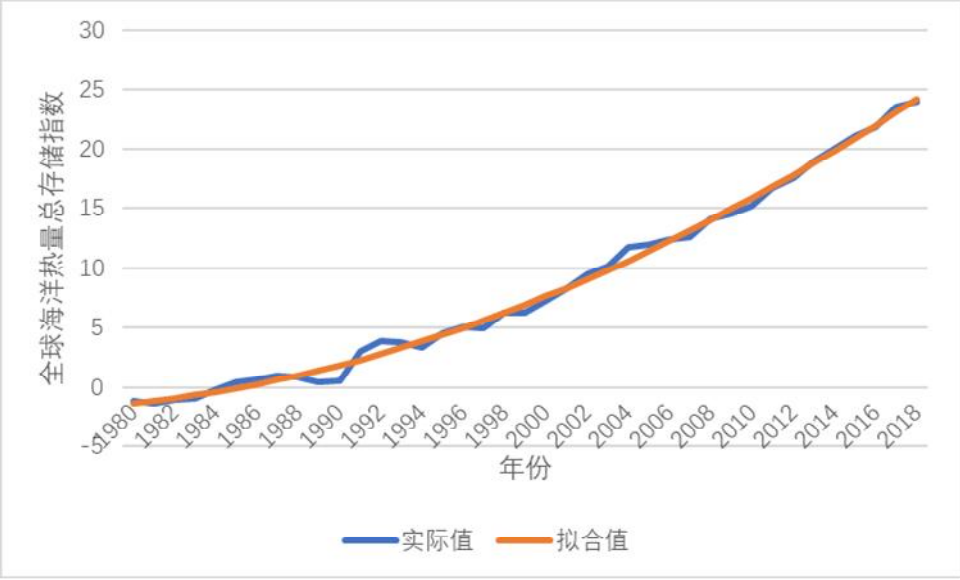


图 5-2 全球海洋热量总存储指数拟合图

表 5-1 全球海洋热量总存储二次回归拟合统计量表

	系数估计值	标准误	t 统计量	P 值
截距项	4.867e+04	3.003e+03	16.21	<2e-16 ***
year	-4.936e+01	3.004e	-16.43	<2e-16 ***
$year^2$	1.251e-02	7.514e-04	16.66	<2e-16 ***
R-square:0.9954		RSE:0.5311	p-value:<2.2e-16	

从表 5-1 中可以看出，全球海洋热量存储的二次模型是非常显著的。在过去 30 多年中，全球海洋总热量以二次函数的形式稳定上涨。其中模型的一次项、二次项系数显著，模型的调整后 R-square 为 0.9954，说明以时间为自变量的二次函数模型能解释 99.5%的海洋总热量的增长。

因此，在将海洋总热量的年度数据转化为月度数据的时候，本文中采用了二次函数插值，从而得到近似的海洋总热量月度数据。

5.3 模型的介绍

这一部分将介绍本文提出的 XCSM 模型的构建过程，以及 XCMM 模型基于的三个模型，分别是 XGBoost、协整模型和 SV。

5.3.1 XGBoost 简介

XGBoost 算法由 Chen and Guestrin（2016）提出，现在该算法已成为众多数据科学竞赛（如 Kaggle）中的常被使用的算法。该算法与 GBDT 同为 Boosting 类算法，但是在 GBDT 基础上有许多改进，例如：可以实现并行运算；可实现内存外计算；比其他 Boosting 算法快很多等。

其中，XGBoost 与 GBDT 最大的区别在于 XGBoost 的每一步迭代时，在目标函数里加入了正则项，可以避免过拟合。XGBoost 的优化目标为式（5-1）。其中“正则项”为式（5-2），符号说明见表 5-1。

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (5\text{-}1)$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (5\text{-}2)$$

表 5-1、XGBoost 公式符号说明

符号	说明
f_t	第t步迭代时需要拟合的树
T	这棵树包含T个叶节点
w_j	第j个叶节点上的估计值
γ	调优参数
λ	调优参数

5.3.2 协整模型简介

协整模型是对时间序列数据建立的一种模型，该模型不像 AR、MA、ARMA 等时间序列模型需要平稳的时间序列才能进行分析，协整模型处理的数据可以是两个或两个以上的非平稳时间序列，但是虽然每个时间序列都是非平稳的，但是它们的某种线性组合却能表现出平稳性，也就是说这些变量之间存在长期稳定关系，即协整关系。

5.3.3 随机波动率模型简介

随机波动率模型是一种时间序列模型。在时间序列中，风险常常以波动率的形式体现。波动率是用于衡量特定区间内时间序列变化程度的指标，反映了市场的变化情况。20 世纪 80 年代初，Engle 在时间序列领域中提出了全新模型——ARCH 模型，以通货膨胀指数作为研究对象进行了实证研究。 80 年代中期，Bollserlev 拓展了 Engle 的模型，给出了适用面更广、根据普遍性的 GARCH 模型，而如今 GARCH 模型已经成为了时间序列分析主流模型之一。 同年，Taylor 提出了与 GARCH 模型有很多相似，但理论基础完全不同的随机波动率 SV 模型，该模型能够在另一个方面很好地反映时间序列的条件异方差性，即随机波动项的变化的方差是随机变量。对于随机波动率模型，常常采用 MCMC 方法，利用蒙特卡洛模拟的思想，进行马尔科夫链上的建模，同时利用贝叶斯方法估计参数后验分布的优势，通过灵活的抽样方法解决 SV 模型之前不能被很好处理的的参数估计上的问题。

5.3.4 XCSM 模型的构建

XCSM 模型的流程图如图 5-1.构建的思路在于希望能综合 XGBoost 的预测准确性与协整模型的可解释性，可以看出，XCSM 模型有以下几个方面的优点：

- XCSM 中的 XGBoost 组件有为特征进行重要性排序的功能，有助于我们选取重要的影响气候变化的因素。
- XCSM 中的 XGBoost 组件有较强的预测能力。
- XCSM 中的协整模型部分考虑时间维度的变化，并且易于解释，有助于得出影响天气因素的结论。
- 综合 XGBoost 和协整模型以及 SV 的预测结果，使得预测结果更加稳健。

5.4 XCSM 的使用过程与结果分析

5.4.1 XCSM 自变量重要性分析

本部分搜集到的数据总量为 437 条数据，考虑到其数据量不大，所以训练集数据所占比例需要调高，这里按照 8：2 比例切分训练集和测试集。拆分方式是随机进行拆分，值得注意的是，虽然数据是存在时间顺序的，但是存在时间以及月份字段来存放时间信息，也就是即使是随机打乱样本数据的顺序，但是时间关系并没有损失。所以此处对数据进行随机拆分也不乏合理性。

在 XCSM 的训练过程中，我们需要为其选择合适的超参数，选择的依据是使得模型在测试集上的均方误差最小。对于 XGBoost 模型，一个重要的参数是学习率，搜索的方法是网格搜索，尝试在 0.1000，0.2125，0.3250，0.4375，0.5500，0.6625，0.7750，0.8875，1.0000 中寻找使得测试误差小的学习率，在学习率为 0.6625 时，测试集上的均方误差最小，为 0.01137757。

该算法也可以得到变量的重要性排序，在条形图 5-3 中展示。可以看出时间、二氧化碳平均浓度、ENSO 指数等对于气温预测有重要影响。将重要的变量尝试建立协整模型与 SV 模型。

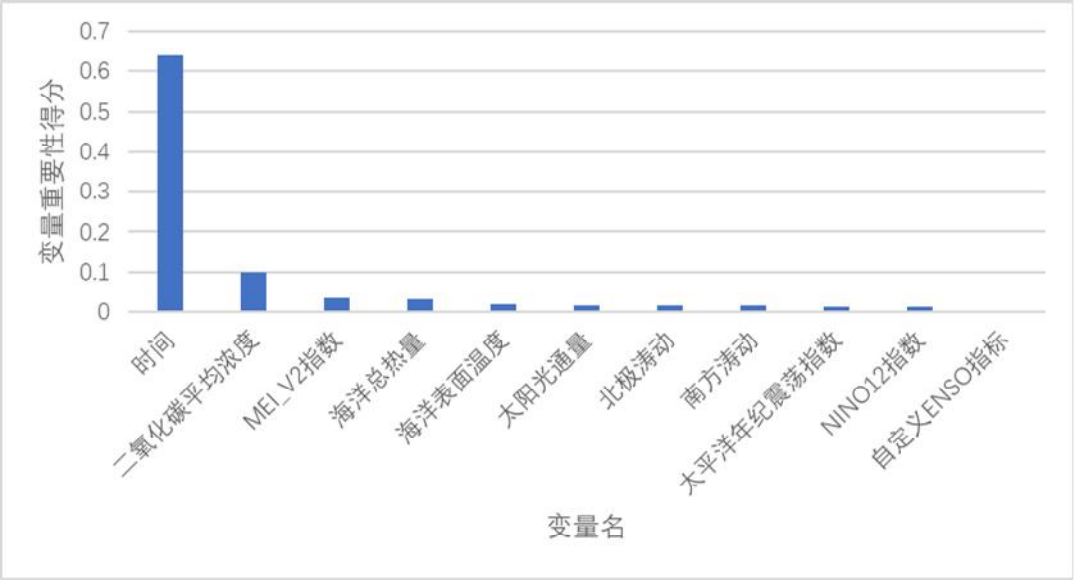


图 5-3 变量重要性排序

5.4.2 随机波动率模型

为了提高模型预测的稳健性，XCSM 借鉴 Bagging 的思想，将随机波动率模型与 XGBoost 和协整进行结合，从而有效提高模型预测能力。

在通过 XGBoost 对影响全球地表平均温度的变量进行重要性分析后，本文认为二氧化碳平均浓度、MEI V2 指数、全球海洋热量存储以及海平面温度四个变量对全球地表平均温度有主要的影响。在图 5-3 中也可看出，其他变量影响程度相对较弱。

因此，在筛选重要变量后，将重要变量通过随机波动率模型进行预测是 XCSM 的下一个流程。本节中以 MEI V2 指数为例阐述随机波动率模型的建模效果与预测能力。

5.4.2.1 时间序列确定性分析与短记忆序列建模

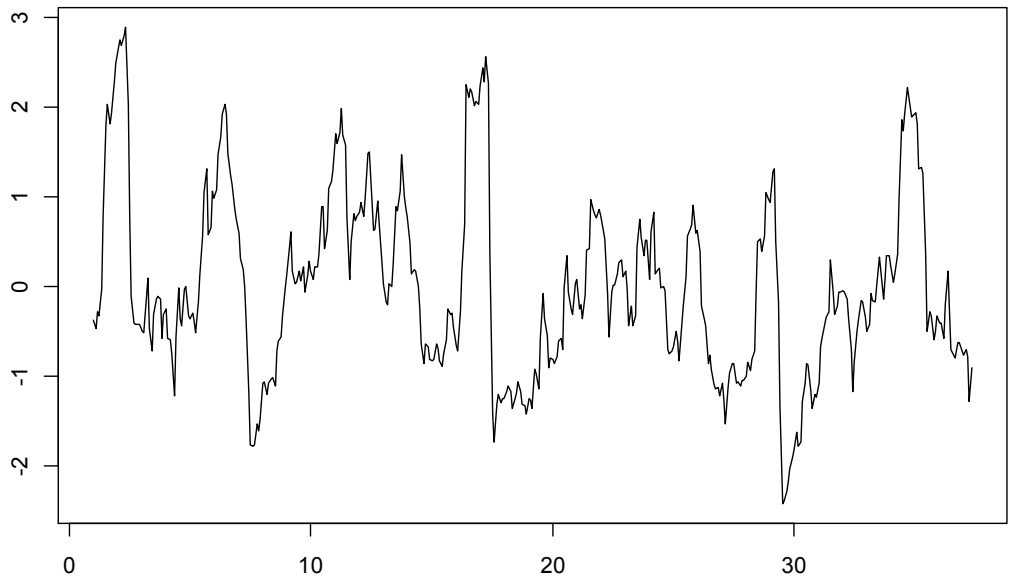


图 5-4 MEI V2 指数的时序图

从图 5-4 中不难直观看出 MEI 指数具有周期性特征，不是平稳序列，因此需要进行确

定性分析。这里使用 Holt-Winters 方法剔除时间序列的周期性与趋势性，分离随机性序列。Holt-Winters 算法是三次指数平滑法。相较于比一次、二次平滑指数模型，三次指数平滑法更多考虑了季节性的因素，适用于季节变化的时间序列。拟合优劣程度与历史数据变化是否稳定有关，如果历史数据变化存在一定规律该算法往往能够捕捉到这一规律。

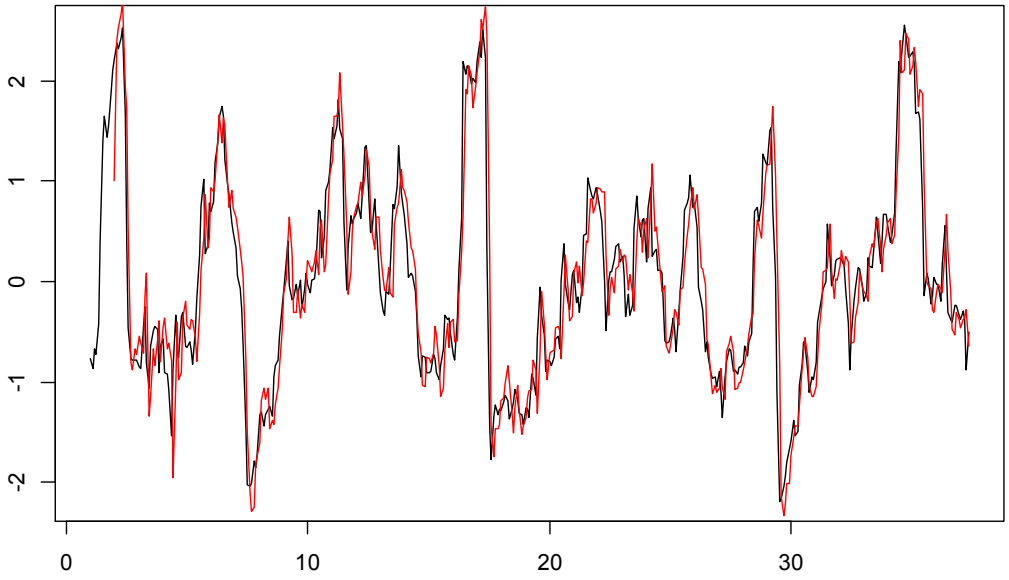


图 5-5 MEI V2 的 Holt-Winters 拟合图

图 5-5 中，黑色序列为剔除趋势性后的 MEI 序列，红色序列为 Holt-Winters 的确定性分析对其进行拟合的结果。可以清楚看到，在拟合方面，Holt-Winters 能很好地捕捉到 MEI 序列的趋势性和周期性特征。

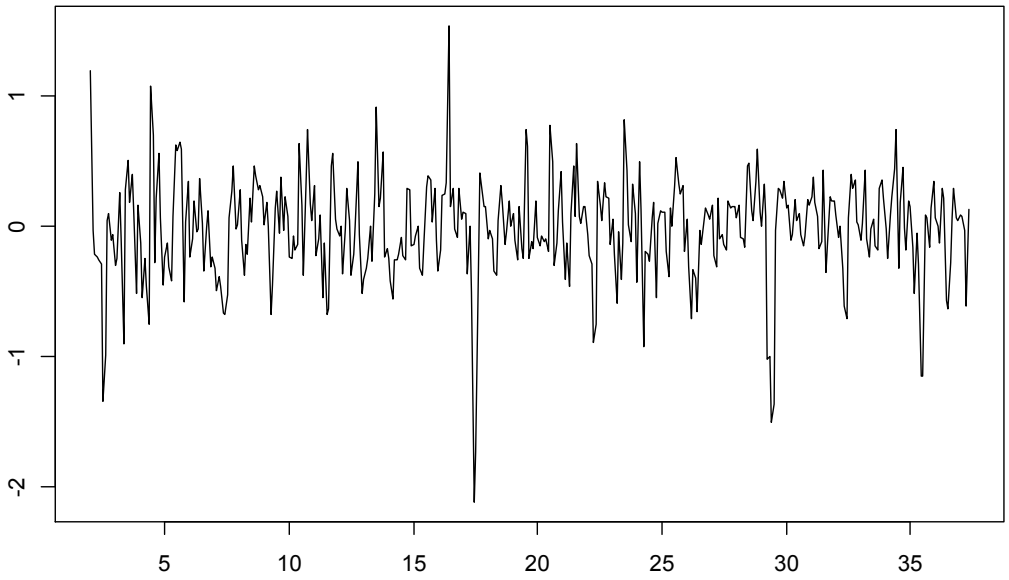


图 5-6 MEI V2 指数的随机性因素时序图

在对 MEI 指数进行了确定性分析后，需要判别其剩余的随机性序列是否是白噪声、是否是平稳的时间序列。对序列是否是白噪声序列的检验可以使用 Box-Pierce 检验。而单

位根检验常常使用 Augmented DF 检验与 PP 检验，平稳性检验常使用 KPSS 检验。

如表 所示，Box-Pierce 统计量 p 值小于 0.05，拒绝原假设，认为随机性序列不是白噪声。ADF 统计量与 PP 统计量 p 值小于 0.05，拒绝单位根原假设，认为随机性序列平稳；KPSS 统计量 p 值大于 0.1，接受原假设，接受随机性序列平稳。

表 5-2　MEI V2 指数纯随机检验及平稳性检验统计量表

统计检验	检验统计量	P 值
Box-Pierce test	X-squared=74.809	<2.2e-16
ADF test	Dickey-Fuller=-8.664	0.01<0.05
Phillips-Perron test	Dickey-Fuller z(alpha)=-234.18	0.01<0.05
KPSS test	KPSS Level=0.0217	0.1>0.05

对于非白噪声平稳时间序列，一般首先使用 ARMA 簇模型进行平稳时间序列建模。如 ACF 和 PACF 图所示，随机性序列平稳，且非白噪声。对 ARMA 模型进行定阶，定为 ARIMA(0,0,1)(2,0,0)[12]模型。

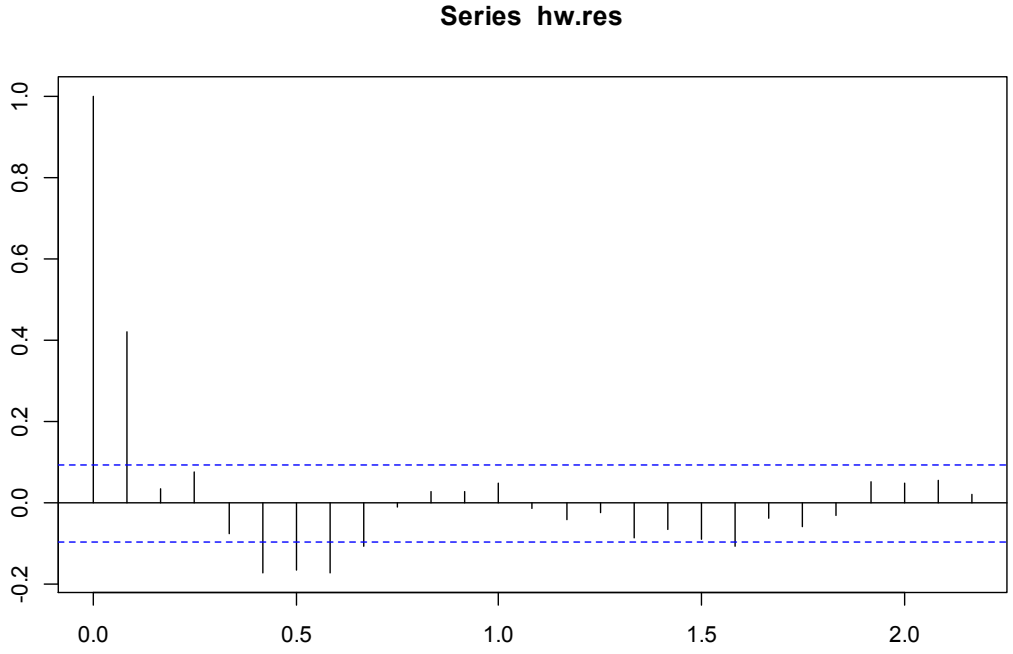


图 5-7 MEI V2 随机性序列 ACF 图

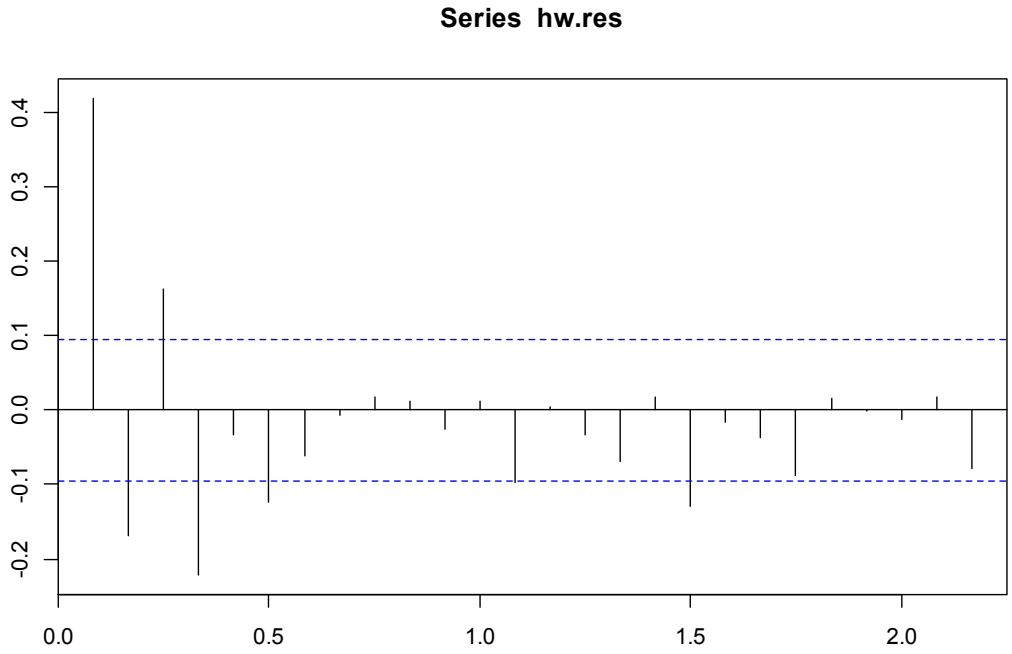


图 5-8 MEI V2 随机性序列 PACF 图

表 5-3 MEI V2 随机性序列 ARMR 族定阶统计量表

ARIMA(0,0,1)(2,0,0)[12]			
变量名	Ma1	Sar1	Sar2
点估计	0.6258	0.0686	-0.0134
标准误	0.0397	0.500	0.0514
AIC=287.26 BIC=287.36 ME=-0.0136 RMSE=0.336 MAE=0.251 MASE=0.6748			

5.4.2.2 厚尾随机波动率模型建模

在 ARMA 模型定阶后，对残差序列进行了 Ljung-Box 检验。如图 5-9 所示，p-value 在 Lag<5 时明显通过了检验。随后对残差序列进行随机波动率模型建模。

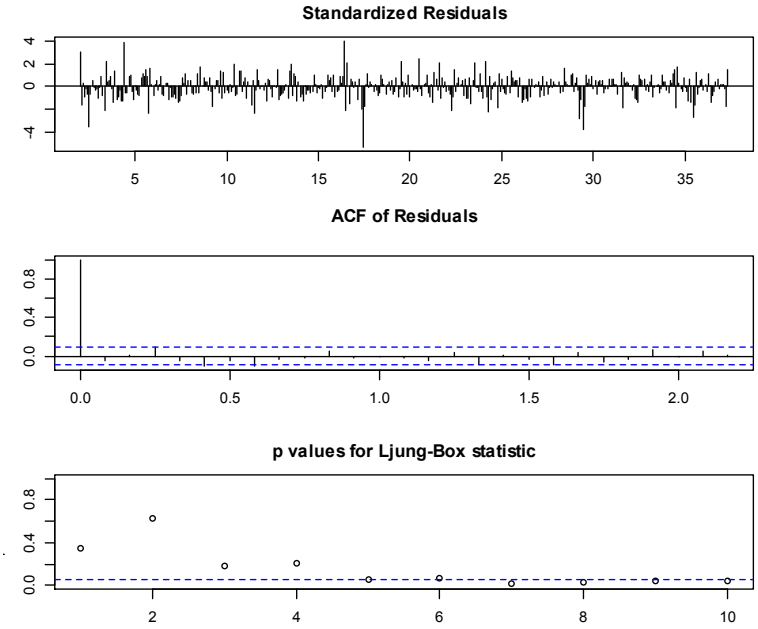


图 5-9 残差检验图

首先对 MEI 指数随机性序列进行正态性检验，判断使用厚尾随机波动率模型还是正态随机波动率模型。如 QQ 图所示，MEI V2 指数随机性序列具有显著的厚尾效应，使用厚尾随机波动率模型进行建模更为合适。

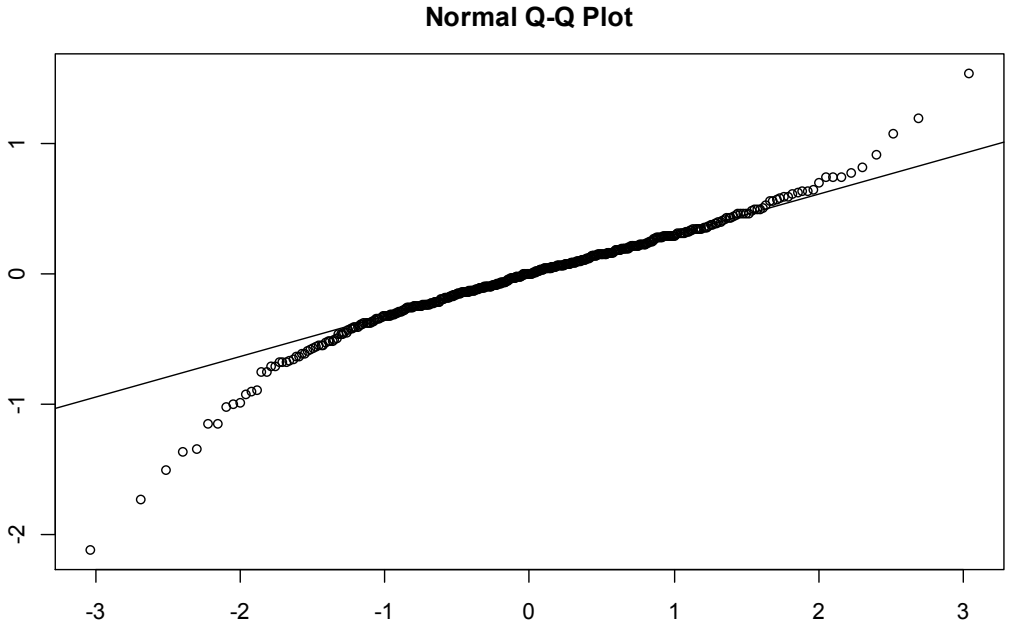


图 5-10 MEI V2 指数随机性序列 Q-Q 图

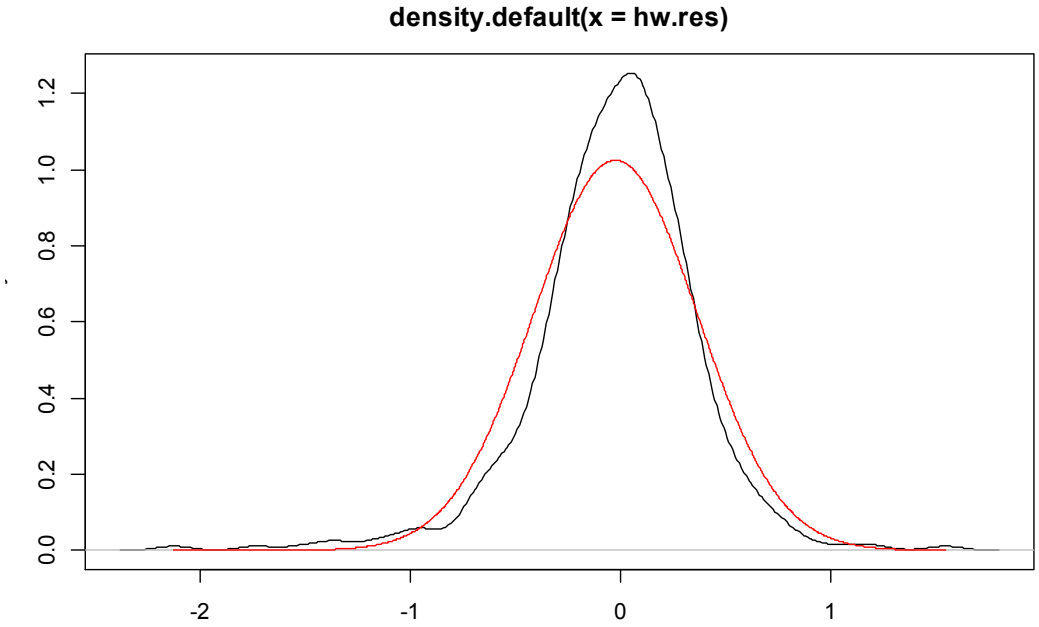


图 5-11 MEI V2 指数随机性序列概率密度分布图

SV 模型参数的先验密度参照 Kim, Shephard, Eric 等的研究论文选取。所建立 SV 模型架构如下图所示。

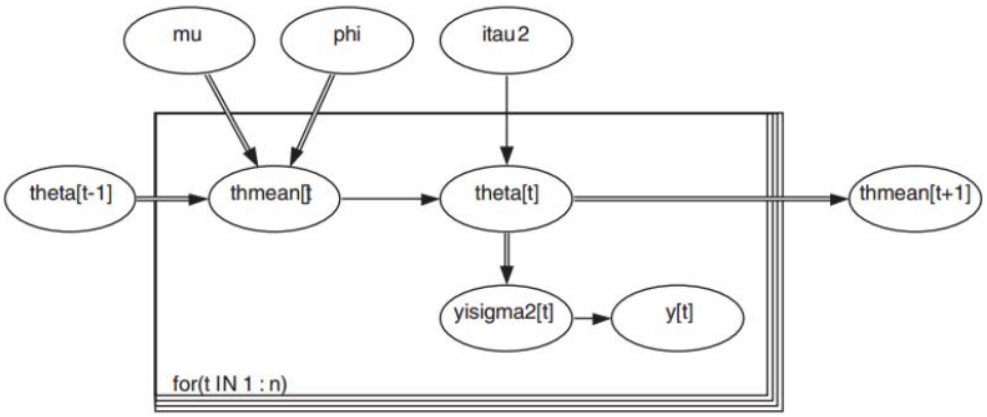


图 5-12 随机波动率模型结构

由于初始参数过多， OpenBUGS 无法自动生成，因此手动载入参数。在模型基础上用 MCMC 方法产生 30000 个数据，并燃烧前 20000 个。

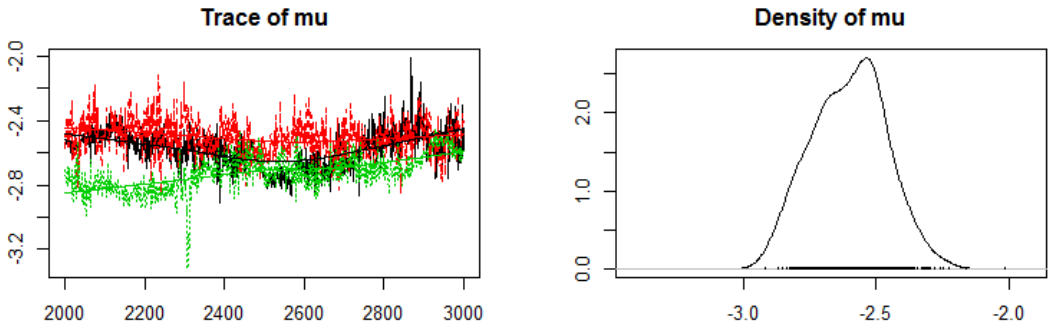


图 5-13 SV-t 模型参数绞合链图

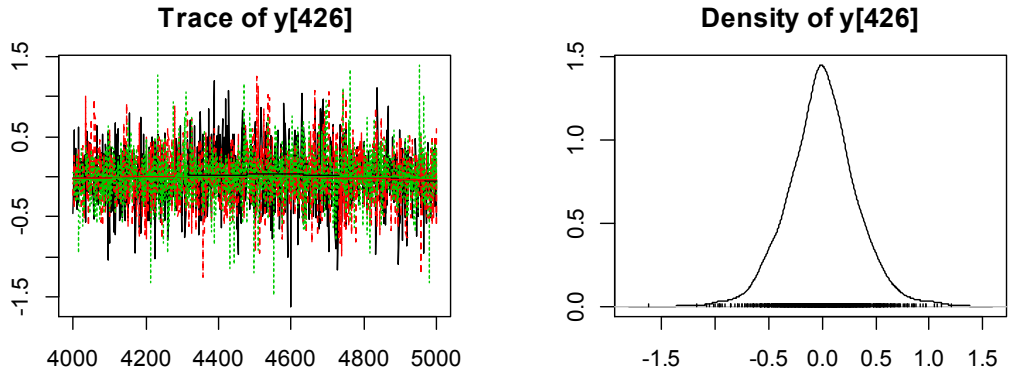


图 5-14 SV-t 模型序列预测图

首先看 MC_error。由于模型参数多(迭代的 theta 有 1950 个),因此模型 MC_error 较高依然可以接受。可以看到， μ, ϕ, τ, ˆ y 的 MC_error 均小于 sd 的十分之一，虽然部分参数 MC_error 大于 sd 的 0.05。我们认为 MCMC 算法得到的模拟的时间序列链不是发散的。同时观察各个样本和参数估计的方差，可以看到参数估计的波动是较小的，因此得出结论参数估计是合理的。

同时，由于 ϕ 接近于 1，可见市场的波动持久性特征很明显。

对于密度图，主要关注 y 预测值的分布。图中可见其服从一个对称分布。经过后文中的检验可以得到其分布厚尾尖峰的结论。

从四个参数的分位数图中可以看出，四个参数的极值都非常稳定，没有太多过大或者过小的异常值产生。首先观察 τ ,y。其轨迹图的形状一直在常数附近震动徘徊，可能具有一些纯随机序列的特征。此外，对于 μ,ϕ，这两个参数都在同一节点有明显的断崖式变化，在这个点之外的地方均非常稳定。我们认为是在这个点处，MCMC 产生了一个非常偏离均值的离群值，使得参数发生改变，参数的传递使得两个参数一同发生了变化。值得注意的是，在该点处 τ 和 y 均没有产生巨大改变。由于模型关心的是 y 参数，可以认为模型是非常稳健的。

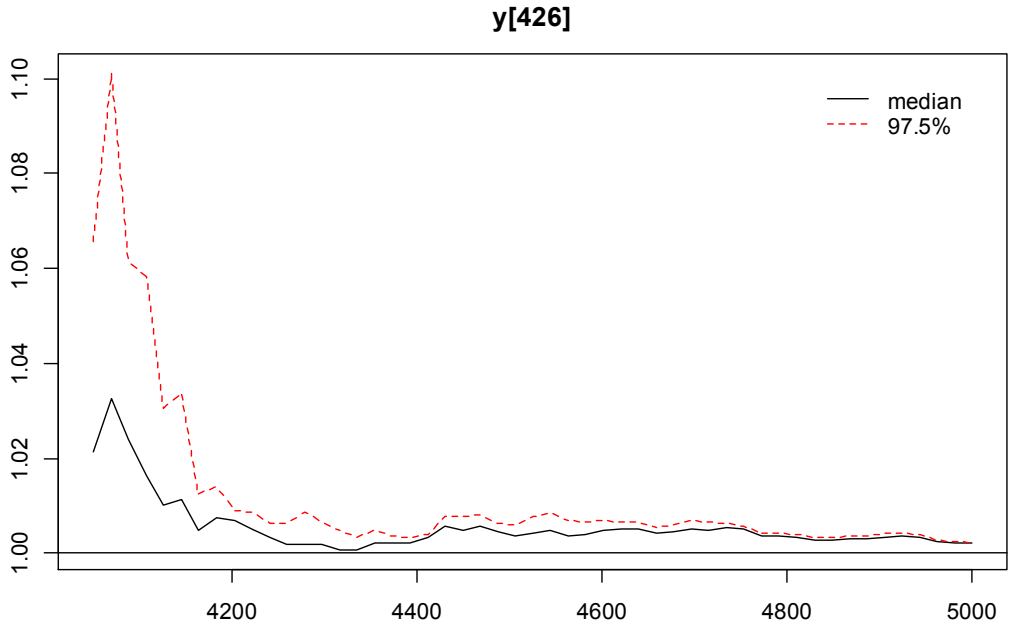


图 5-15 厚尾随机波动率模型 Gelman 检验

如上图所示，随机波动率模型在 Gelmen 检验下相对稳健。

最后，对 MEI V2 指数进行基于 ARMA-SV 模型的预测。预测结果如下图所示。其中橙色折线为 1982 年 1 月至 2018 年 5 月的 MEI V2 指数历史数据，蓝色直线为对未来 25 年 MEI V2 指数的预测。由于样本外预测虽能通过随机波动率进行波动率的估计与序列的预测，但是预测的序列在震荡上，与其他时间序列短记忆预测模型一样，相较于原始序列将处于较低的状态。

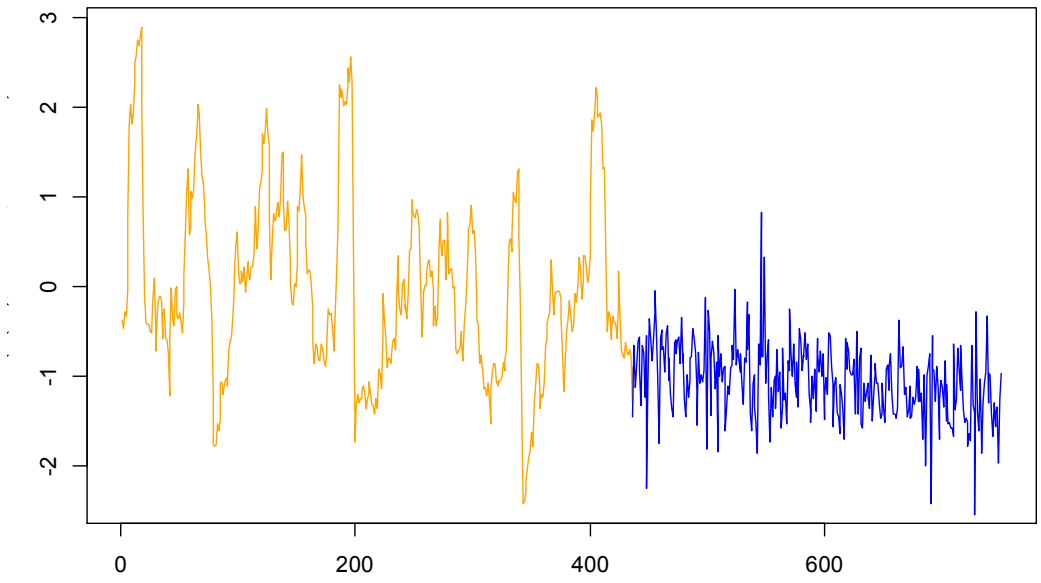


图 5-16 MEI V2 序列的 SV 模型预测

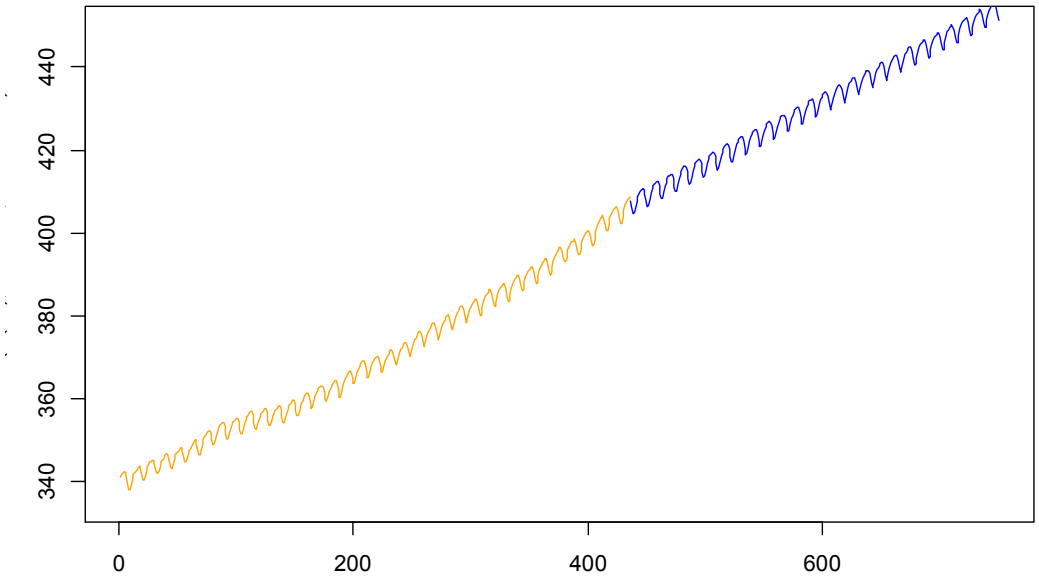


图 5-17 CO2 浓度的 SV 模型预测

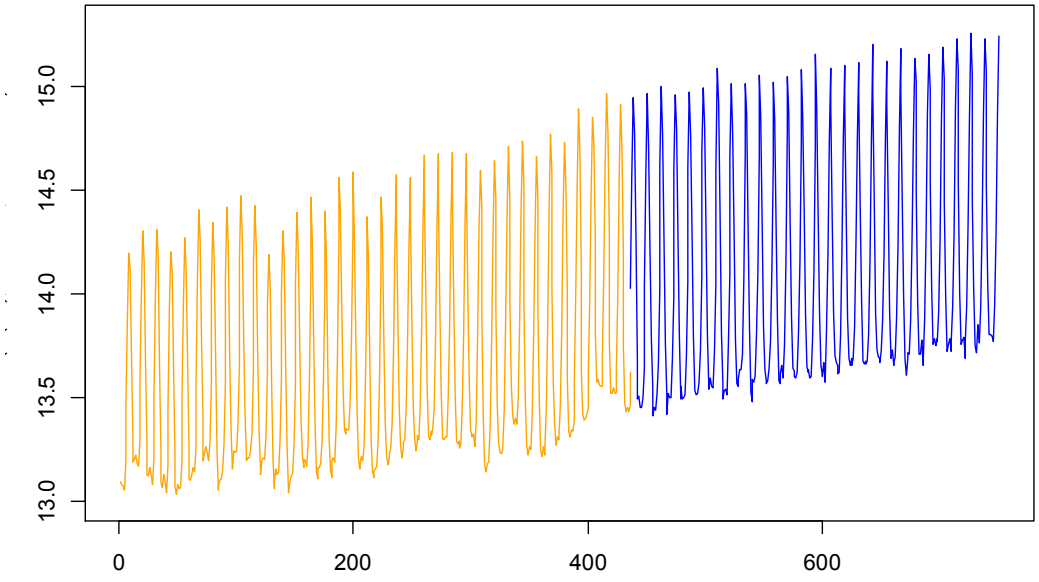


图 5-18 全球海洋表面温度的 SV 模型预测

如图 5-18 所示，使用 SV 模型对 CO2 浓度与全球海洋表面温度进行了预测，效果良好。

5.4.3 XGBoost-SV 模型预测

将随机波动率模型与 XGBoost 模型进行结合，可得到如下图预测。

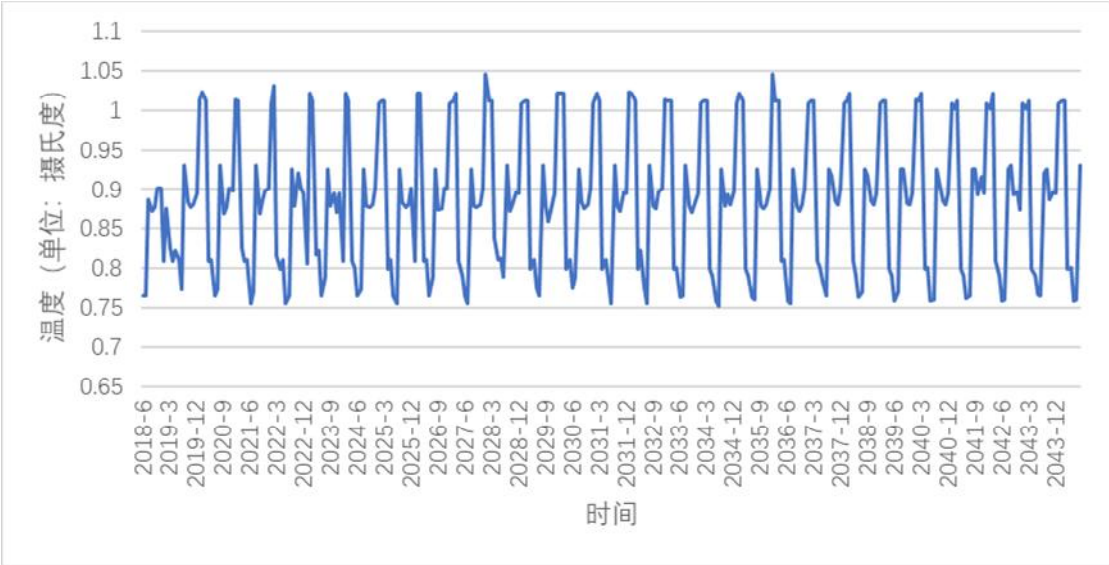


图 5-19 SV-XGBoost 地表平均温度 25 年预测图

5.4.4 协整模型

如前文中所述，多个时间序列可做协整分析的前提条件是多个序列同阶单整。因此需要对模型中的全球地表平均温度因变量和四个自变量进行同阶单整的检验。这里以对地表平均温度的检验为例。

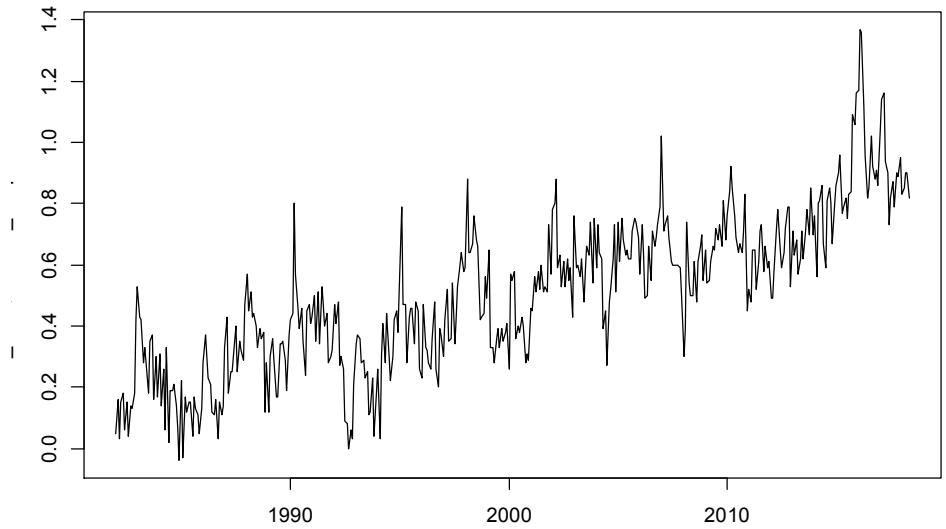


图 5-20 全球地表平均温度时序图

从全球地表平均温度时序图中不难看出，过去 30 多年内地表平均温度具有明显的趋势性。经过 KPSS 平稳性检验，p 值较小，拒绝原假设，认为序列不平稳。

表 5-4 GMST 平稳性检验统计量

检验名	原假设	检验统计量	P 值
KPSS test	序列平稳	KPSS Level=6.829	0.01<0.05

经过一阶差分后，重新检验序列平稳性。通过 ADF, PP ,KPSS 检验后可得，一阶差分

后的 GMST 序列平稳。因此 GMST 序列一阶单整。

表 5-5 GMST 一阶差分序列平稳性检验统计量

统计检验	检验统计量	P 值
Box-Pierce test	X-squared=62.932	=2.109e-15<0.05
ADF test	Dickey-Fuller=-9.898	0.01<0.05
Phillips-Perron test	Dickey-Fuller z(alpha)=-526.16	0.01<0.05
KPSS test	KPSS Level=0.00984	0.1>0.05

同样，将 CO2 浓度，MEI V2 指数，海洋表面温度及海洋热量存储进行相同的检验，可得四个自变量均为一阶单整。因此，所有因变量及自变量均同阶单整，可进行协整分析。

表 5-6 GMST 协整模型显著性检验

	系数估计值	标准误	t 统计量	P 值
截距项	-8.707	1.327	-6.56	<<0.05 ***
CO2_average	0.024	0.0034	7.02	<<0.05 ***
MEI_V2	0.052	0.0062	8.33	<<0.05 ***
ocean_heat	-0.0318	0.0087	-3.63	0.0003 ***
sea_surface_temperature	0.0453	0.0160	2.82	0.0049 **
R-square:0.7479		RSE:0.1271	p-value:<2.2e-16	

在进行协整建模后，对模型进行检验。可以看到四个因变量的系数远小于 0.05，系数显著。同时模型的 R2 达到 0.75，调整后的 R2 也达到了 0.746，说明模型显著地解释了 75%左右全球地表平均温度的变动情况。

表 5-7 GMST 协整模型检验

统计检验	检验统计量	P 值
ADF test	Dickey-Fuller=-5.21	0.01<0.05
Phillips-Perron test	Dickey-Fuller z(alpha)=-225.26	0.01<0.05
KPSS test	KPSS Level=0.108	0.1>0.05
Jarque Bera test	X-squared=8.503	0.014<0.05

最后检验协整模型建立时是否存在伪回归现象。使用 DubinWatson 统计量与 R2 统计量进行比较验证是否有伪回归存在。DW 统计量值为 0.96，高于 R2 统计量。因此不存在伪回归现象。

表 5-8 协整模型伪回归检验

Durbin-Watson test		
DW 统计量	P 值	R-squared
0.96	<2.2e-16	0.7479

协整模型相较于 XGBoost 的预测模型而言，解释性与鲁棒性都更强。由于协整模型建立了 GMST 因变量与四个自变量之间的线性关系，因此各变量之间关系非常明确。由于不

存在二次项与交互项，因此当其他变量确定的情况下，从二氧化碳浓度到海洋热量总存储对气候变化的影响都可以清晰地分析出来，且各个变量对气候变化的影响都是互不干扰的。这样的模型更具有代表性与稳健性。

如下图所示，通过协整模型与随机波动率模型的预测，本节建立了 Cointegration-SV 预测模型。

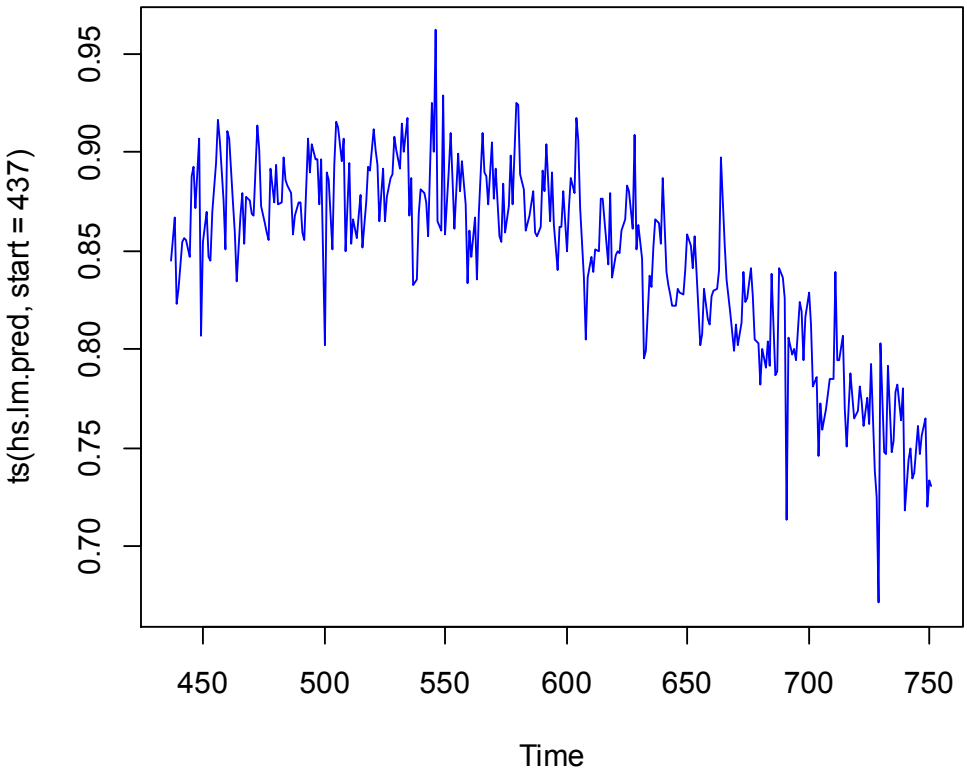


图 5-21 全球地表平均温度 Cointegration-SV 预测

5.4.5 XCSM 模型预测

有了可解释性强的协整模型与预测性强的 XGBoost 模型后，需要将其进行结合，生成最终的预测。本文认为，两个模型在对 XCSM 模型最终预测的贡献上应该是等价的，因而 XCSM 模型中两者的成分与所表述的信息应当是相同的。

通常研究中，往往将数据的方差、标准差作为其信息量的表征。本文以标准差作权重，权衡 XCSM 模型的成分构成。

XCSM 预测结果如下图所示。其中黑色折线为 1982 年 1 月至 2018 年 5 月 GMST 历史数据，蓝色折线为 2018 年 6 月一直到至今 25 年后的将来的预测。

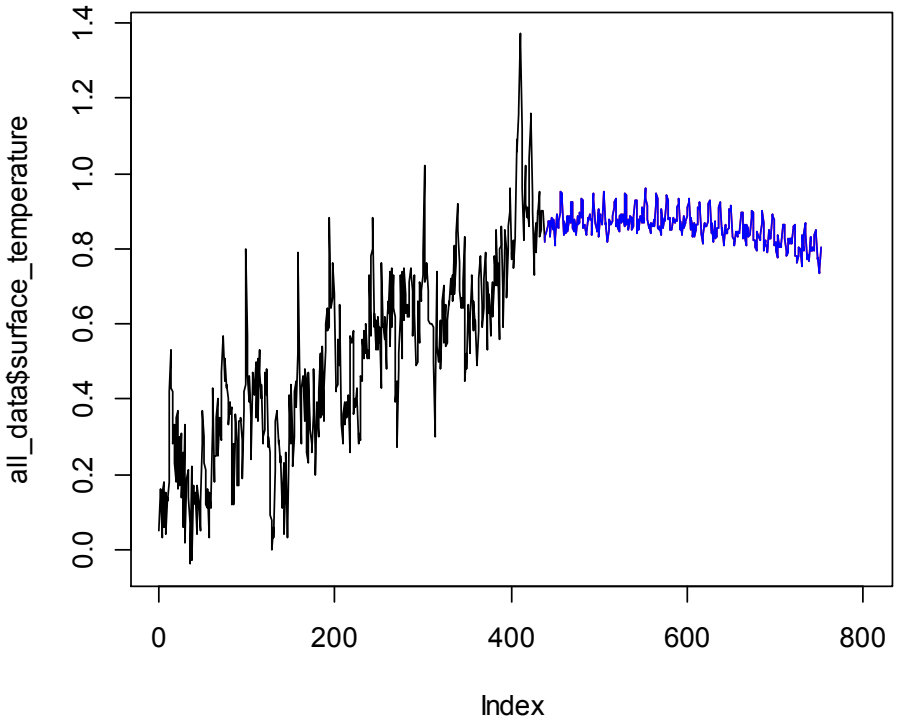


图 5-22 XCSM 模型全球地表温度预测

由于气候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，因此需要对全球地表温度进行进一步处理，以得到气候变化的预测。参考气候学文献，本文对 GMST 取季度为单位的指数平滑，得到未来 25 年气候变化的预测。

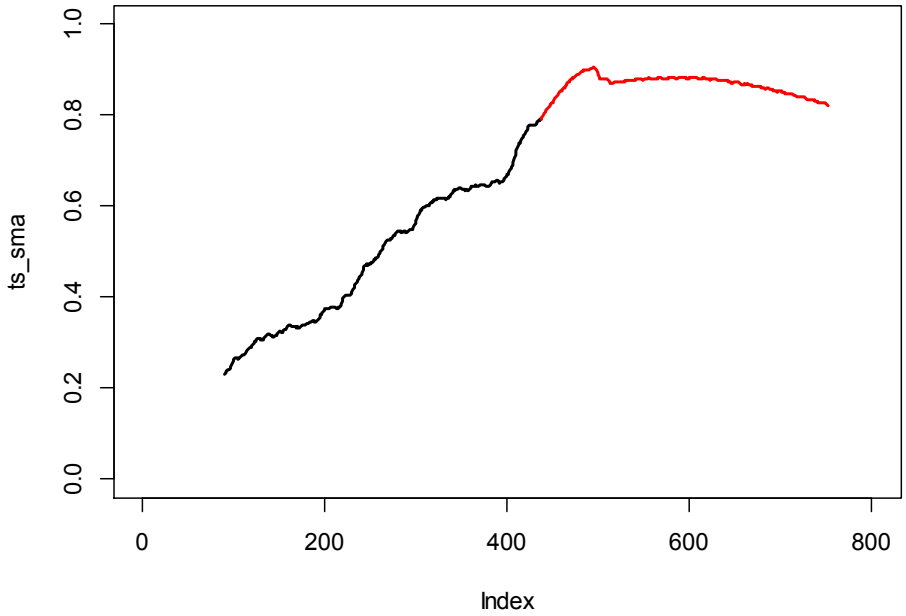


图 5-23 XCSM 模型气候变化预测

如图 5-23 所示，全球变暖确实于 2010 年左右进入停滞阶段。在 2015~2023 年左右，全球变暖将持续；根据 XCSM 模型预测，在 2023 年后，全球变暖将陷入较长的停滞期，全球气温将保持平稳，长期有下降趋势。

6 问题三 局地极寒模型

6.1 问题的描述与分析

2019 年 1 月北美遭遇了极寒天气，美国及加拿大多个州因自然灾害损失惨重。极寒天气的出现看上去与全球变暖之间是矛盾的，因为对气候学不甚了解的人会认为在全球升温的大环境下出现大范围的极寒天气是不合理的，因而极寒天气的出现可以直接否定全球变暖的假说。这种观点是错误的，因为极寒天气只是局地现象，不能以点概面，以北美洲中部的极端天气概括全球范围内气候的整体变化。

局地极寒的产生原因是相当复杂的，学术界至今没有达成完全统一的共识。一般认为，北美洲的极寒天气的成因很大程度上归咎于北极涛动和极地涡旋。其中不少学者认为北极涛动一定程度上引发了极地涡旋的产生。另一方面，不少学者认为北极涛动与厄尔尼诺南方涛动之间有密切的联系。而如前章所述，ENSO 耦合了赤道太平洋的海-气系统，将大气热量与海水热量紧密结合在一起。通过这个气-海-气链条，温室气体排放、全球变暖与局地极寒天气之间有密切的联系。

要理解极寒天气的产生以及极寒天气与全球变暖之间的矛盾，最直观的做法是建立极寒天气与全球变暖等自然气候要素之间的相互关系，从他们之间的关系中反映出真正的全球气候整体模型。其中的难点在于如何通过全球的整体数据，建立局地极寒的模型。这点是具有难度的，原因在于在不获得局地数据的约束下，通过宏观因素来对局地地区进行分析是困难的，而如果获得局地数据，那么全球变暖与局地极寒之间的矛盾将被弱化，所解决的也不再是具体极寒与全球变暖之间的问题了。

6.2 数据的准备

6.2.1 变量选取

相较于第二章中的气候变化预测模型而言，局地极寒模型的构成更为复杂。原因在于气候变化预测模型是全球性的模型，能从整体上影响气候变化的因素较少，影响力较小的因素在全球范围内体现不出其作用。同时，为了宏观模型的稳健性与可解释性，大量不甚重要的影响因素被剔除。另一方面，局地的气候变化瞬息万变。能对局地产生影响的气候要素众多，稍有遗漏模型解释性会大幅降低；同时，在不获得局地数据的约束下，通过宏观因素来对局地地区进行分析是困难的，而如果获得局地数据，那么全球变暖与局地极寒之间的矛盾将被弱化，研究方向就会发生转变。

为表征局地极寒，本文采用加拿大 Nunavut 省 30 个基站的测量结果进行交叉互补，并取平均，作为努省月平均气温。将年度月最低温作为努省寒冷指标。在自变量方面，主要分为六个部分。第一部分是厄尔尼诺指标及南方涛动指标，分为 MEI V2, NIÑO1&2, NIÑO3, NIÑO4, ENSO SST, Southern Oscillation 六个自变量；第二部分是太平洋年代际震荡；第三部分是北极涛动；第四部分是 CO2 浓度；第五部分是地表平均温度以及全球海洋热量存储指数；第六部分是太阳光通量。

6.2.2 数据来源

MEI V2, NIÑO1&2, NIÑO3, NIÑO4, Southern Oscillation, 太平洋年代际震荡, 北极涛动，CO2 浓度，地表平均温度，全球海洋热量存储指数以及太阳光通量主要来自于 NOAA、NASA、NCAR、世界银行和 Berkeley 发布的公开数据集。此外，ENSO SST 数据来源于 ESRL 的时空 SST 数据，通过数据预处理获得。

6.2.3 数据预处理

Shineng Hu 在 The extreme El Niño of 2015-2016 and the end of global warming hiatus 中采用 160° E~90° W, 5° S~5° N 的太平洋赤道中部表面温度作为 ENSO 的特征。本文中引用 Hu 的做法，通过对 ESRL 提供的 SST 数据进行处理计算得到 ENSO SST 指标。

6.3 模型的介绍

6.3.1 研究对象选取

本章中，局地极寒天气模型主要研究对象为加拿大 Nunavut 省。

对于极寒现象的研究，采集并使用的数据为加拿大努纳武特行政区的数据样本。选择努纳武特行政区是由于北美洲受到北极涛动的影响较大。其次努纳武特行政区位于加拿大东部的北极地区，而加拿大西部有科迪勒拉山系作为天然屏障，北极的冰空气团往南移动的过程中必先经过努纳武特行政区。再者加拿大大部分区域为温带大陆性气候，西部沿海区域为温带海洋性气候，邻近科迪勒拉山系地区为高原山地气候，而东北地区的努纳武特为寒带气候，利用位于寒带气候区的努纳武特行政区数据样本可以得到极寒现象时加拿大的最低温度样本。综合以上因素，因而选择努纳武特行政区的数据样本作为分析局地地区极寒气候的对象。

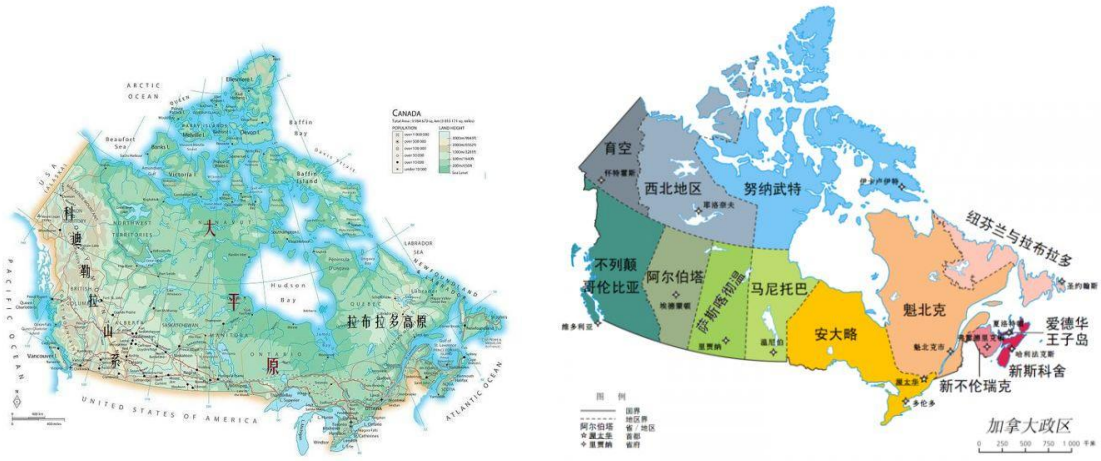


图 6-1 加拿大地形图与行政区划分

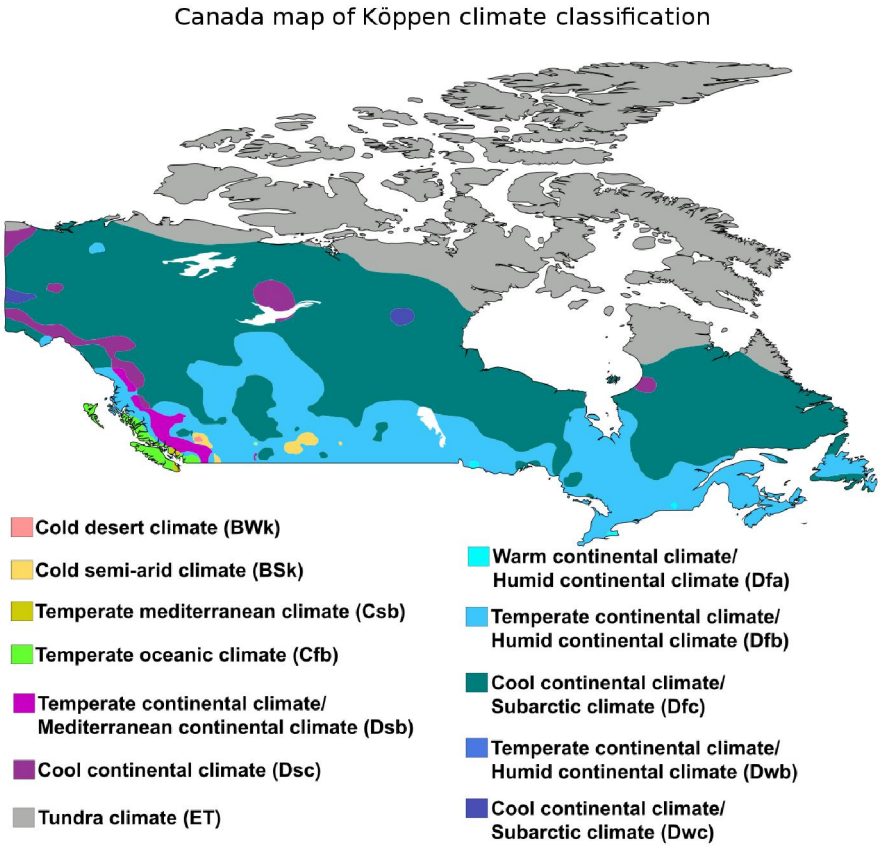


图 6-2 加拿大气候类型分布图

6.3.2 因子分析法简介

因子分析法起源于 20 世纪早期 K.Pearson，C.Spearman 和其他一些学者为定义和测定智力所做的努力，是主成分分析的推广和发展。它将具有错综复杂关系的变量总和为数量较少的几个因子，以再现原始变量和因子之间的相互关系，同时根据不同因子还可以对变量进行分类。

因子分析通过分析事件的内在关系，抓住主要矛盾，找出主要因素，使多变量的复杂问题变得易于研究和分析。同时它消除了指标间的相关问题和评价的主观性。因子分析法不仅可以给出排名顺序，还可以探究影响排名次序的因素，从而进一步改善努力的方向，这是其他综合评价方法所不具备的。

因子分析的基本思想是通过变量的相关系数矩阵或协方差矩阵内部结构的研究，找出能控制所有变量的少数几个随机变量去描述多个变量之间的相关关系。然后根据相关性大小把变量分组，使得同组内的变量之间相关性较高，不同组的变量之间相关性较低。每组变量代表一个基本结构，这个基本结构称为公共因子或主因子。

从一些错综复杂的变量中找出几个主因子，抓住这些主因子就可以帮助我们对复杂的变量关系进行分析和解释。

6.3.3 多元回归模型简介

在现实世界中存在大量这样的问题：两个或多个变量之间有一些联系，但没有确切到可以严格决定的程度。例如，人的身高 x 和体重 Y 有联系，一般表现为 X 大时，Y 也倾向于大，但由 X 并不能严格决定 Y。一种农作物的亩产量 Y 与其播种量 X1，施肥量 X2 有联

系，但 X1、X2 不能严格决定 Y。工业产品的质量指标 Y 与工艺参数和配方等有关系。但后者也不能严格决定 Y。在以上诸例以及相似的例子中，Y 通常称为因变量或预报变量，X1、X2 等称为自变量或预报因子。因变量与自变量的称呼借用于函数关系。回归分析着重在寻找变量间近似的函数关系。在回归分析中，因变量总是随机变量，对于自变量则情况较复杂：有随机的，如人的身高体重的那个例子；也有非随机的，农作物中的播种量与施肥量即是。

在大多数的实际问题中，影响因变量的因素不是一个而是多个，虽然自变量和因变量之间没有严格的、确定性的函数关系，但可以设法找出最能代表它们之间关系的数学表达形式。我们称这类问题为多元回归分析。

设变量 Y、xi （ i=1,2,…n)是相关变量，并在实际工作中获得了它们的 n 组观测值。多元回归分析是通过这 n 组观测值研究变量间相互关系的一种数理统计方法。它能够把隐藏在大规模原始数据群体中的重要信息提炼出来，把握住数据群体的主要特征。

6.4 因子分析的使用过程与结果分析

6.4.1 自变量相关系数分析

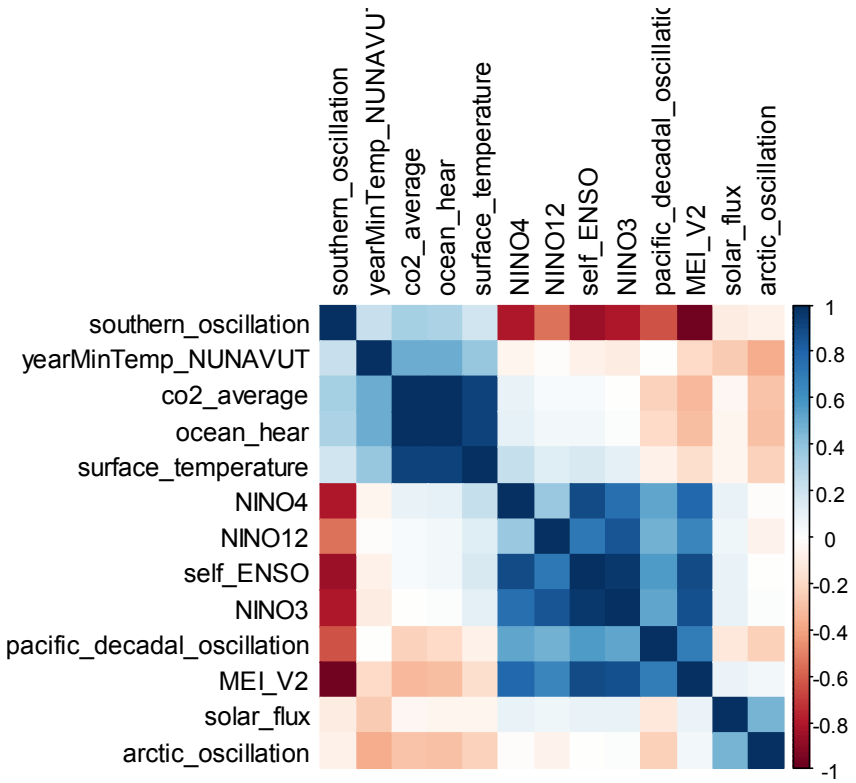


图 6-3 变量相关系数矩阵

由于模型的自变量是过量的，因此在做模型前需要先对各个自变量之间的相关性、共线性有初步的了解。通过绘制自变量相关系数矩阵，可以直观看出各个变量之间的相关性。

从上图第一列与中间部分深色矩阵中不难看出，Southern Oscillation 与 NIÑO12, NIÑO3, NIÑO4, MEI V2, ENSO SST 之间有非常强的相关性。由图中左上角深蓝色矩阵可以看出，CO2 浓度与海洋热量存储、海洋表面温度之间具有很强的相关性。

通过自相关矩阵，可以很直观地将不同变量进行聚类，其中 Southern Oscillation,NIÑO12, NIÑO3, NIÑO4, MEI V2 与 ENSO SST 可统一归为厄尔尼诺指数类，CO2 浓度，海洋热量存储与海洋表面温度可归为环境热量类。

6.4.2 主成分分析

在使用因子分析前，先使用主成分分析进行重要变量的选择以及自变量正交化，从而从直观上得到各个变量的区别。

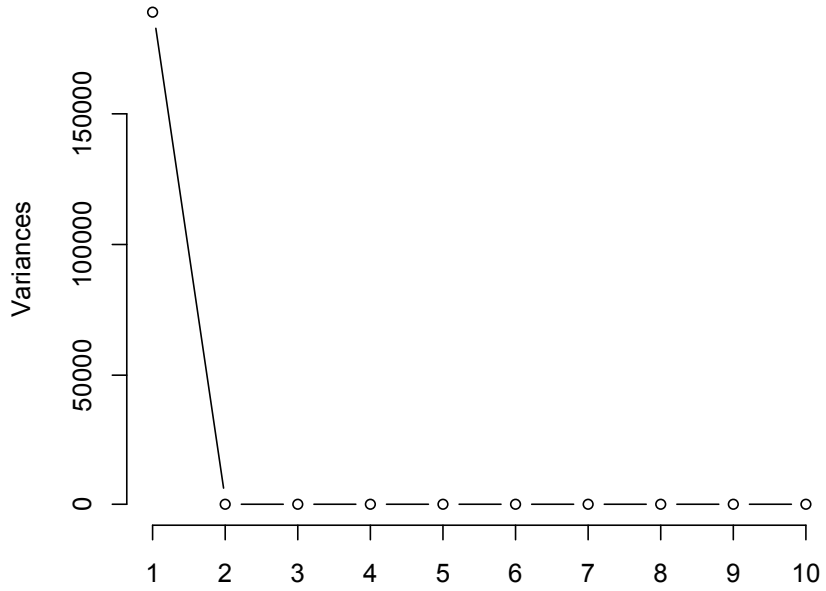


图 6-4 自变量主成分贡献程度图

从上图中可以看出第一个主成分对全部信息的贡献率是最大的，几乎涵盖了所有自变量信息的全部内容。

表 6-1 自变量主成分分解图

变量名	第一主成分	第二主成分	第三主成分
arctic_oscillation	-1.000	-0.014	0.000
MEI_V2	0.000	0.012	-0.394
NIÑO3	0.000	0.000	-0.354
NIÑO4	0.000	-0.002	-0.250
NIÑO12	0.000	-0.001	-0.402
pacific_decadal_oscillation	0.000	0.000	-0.335
solar_flux	0.000	-0.002	-0.022
southern_oscillation	0.000	-0.017	0.512
co2_average	0.013	-0.930	0.026
surface_temperature	0.000	-0.010	-0.019
ocean_heat	0.005	-0.366	-0.110
self_ENSO	0.000	-0.001	-0.327

上表分析了各个自变量对主成分的贡献程度。由于第一主成分贡献率极高，因此主要分析第一主成分。第一主成分中北极涛动的影响接近-1.000，几乎构成了全部第一主成分。此外，二氧化碳浓度及海洋热量对第一主成分也有微小贡献。第二主成分主要由二氧化碳浓度及海洋热量构成。而在此之中，地球表面平均温度对主成分的构建几乎不起作用。

6.4.3 因子模型结果分析

本节因子分解采用了以主成分分析为基础的，带 Varimax 旋转的因子分解方法。相较于极大似然估计法，主成分分析法在因子信息量提取中有更大的优势，但是在模型估计上有所欠缺。为了能更好找到贡献更大的因子，同时与预处理中的主成分分析进行比较，本节采用了 PCA 方法来估计分解。采用 Varimax 旋转方法，可使因子分解系数解释性更强。

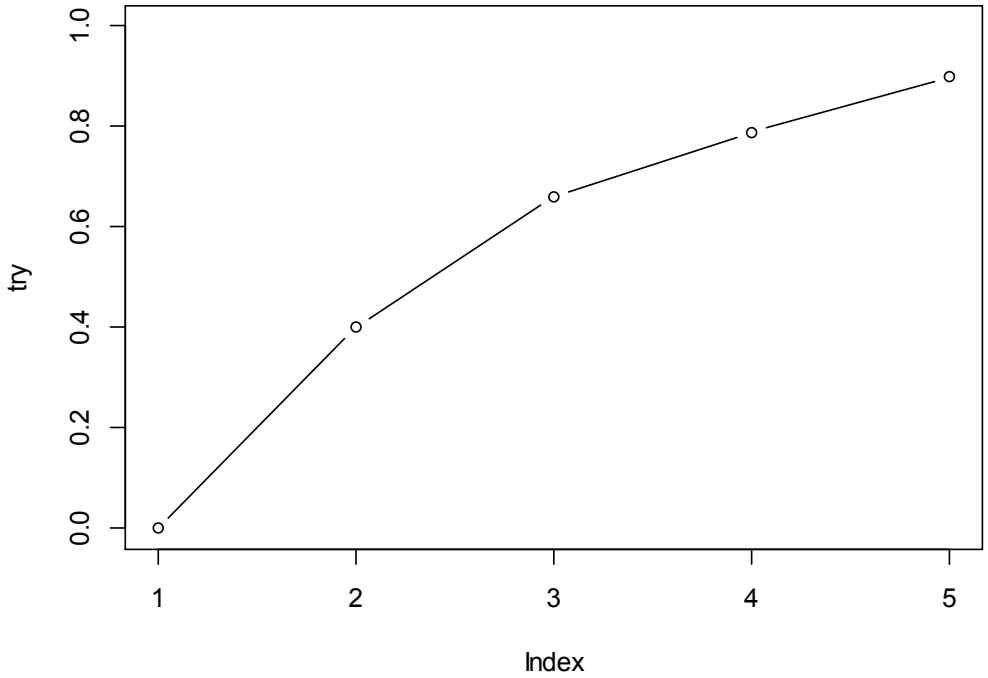


图 6-5 因子分析累计方差占比

四个因子的累计方差占比依次达到了 40%，66%，80%，79%。其中第一因子方差占比最多，占据了接近一半的信息量；第二因子方差、第三因子方差与第四因子方差分别达到了 26%，14%和 11%。与主成分分析模型相比，因子分析模型方差占比较为均衡，前四个因子作用都相对适中。同时，剩余信息占比比第四因子方差占比还小，说明因子分析较合理。学界一般以 75%~85%的累计方差占比作为分界线，界定因子分解是否全面。局地极寒模型的因子分解全面且稳健。

表 6-2 因子分析统计量表

	RC1	RC2	RC3	RC4	h2	u2	com
arctic_oscillation	-0.01	-0.24	0.83	-0.06	0.75	0.250	1.2
MEI_V2	0.91	-0.25	0.03	0.29	0.98	0.025	1.4
NINO3	0.81	0.08	0.08	0.54	0.96	0.039	1.8
NINO4	0.96	0.20	0.06	-0.10	0.97	0.026	1.1
NINO12	0.47	0.07	-0.01	0.88	0.99	0.010	1.5
pacific_decadal_oscillation	0.68	-0.20	-0.37	0.19	0.67	0.330	1.9
solar_flux	0.05	0.02	0.84	0.07	0.71	0.288	1.0
southern_oscillation	-0.93	0.25	-0.06	-0.14	0.94	0.056	1.2
co2_average	-0.10	0.99	-0.07	0.01	0.99	0.013	1.0
surface_temperature	0.06	0.96	-0.06	0.04	0.93	0.068	1.0
ocean_hear	-0.08	0.98	-0.09	0.03	0.98	0.021	1.0
self_ENSO	0.92	0.13	0.06	0.33	0.97	0.028	1.3

从表 6-2 中，可以直观感受到各因子的含义。Factor1 的得分主要来自于 MEI V2, NIÑO3, NIÑO4, Southern oscillation 与 ENSO SST，此外太平洋年代际震荡与 NIÑO1+2 也对其有

一定的贡献。NIÑO3，NIÑO4 与 ENSO SST 都是对太平洋热带中部进行分析所得到的数据。因此，将 Factor1 命名为太平洋热带中部厄尔尼诺系数，简称 ENSO 系数。Factor2 的得分主要来自于 CO2 浓度，全球地表平均气温与海洋热量存储。这三个指数的在因子 2 上的得分都接近+1，因此，将 Factor2 命名为热量系数。Factor3 的得分主要来自于北极涛动与太阳光通量。很多学者认为，太阳光通量与北极涛动之间有很强的相关性与气候学上的联系，日照强度改变对北极冰-水-气系统结构的影响是巨大的。因此，将 Factor3 定义为北极涛动指数。

Factor4 的得分主要来自于 NIÑO1+2，其次来自于 NIÑO3。NIÑO1,2,3 所研究的对象是有区别的。Niño1+2 的地区是 NiñoSST 地区中最小，最东部的地区，与南美洲沿海地区相对应，该指数往往具有 NiñoSST 指数的最大方差。Trenberth and Stepaniak（2001）定义了 TNI 指数，即 Niño1 + 2 和 Niño4 地区之间归一化 SST 异常的差异。TNI 测量了赤道中太平洋和东部赤道太平洋之间海表温度异常的梯度。当 SST 梯度特别大时（例如，由于 Niño4 区为正异常，而 Niño1 + 2 区为负异常），一些研究人员将该事件归类为“太平洋中部厄尔尼诺中心”。因此，将 Factor4 定义为 NIÑO 极东指数。

6.4.4 多元非线性回归模型

基于 Trenberth and Stepaniak 的研究，有理由认为各个因子对局地极寒天气的影响不是独立的，很可能有交互效应存在。因此，本文建立了多元非线性回归方程，以研究各个因子对局地极寒天气的影响程度。

通过 AIC 准则，建立了如下多元非线性回归模型。

$$\begin{aligned}\text{Freezing} = {} & \text{factor1} + \text{factor2} + \text{factor3} + \text{factor4} + \text{factor1:factor2} + \text{factor1} \times \text{factor3} \\ & + \text{factor1} \times \text{factor4} + \text{factor1} \times \text{factor2} \times \text{factor4}\end{aligned}$$

表 6-3 多元非线性回归模型 1 模型显著性统计量表

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-35.21979	0.40482	-87.000	< 2e-16	***
factor1	-1.14534	0.44875	-2.552	0.01623	*
factor2	1.36653	0.42481	3.217	0.00318	**
factor3	-1.02629	0.42339	-2.424	0.02182	*
factor4	0.03038	0.42453	0.072	0.94345	
factor1:factor2	-1.44498	0.50960	-2.836	0.00825	**
factor1:factor4	1.32247	0.53286	2.482	0.01911	*
factor1:factor2:factor4	1.75322	0.76550	2.290	0.02947	*

虽然该多元非线性模型的 AIC 较低，但是 factor4 的显著性极低，达到了 0.94，几乎可以认为 factor4 对模型没有线性贡献。通常来说，从一般规律考虑，当模型中含有某变量的交互项时，该交互项不可省略。

为了定义每个厄尔尼诺或拉尼娜事件的独特性，Trenberth and Stepaniak（2001）认为，NIÑO3,4 指数应与他们引入的称为反 NIÑO 指数（TNI）的指数结合使用。TNI 被定义为 Niño1 + 2 和 Niño4 地区之间归一化 SST 异常的差异。因此，TNI 测量了赤道中太平洋和东部赤道太平洋之间海表温度异常的梯度。当 SST 梯度特别大时（例如，Niño4 区为正异常，而 Niño1 + 2 区为负异常），一些研究人员将该事件归类为“太平洋中部厄尔尼诺中心”。

基于 Trenberth and Stepaniak 的研究，有理由认为 NIÑO 极东因子不单独产生影响，而是与其他变量（尤其是 ENSO 因子）结合产生影响。综上，从多元非线性模型中剔除 NIÑO 极东因子的一次项，提出第二个多元线性回归模型。模型结构如下：

$$Freezing = factor1 + factor2 + factor3 + factor1 \times factor2 + factor1 \times factor3 + factor1 \times factor4 + factor1 \times factor2 \times factor4$$

表 6-4 多元非线性回归模型 2 模型显著性统计量表

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             -35.1985     0.3885 -90.594  < 2e-16 ***
factor1                  -0.9848     0.4477  -2.200  0.03594 *
factor2                   1.3640     0.4036   3.380  0.00209 **
factor3                  -1.1038     0.4066  -2.715  0.01106 *
factor1:factor2          -1.4203     0.4928  -2.882  0.00736 **
factor1:factor3          -0.5650     0.3945  -1.432  0.16280
factor1:factor4           1.2462     0.5048   2.469  0.01970 *
factor1:factor2:factor4   1.8628     0.7006   2.659  0.01264 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.214 on 29 degrees of freedom
Multiple R-squared:  0.5596,    Adjusted R-squared:  0.4533
F-statistic: 5.265 on 7 and 29 DF,  p-value: 0.0005823
```

模型系数均在 α =0.2 下显著。除了 ENSO 与北极涛动因子的交互项外，其他系数均在 α =0.05 下显著。模型 R2 系数达到 0.56，调整后 R2 达到 0.45，说明即使从宏观因素出发建立模型，也能解释加拿大 Nunavut 地区 56%极寒天气的形成。

从模型系数来分析各因素对局地极寒天气的影响，首先看线性项。ENSO 因子与局地极寒负相关，说明 ENSO 因子越大，局地极寒越严重。说明厄尔尼诺-南方涛动剧烈的年份，即厄尔尼诺极端气候出现、太平洋中部升温与局地严寒之间有相关性。其次，AO 因子与局地极寒负相关，说明北极涛动剧烈的年份，局地极寒严重。日照通量改变与北极涛动带来的极地漩涡在分裂后会给北美洲带来强烈的降温与极寒天气，这在学界是公认的理论。

单独的热量因子对努州局地温度正相关，说明当其他条件不变时，全球热量存储的增加确实会使得局地严寒得到缓解。在问题二的分析中，本文得到了全球热量存储随时间稳定增加的结论。因此，在非厄尔尼诺年，全球的热量增加抑制了努州局地严寒的产生。但当 ENSO 因子增加，出现厄尔尼诺现象的时候，由于 ENSO 与热量因子的交互作用高于单独的热量因子的作用，此时局地严寒反而会因为全球热量的增加而加剧。

如果将 ENSO 与 NIÑO 极东因子交互项看作 ENSO Gradient 因子，那么 ENSO Gradient 因子的改变与大气系统及局地极寒气候之间有较强的联系。同时，由 ENSO Gradient 与热量因子结合形成的复杂气候变化也对局地极寒气候有显著的影响。可以看出局地极寒气候所受影响的因素是复杂多样的。

ENSO 与北极涛动因子的交互作用也对局地极寒气候有影响。这证明了 AO 与 ENSO 的联动作用存在，同时也反映了全球气候变化与局地气候异常的连锁性与复杂性。

6.4.4.1 对北极涛动因子的理解

北极涛动因子的得分主要来自于北极涛动与太阳光通量。为了理解这两个变量为何同属于一个因子，需要对这两个变量进行分析，以提供理论依据，便于理解。

对北极涛动与太阳光通量进行线性回归后，得到以下结果：

表 6-5 北极涛动与太阳光通量线性回归模型统计量

	系数估计值	标准误	t 统计量	P 值
截距项	1152.0	64.5	17.860	<2e-16 ***
solar_flux	498.5	161.5	3.086	0.00395 **
R-square:0.1915			p-value:0.003947<0.05	

虽然模型可解释性差，但是由于回归斜率项显著，一定程度上可以认为北极涛动与太阳光通量之间存在线性关系。由自然规律可推得太阳光通量增加与北极涛动变化之间的因果联系。

6.4.4.2 对太平洋年代际震荡的理解

在四个不同的因子中，太平洋年代际震荡都取得了不可忽略的得分。从 ENSO 与 NIÑO 极东因子的正得分可知，当太平洋年代际震荡逐渐从负相转为正相时，全太平洋的厄尔尼诺涛动将受影响而频繁活动。同时，整个热量系统，包括二氧化碳、海洋表面温度和海洋热量存储都有上升趋势。

在因子分解模型中，在所有变量中，太平洋年代际震荡的共性方差最小，仅为 0.67，说明其特征十分复杂，难以通过四个因子进行刻画。进一步通过气象学知识可认定，太平洋年代际震荡可被认定是模型外生变量。

6.5 全球变暖和局地极寒现象的出现之间是否矛盾？

全球变暖既可以定义为全球地表平均温度的增加，也可以定义为全球海-地-气系统的热量增加。无论是哪种定义，全球变暖与局地极寒现象的出现之间都是不矛盾的。

当全球变暖定义为全球地表均温时，由因子分析，热量因子与 GMST 之间正相关，因此，矛盾消解为全球体系热量增加与局地极寒现象之间的矛盾。

根据学界研究表明，全球海地气系统的热量绝大多数存储于海洋中。因此，全球热量增加即热量因子增加。当全球变暖定义为全球热量增加时，根据 6.4.3 节中多元非线性回归模型的分析结果，厄尔尼诺现象的加剧、缓和与热量因子增加的结合将在很大程度上影响局地极寒现象。

综上，全球变暖在厄尔尼诺现象的影响下，导致了局地极寒现象的产生。

7 问题四 解释“全球变暖”

本节将用通俗易懂的方式去说明全球变暖和局地极寒之间的相关性，并且对“全球变暖”进行概念理解。本节将对全球变暖所造成的气候影响进行逐步的解说和分析，并从前文中建立的气候变化预测模型、极端气候模型、北极涛动以及厄尔尼诺现象来说明两者之间的关联性，与此同时，寻找一个能体现出现今气候转变的趋势和复杂性的新概念去取代“全球变暖”，且新的概念能体现出现今气候转变的趋势和复杂性，让非专业人士能重新认识和理解全球气候变化的转变和未来的趋势。

7.1 全球变暖与局地极寒的关联性

在工业革命后，由于温室气体的大量排放，全球全年平均气温急速增长，直至 21 世纪后，全球全年平均气温增长速度放缓。与此同时，海洋温度却有不断升高的趋势，这说明了全球全年平均气温增长停滞的原因在于温室气体所产生的大量能量被转移到海洋之中。

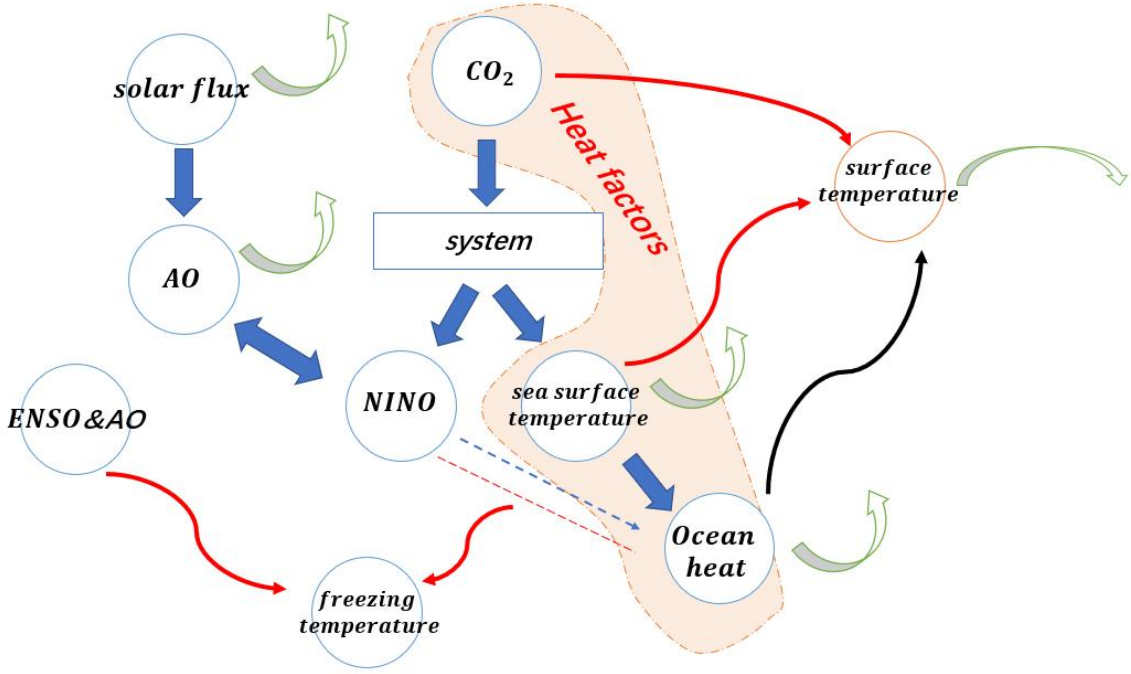


图 7-1 全球气候变化因素结构图

由第六节因子分析结论，以二氧化碳为代表的温室气体浓度、海洋表面温度及海洋热量存储指数同归属于热量指数范畴。同时，CO2 浓度与海洋表面温度、海洋热量存储之间有高度正相关。由联合国世界气象组织发布的温室气体排放量数据可知，全球范围内二氧化碳浓度的增加主要原因是碳排放量的逐年增长。因此，可作出因果推断，温室气体的增加影响了海洋表面温度与海洋热量存储。温室效应影响了整体的气候系统，导致了厄尔尼诺-南方涛动的产生，同时也通过热对流的方式改变了海洋表面温度。海洋热传导以及厄尔尼诺-南方涛动的发生将热量向海洋深处传递。根据第五节模型分析结论，二氧化碳浓度、海洋表面温度、海洋热量存储均有逐年上升趋势，三者同归属于热量系统，但是对地表平均气温的影响不同。其中二氧化碳浓度、海洋表面温度对 GMST 有正向影响，而海洋热量的增加与 GMST 间有负相关性。又根据第五节中的分析，海洋热量存储增长较快，导

致气候变暖产生了停滞。

尽管地表气候变暖产生了停滞，但由于存在温室效应，影响地气系统吸收与发射的能量平衡，热量依旧不断在地气系统累积。也就是说，表象的大气变暖停滞掩盖不了背后的系统热量增加。

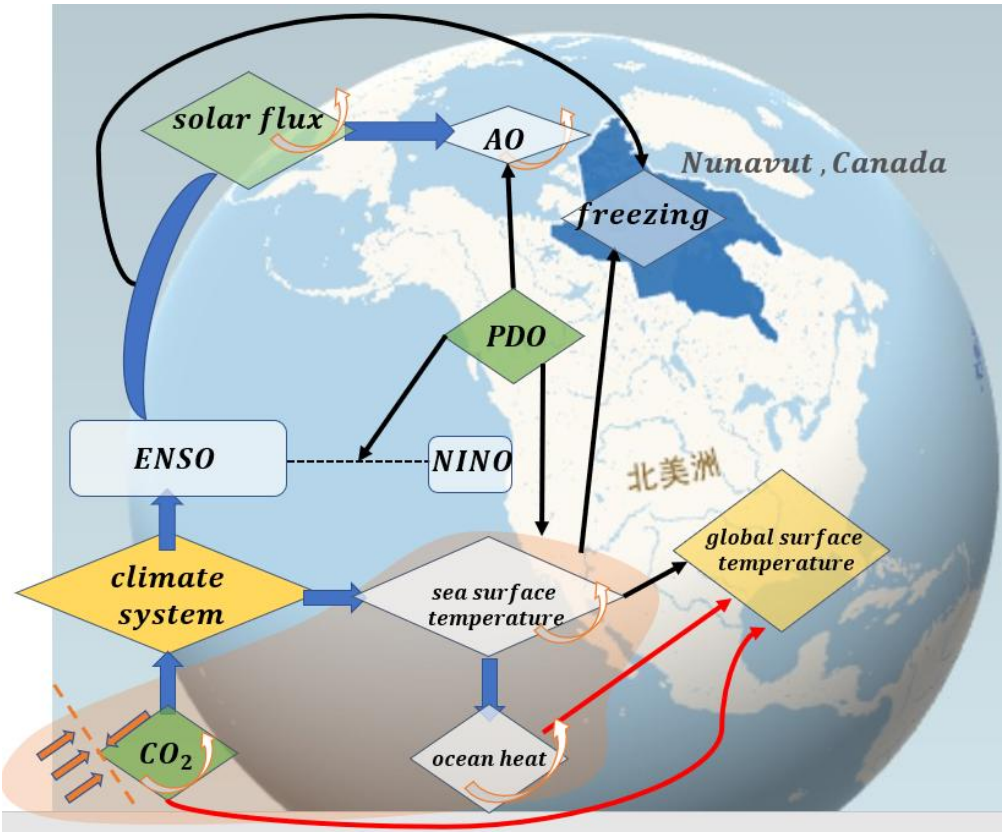


图 7-2 全球气候变化模拟图

除了 CO2 浓度的增加是气候变化的元因外，太阳光通量的增加是局地极寒的元因之一。近年来，随着日照通量全球气温的逐渐上升，北极的海冰融化，上升的大气温度增加了北极的气压变化与气流变动的不稳定性，同时也加剧了能限制冷空气流动的逆时针旋转极地旋涡在北半球冬季的不稳定情况。北极的高气压使冷空气团受到挤压往南方移动，造成北半球地区局地极寒的现象此现象称为北极涛动，2019 年冬季在美国和加拿大出现大规模的极寒现象就是北极涛动使北极冷空气转移到北美地区的例子。

Nakamura 等认为当北极涛动为正位相时，在赤道西太平洋会出现风向异常的情况，而其后该异常情况将会促使厄尔尼诺现象的发生，由此北极涛动所造成的气流异常与厄尔尼诺现象有一定的相关性，而同时伴随着厄尔尼诺现象，会在全球范围引起各种类型的极端气候现象。根据第六节的结论，ENSO-AO 的泛空间海-气系统影响着局地极寒的产生。

此外，根据第六节的结论，ENSO 与全球气候变暖的交互作用在一定程度上影响着局地极寒的产生。在地表均温不断上涨的年份，一旦出现厄尔尼诺现象，将剧烈影响局地气候。

全球变暖了，极端气象的发生频率也随之增多，极寒就是极端气象的其中一种体现，因此才会产生了全球气温变暖了，局地天气却异常寒冷的状态。

7.2 全球变暖的新概念

在全球变暖的概念下，可能会引至人们觉得地表温度在不断上升的误会。在现今全球变暖停滞的状态下，虽然地表的温度没有明显的上升，但被海洋所吸收的能量却有上涨的

趋势，地气及海气系统的总体能量仍旧在增加，而全球变暖真正应该表达的是地球整体热量的增长。由于全球变暖这个概念容易对现今复杂的气候状况出现错误的理解，也在全球变暖停滞下无法对未来的气候趋势作出准确的描述，因此选择用“全球气候异常”来替代全球变暖的概念，能更加体现出气候的未来趋势和复杂性。

全球气候异常是全球变暖概念的一个扩大，从工业革命后全球全年平均温度上升的地表温度异常，到 21 世纪后出现的全球变暖停滞，海洋温度开始上升的海洋温度异常，都是气候异常的一种体现。而且从 21 世纪起，极端气候的频发、强度的增加和规模的扩大，都充分体现了气候从时间到空间上的异常。在气候变化预测模型的预测结果来看，在未来全球平均海洋温度会有持续上升的趋势，但地表温度的上升趋势会暂缓甚至会轻微降低，而持续上升的海洋温度会相应伴随着更为频繁的全球极端气象，因此全球气候异常相较全球暖化可以充分描述全球未来气候变化的趋势。另一方面，极端气象使原来地区的气候特征在短期间内发生剧烈的改变，诸如严重的干旱、涝灾、热浪及极寒，为气候系统带来了很大的复杂性，而在全球范围内极端气象的频发使气候系统的复杂性大为增加，因此亦可视之为全球气候异常的一种情况。

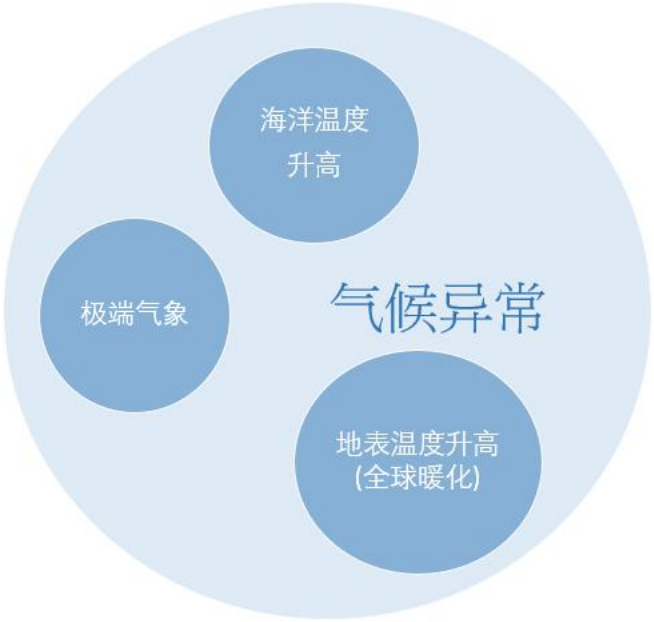


图 7-3 概念示意图

综上所述，全球气候异常将全球变暖的概念扩充，使气候观察角度从单一扩展为多元，更加充分体现出了未来地气及海气系统的热量储存上升的趋势，亦描述了极端气候的复杂性，因此，通过全球气候异常这个新的概念，可以使人们重新认识到当今气候环境的转变，同时加强人们对减能节排的意识，并通过人们对气候的关注以督促政府制定限制温室气体排放政策。

8 模型评价

8.1 模型优点

根据模型建立与求解过程中对模型以及结果的分析，现总结模型的优点如下。

- 用于解决问题二而提出的 XCSM 模型兼顾了预测准确性与可解释性：XCSM 融合了 XGBoost、协整以及 SV。其中协整模型的可解释性强，XGBoost 的预测能力强，融合了两者的优点，并在最终预测时借鉴了 Bagging 思想进行模型融合，使得预测结果更加稳健。
- 局地极寒因子分析-多元回归模型有吻合的理论支持： Trenberth 提出的 TNI 指数模型与本文提出的因子分析-多元回归模型在 NIÑO Gradient 上有一致的解释。
- 完整统一的地-海-气耦合理论：在气象学理论支持下，结合 XCSM 与因子分析-回归模型所得得出的结论，提出了完整统一的地-海-气耦合理论，分析了全球气候变化与局地极寒天气产生的原因。
- 缺失数据补全：通过多基站数据互补，一定程度上解决了气候时空数据不完整的问题。

8.2 模型缺点

根据模型建立与求解过程中对模型以及结果的分析，现总结模型的缺点如下。

- 未考虑影响因素的延迟效应：部分影响因素对于气候的影响存在延迟效应，模型在这个方面有待改进。
- 气候因素相互影响上未使用气候学模型：论文仅根据学界气候模型结论，建立了有理论依据的统计学模型，未真正建立气候学模型。在模型结构性与仿真性上有所欠缺。

参考文献

[1] Nakamura T , Tachibana Y , Honda M , et al. Influence of the Northern Hemisphere annular mode on ENSO by modulating westerly wind bursts[J]. Geophysical Research Letters, 2006, 33(7):L07709.

[2] Chen T, Guestrin C. XGBoost: a scalable tree boosting system[C]. The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), USA, 2016, 785-794.

[3]Chen X , Tung K K . Varying planetary heat sink led to global-warming slowdown and acceleration[J]. Science, 2014, 345(6199):897-903.

[4]Corinne Le Quéré, Róisín Moriarty, Andrew R M , et al. Global Carbon Budget 2014[J]. Earth System Science Data, 2015, 7(7):47-85.

[5]Hu S , Fedorov A V . The extreme ENSO of 2015–2016: the role of westerly and easterly wind bursts, and preconditioning by the failed 2014 event[J]. Climate Dynamics, 2017.

[6]England M H , Mcgregor S , Spence P , et al. Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus[J]. Nature Climate Change, 2014, 4(3):222-227.

[7]Medhaug I , Stolpe M B , Fischer E M , et al. Reconciling controversies about the 'global warming hiatus' [J]. Nature, 2017, 545(7652):41-47.

[8]Tarasova O A O A , Braathen G O G O , Barrie L A L A , et al. The state of greenhouse gases in the atmosphere using global observations through 2008[C]// Egu General Assembly Conference. EGU General Assembly Conference Abstracts, 2010.

[9]Thirumalai K , Dinezio P N , Okumura Y , et al. Extreme temperatures in Southeast Asia caused by El Ni?o and worsened by global warming[J]. Nature Communications, 2017, 8:15531.

[10]World ocean heat content and thermosteric sea level change (0–2000m), 1955–2010[J]. Geophysical Research Letters, 2012, 39.

[11]Xianyao C , Ka-Kit T . Global surface warming enhanced by weak Atlantic overturning circulation[J]. Nature, 2018, 559(7714):387-391.

[12]陈晓雷. 多元回归分析在水淹层地层水电阻率评价中的应用[D]. 黑龙江大学, 2009.

[13]高凤玲, 崔国民, 黄晓璜. CO2 的温室效应饱和度分析及其大气体积分数预测模型[J]. 上海理工大学学报, 2017(4).

[14]林 王, 文 陈, 音 马, et al. ENSO 和北极涛动对东亚冬季气候异常的综合影响[J]. Chinese Journal, 2013, 58(8):634-641.

[15]刘强, 刘嘉麒. 温室气体浓度变化及其源与汇研究进展[J]. 地球科学进展,

2000(4).
[16]徐一丹，李建平，汪秋云.全球变暖停滞的研究进展回顾[J]. 地球科学进展，2019,34(02):65-80.
[17]王敏. 因子分析法在上市公司绩效评价中的应用[D]. 石油大学（北京），2005.
孙振宇. 多元回归分析与 Logistic 回归分析的应用研究[D]. 南京信息工程大学，2008.

附录

```
#问题一时间序列分析
setwd('E:/1master/华为杯/代码')
library(readxl)
dt=read.csv('seaAndLandTemp.csv')
library(TTR)
sma=SMA((dt$Ocean_Annual),5)
write.csv(sma,'sma.csv')
lines(sma)
x=dt$Ocean_Annual-sma
x=x[5:length(x)]
acf(x)
pacf(x)
Box.test(x, lag = 6)
library(tseries)
library(forecast)
auto.arima(x)
sea=dt$Ocean_Annual
land=dt$Land_Annual
cor.test(sea,land,alternative="greater",method="pearson")

#问题二 xgboost
setwd('E:/1master/华为杯/代码')
library(readxl)
#dt=read.csv("problem2_1.csv")
dt=read.csv("year_min_temp_all_data.csv")
train_index=sample(nrow(dt),0.8*nrow(dt))
nrow(dt)
train_dt=dt[train_index,]
nrow(train_dt)
test_dt=dt[-train_index,]
nrow(test_dt)
library(xgboost)
library(readr)
library(stringr)
library(caret)
library(car)
eta_c=seq(0.1,1,length.out = 9)
mse_c=numeric(9)

for (i in 1:9){
```

```
  xgb1=xgboost(data=data.matrix(train_dt[,-1]),
               label = train_dt[,1],
               eta=eta_c[i],nrounds = 10)
  y_pred=predict(xgb1,data.matrix(test_dt[,-1]))
  mse_c[i]=mean((test_dt[,1]-y_pred)^2)
}

xgb_best=xgboost(data = data.matrix(train_dt[,-1]),
                                    label=train_dt[,1],
                                    eta=0.9,nrounds=10)
model=xgb.dump(xgb_best,with.stats=T)
names=dimnames(data.matrix(train_dt[,-1]))[[2]]
importance_matrix=xgb.importance(names,model=xgb_best)
xgb.plot.importance(importance_matrix[1:15,])
library(xgboost)
library(readr)
library(stringr)
library(caret)
library(car)
xgb_best=xgboost(data = data.matrix(train_dt[,-1]),
                 label=train_dt[,1],
                 eta=0.7750,nrounds=10)

#聚类
setwd('E:/1master/华为杯/代码')
library(readxl)
dt=read.csv("cluster_dt1.csv")
library(Hmisc)
head(dt)
dt$maxTemp=impute(dt$maxTemp,mean)
dt$minTemp=impute(dt$minTemp,mean)
dt$meanTemp=impute(dt$meanTemp,mean)
dt$totalRain=impute(dt$totalRain)
dt$totalSnow=impute(dt$totalSnow)
row.names(dt)=dt$stationName
head(dt)
result=dist(dt[,-1],method="euclidean")
result
km.res=kmeans(dt[,-1],4)
km.res$cluster
km.res$centers

dt$cluster=km.res$cluster
head(dt)
```

```
write.csv(dt,file='cluster.csv')

result_hc=hclust(d=result,method="ward.D2")
library(factoextra)
factoextra::fviz_cluster(km.res,result)

#相关系数图
setwd('E:/1master/华为杯/代码')
dt=read.csv("cluster_dt.csv")
head(dt)
library(Hmisc)
dt$maxTemp=impute(dt$maxTemp,mean)
dt$minTemp=impute(dt$minTemp,mean)
dt$meanTemp=impute(dt$meanTemp,mean)
dt$totalRain=impute(dt$totalRain)
dt$totalSnow=impute(dt$totalSnow)
names(dt)=c("基站名","州名","纬度","经度","海拔高度","最高气温","最低气温","
平均气温","总降雨量","总降雪量")
library(corrplot)
corr=cor(dt[,3:ncol(dt)])
corrplot(corr=corr,method = 'color', order ="AOE",tl.col = 'black')

setwd("F:/Cpipc2019/dataset")
all_data <- read.csv('F:/cpipc2019/dataset/all_data.csv',sep=' ')# mention the
sep
#######################################
# surface_temperature
plot(all_data$time,all_data$surface_temperature,type='l')
#lines(all_data$time,all_data$surface_temperature)
Box.test(all_data$surface_temperature)
library(tseries)
adf.test(all_data$surface_temperature)
pp.test(all_data$surface_temperature)
kpss.test(all_data$surface_temperature)
#对残差进行分析，发现其平稳且不是白噪声

surface_temperature_diff <- diff(all_data$surface_temperature)
Box.test(surface_temperature_diff)
adf.test(surface_temperature_diff)
pp.test(surface_temperature_diff)
kpss.test(surface_temperature_diff)
# 1阶差分后平稳了

##CO2
```

```
plot(all_data$time,all_data$co2_average)
lines(all_data$time,all_data$co2_average)

Box.test(all_data$co2_average)
adf.test(all_data$co2_average)
pp.test(all_data$co2_average)
kpss.test(all_data$co2_average)

co2_diff <- diff(all_data$co2_average)
plot(co2_diff)
lines(co2_diff)

Box.test(co2_diff)
adf.test(co2_diff)
pp.test(co2_diff)
kpss.test(co2_diff)

acf(co2_diff,lag=100)
pacf(co2_diff)

# NINO12
plot(all_data$time,all_data$NINO12)
lines(all_data$time,all_data$NINO12)
Box.test(all_data$NINO12)
library(tseries)
adf.test(all_data$NINO12)
pp.test(all_data$NINO12)
kpss.test(all_data$NINO12)
# stationary

##MEI_V2
plot(all_data$time,all_data$MEI_V2)
lines(all_data$time,all_data$MEI_V2)

Box.test(all_data$MEI_V2)
adf.test(all_data$MEI_V2)
pp.test(all_data$MEI_V2)
kpss.test(all_data$MEI_V2)

MEI_V2_diff <- diff(all_data$MEI_V2)
plot(MEI_V2_diff)
lines(MEI_V2_diff)

Box.test(MEI_V2_diff)
```

```
adf.test(MEI_V2_diff)
pp.test(MEI_V2_diff)
kpss.test(MEI_V2_diff)

acf(MEI_V2_diff,lag=100)
pacf(MEI_V2_diff)


##solar_flux
plot(all_data$time,all_data$solar_flux,type='l')
#lines(all_data$time,all_data$solar_flux)

Box.test(all_data$solar_flux)
adf.test(all_data$solar_flux)
pp.test(all_data$solar_flux)
kpss.test(all_data$solar_flux)
# stationary

##NINO3
plot(all_data$time,all_data$NINO3)
lines(all_data$time,all_data$NINO3)
Box.test(all_data$NINO3)
library(tseries)
adf.test(all_data$NINO3)
pp.test(all_data$NINO3)
kpss.test(all_data$NINO3)
# stationary

##NINO4
plot(all_data$time,all_data$NINO4)
lines(all_data$time,all_data$NINO4)
Box.test(all_data$NINO4)
library(tseries)
adf.test(all_data$NINO4)
pp.test(all_data$NINO4)
kpss.test(all_data$NINO4)

##pacific_decadal_oscillation
plot(all_data$time,all_data$pacific_decadal_oscillation)
lines(all_data$time,all_data$pacific_decadal_oscillation)

Box.test(all_data$pacific_decadal_oscillation)
adf.test(all_data$pacific_decadal_oscillation)
pp.test(all_data$pacific_decadal_oscillation)
```

```
kpss.test(all_data$pacific_decadal_oscillation)

pacific_decadal_oscillation_diff <- diff(all_data$pacific_decadal_oscillation)
plot(pacific_decadal_oscillation_diff)
lines(pacific_decadal_oscillation_diff)

Box.test(pacific_decadal_oscillation_diff)
adf.test(pacific_decadal_oscillation_diff)
pp.test(pacific_decadal_oscillation_diff)
kpss.test(pacific_decadal_oscillation_diff)

acf(pacific_decadal_oscillation_diff,lag=100)
pacf(pacific_decadal_oscillation_diff)

qqnorm(pacific_decadal_oscillation_diff)
jarque.bera.test(pacific_decadal_oscillation_diff)
# after DIFF, turned out to be a Gaussion Noise

##southern_oscillation
plot(all_data$time,all_data$southern_oscillation)
lines(all_data$time,all_data$southern_oscillation)

Box.test(all_data$southern_oscillation)
adf.test(all_data$southern_oscillation)
pp.test(all_data$southern_oscillation)
kpss.test(all_data$southern_oscillation)

southern_oscillation_diff <- diff(all_data$southern_oscillation)
plot(southern_oscillation_diff)
lines(southern_oscillation_diff)

Box.test(southern_oscillation_diff)
adf.test(southern_oscillation_diff)
pp.test(southern_oscillation_diff)
kpss.test(southern_oscillation_diff)

acf(southern_oscillation_diff,lag=100)
pacf(southern_oscillation_diff)
#good!

##ocean_heat
#need no test

#sea_surface_temperature
```

```
plot(all_data$time,all_data$sea_surface_temperature)
lines(all_data$time,all_data$sea_surface_temperature)

Box.test(all_data$sea_surface_temperature)
adf.test(all_data$sea_surface_temperature)
pp.test(all_data$sea_surface_temperature)
kpss.test(all_data$sea_surface_temperature)

sea_surface_temperature_diff <- diff(all_data$sea_surface_temperature)
plot(sea_surface_temperature_diff)
lines(sea_surface_temperature_diff)

Box.test(sea_surface_temperature_diff)
adf.test(sea_surface_temperature_diff)
pp.test(sea_surface_temperature_diff)
kpss.test(sea_surface_temperature_diff)

acf(sea_surface_temperature_diff,lag=100)
pacf(sea_surface_temperature_diff)

################################################
#回归模型 1 co2
series2=cbind(all_data$surface_temperature,all_data$co2_average)
library(MTS)
MTSplot(series2)
m1=lm(all_data$surface_temperature~all_data$co2_average)
summary(m1)  # R^2=0.7031

#残差检验
res=m1$residuals
adf.test(res)           ## null hypothesis is that x has a unit root
pp.test(res)              ## null hypothesis is that x has a unit root
kpss.test(res)           ## Null hypothesis is stationarity
plot(density(res))
library(tseries)
jarque.bera.test(res)
library(urca)
summary(ur.df(res,type="none",selectlags="AIC")) #协整检验

#伪回归
library(car)
durbinWatsonTest(m1)
library(lmtest)
dwtest(m1)      #D 值>R2,没有伪回归
```

```
####

#回归模型 2 co2+MEI_V2
series2=cbind(all_data$surface_temperature,all_data$co2_average,all_data$MEI_
V2)
library(MTS)
MTSplot(series2)
m2=lm(all_data$surface_temperature~all_data$co2_average+all_data$MEI_V2)
summary(m2)  # R^2=0.7031

#残差检验
res=m2$residuals
adf.test(res)           ## null hypothesis is that x has a unit root
pp.test(res)              ## null hypothesis is that x has a unit root
kpss.test(res)           ## Null hypothesis is stationarity
plot(density(res))
library(tseries)
jarque.bera.test(res)
library(urca)
summary(ur.df(res,type="none",selectlags="AIC")) #协整检验

#伪回归
library(car)
durbinWatsonTest(m2)
library(lmtest)
dwtest(m2)      #D 值>R2,没有伪回归

#回归模型 3 co2+MEI_V2+southern_oscillation
series3=cbind(all_data$surface_temperature,all_data$co2_average,all_data$MEI_
V2,all_data$southern_oscillation)
library(MTS)
MTSplot(series3)
m3=lm(all_data$surface_temperature~all_data$co2_average+all_data$MEI_V2+all_d
ata$southern_oscillation)
summary(m3)  # R^2=0.7031

m3A=lm(all_data$MEI_V2~all_data$southern_oscillation)
summary(m3A)  # R^2=0.7031
#MEI_V2 and southern_oscillation are highly correlated.



#回归模型 4 co2+MEI_V2+ocean_heat
series2=cbind(all_data$surface_temperature,all_data$co2_average,all_data$MEI_
```

```
V2,all_data$ocean_heat)
library(MTS)
MTSplot(series2)
m2=lm(all_data$surface_temperature~all_data$co2_average+all_data$MEI_V2+all_d
ata$ocean_heat)
summary(m2)  # R^2=0.7031

#残差检验
res=m2$residuals
adf.test(res)           ## null hypothesis is that x has a unit root
pp.test(res)              ## null hypothesis is that x has a unit root
kpss.test(res)           ## Null hypothesis is stationarity
plot(density(res))
library(tseries)
jarque.bera.test(res)
library(urca)
summary(ur.df(res,type="none",selectlags="AIC")) #协整检验

#伪回归
library(car)
durbinWatsonTest(m2)
library(lmtest)
dwtest(m2)      #D 值>R2,没有伪回归

#回归模型 4 co2+MEI_V2+ocean_heat+sea_surface_temperature
series2=cbind(all_data$surface_temperature,all_data$co2_average,all_data$MEI_
V2,all_data$ocean_heat,all_data$sea_surface_temperature)
library(MTS)
MTSplot(series2)
m2=lm(surface_temperature~co2_average+MEI_V2+ocean_heat+sea_surface_temperatu
re,data = all_data)
summary(m2)  # R^2=0.7031

#残差检验
res=m2$residuals
adf.test(res)           ## null hypothesis is that x has a unit root
pp.test(res)              ## null hypothesis is that x has a unit root
kpss.test(res)           ## Null hypothesis is stationarity
plot(density(res))
library(tseries)
jarque.bera.test(res)
library(urca)
summary(ur.df(res,type="none",selectlags="AIC")) #协整检验
```

```
#伪回归
library(car)
durbinWatsonTest(m2)
library(lmtest)
dwtest(m2)      #D 值>R2,没有伪回归


########################################
########################################
########################################
# 数据处理
factors                                                                    <-
cbind(predict.co2,predict.MEI_V2,predict.ocean_heat,predict.sea_surface_tempe
rature)
factors <- data.frame(factors)
###

colnames(factors)=c('co2_average','MEI_V2','ocean_heat','sea_surface_temperat
ure')

coin_pred <- predict(m2, factors, interval="prediction", se.fit = TRUE)
pred1 <- coin_pred$fit
hs.lm.pred <- pred1[,1]

# forecast
plot(ts(hs.lm.pred,start=437),col="blue",type='l')

par(mfrow=c(1,1))
plot(all_data$surface_temperature,col="black",type='l',xlim=c(1,800))
#lines(hw.forecast,col="red")
#lines(ts(hs.lm.pred,start=437),col="blue",type='l')# MCMC

xgboost2_pred_temp <- read.csv('xgboost2_pred_temp.csv')
xgboos.pred <- xgboost2_pred_temp$surface_temperature
weight1 <- sd(xgboos.pred)/(sd(hs.lm.pred)+sd(xgboos.pred))
final_pred <- weight1*hs.lm.pred+(1-weight1)*xgboos.pred
lines(ts(final_pred,start=437),col="blue",type='l')

## smoothing
ts_sma=SMA(all_data$surface_temperature,90)
plot(ts_sma,xlim=c(1,800),ylim=c(0,1),type='l',lwd=2)

#pred_sma <- SMA(ts(final_pred,start=437),10)
#plot(pred_sma,xlim=c(1,800),ylim=c(0,1),type='l')
```

```
time_all <- SMA(c(all_data$surface_temperature,final_pred),90)
lines(window(ts(time_all),start=437),col=2,lwd=2)
#lines(ts(time_all),col=3)
#predict.co2
window(y.predict.t,start=437,end=437+25*12-1)+hw.forecast+hs.lm.pred
#plot(factors$co2_average)
#plot(factors$MEI_V2)
#plot(factors$ocean_heat)
#plot(factors$sea_surface_temperature)

write.csv(factors,"factors.csv")
write.csv(hs.lm.pred,"mcmc_pred_temp.csv")
```