**五、模型的建立与求解**

5.1 数据预处理

5.1.1 数据筛选

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

最终筛选出的7个城市的数据具有全球代表性：

1. 所选地点能代表全球主要气候类型。

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

1. 所选地点在全球范围内达到基本均匀分布。

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

5.1.2 缺失数据的处理

筛选后的数据仍然存在部分缺失值，但此处本文没有对缺失值直接进行简单的删除处理而是选择补全数据。因为7个城市中仅有印度新德里存在数据缺失，具体时期是1855年3月、1856年5月、1858-1869年，可见缺失值处在时间序列的中部。为确保时间序列的连续性和后续数据分析数据的完整性和结果的准确性，我们进行了对空缺数据的平均值插值补全。由于温度数据属于稳定数据，在十年间隔中不会出现快速增长或下降的剧烈幅度变化，因此平均值插值补全方法具有可行性。

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

其中t2表示缺失月份平均温度，t1表示同地区前一非空缺值年份的同月份平均温度，t3表示同地区后一非空缺值年份同月份平均温度。如：t2表示1855年3月新德里温度，t1表示1854年3月新德里温度，t3表示1856年3月新德里温度。

得到修正后的数据如下表所示：

表5-1-1 缺失颜色数据修补

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

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

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

5.1.3 世界气温的表示

本文使用筛选和补全后的数据计算世界气温，年度世界气温计算公式如下：

月度世界气温计算公式如下：

其中表示年度世界气温，表示年度世界气温。该年表示地区第月的月平均气温。

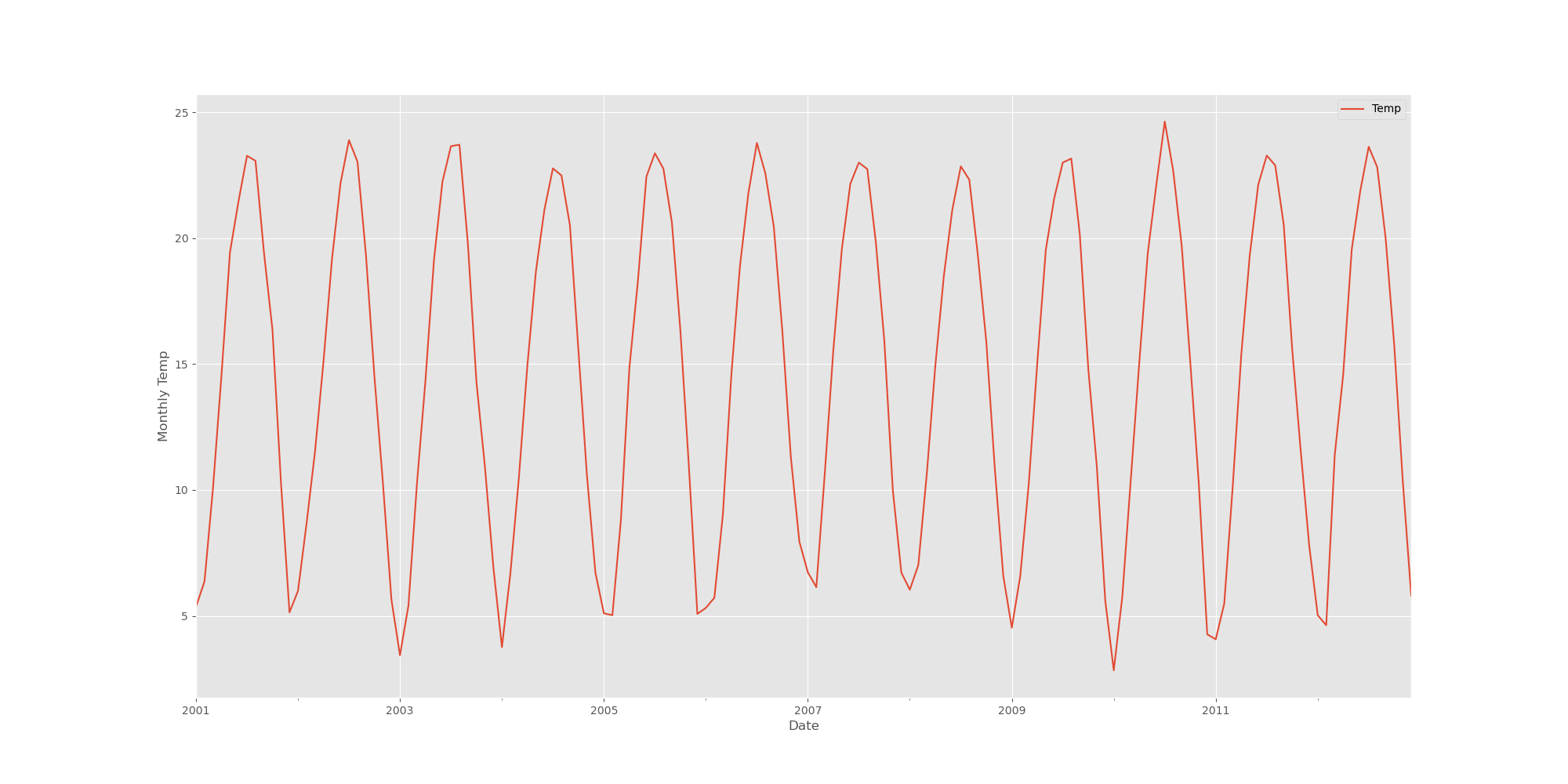
下表为1851-2012年年度世界气温的节选展示，完整版见附件表【】【】【】，

|  |  |
| --- | --- |
| Year |  |
| 1860 | 12.821 |
| 1880 | 13.437 |
| 1900 | 13.713 |
| 1920 | 13.784 |
| 1940 | 13.067 |
| 1960 | 13.789 |
| 1980 | 13.554 |
| 2000 | 14.672 |

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

|  |  |
| --- | --- |
| Month |  |
| 2001-01-01 | 5.36 |
| 2001-02-01 | 6.37 |
| 2001-03-01 | 10.08 |
| 2001-04-01 | 14.56 |
| 2001-05-01 | 19.44 |
| 2001-06-01 | 21.46 |
| 2001-07-01 | 23.27 |
| 2001-08-01 | 23.07 |
| 2001-09-01 | 19.42 |
| 2001-10-01 | 16.35 |
| 2001-11-01 | 10.29 |
| 2001-12-01 | 5.14 |

2001年1月至2012年12月的月度世界平均气温情况绘制如图：

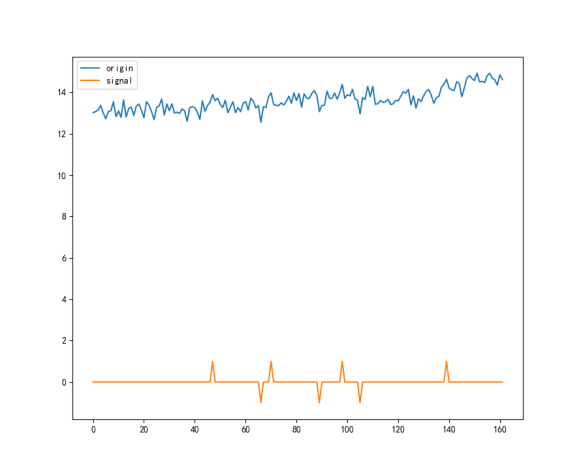


5.1.4 异常数据的处理

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

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

利用Smoothed z-score方法对5.1.3中的世界气温异常值识别代码见附件【】【】【】，识别结果如下图：



结果显示1894年、1917年、1921年、1940年、1949年、1956年、1990年为异常点，对上述年份数据剔除并使用5.1.2的缺失数据补齐方法修正结果，最终结果见附件【】【】【】

5.2.2.1 模型准备

基于历史数据，对全球温度水平的过去进行描述并且对未来进行预测，需要我们对全球不同年份的平均气温利用模型，进行数据处理与分析。原始数据有来自很多国家不同的100个地区的平均月气温，跨度从1851年1月到2013年9月。这些数据过于庞大且复杂，并不利于模型的建立以及模型对数据的描述与预测。所以我们用前文提到处理好的数据，结合搜索到的数据集合成1851年到2022年(截止到十月)每年平均温度放入模型进行分析与预测。

由于研究的内容是温度随着时间的变化规律及根据规律预测，在准备建立模型的过程中更倾向于使用时间序列模型，故引入三个模型：Gray Forecast [Model](https://so.csdn.net/so/search?q=Model&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/qq_39798423/article/details/_blank)(简称为GF)、Autoregressive Integrated Moving Average Model(简称为ARIMA)、Simple Exponential Smoothing Prediction Model(简称为SESP).其中，ARIMA Model、SESP Model为基于时间序列的预测模型。

根据数据集以及前文分析，对全球温度水平过去的描述需要1851年至2022年的全体数据；若要预测2050年及2100年的全球温度水平，由于全球平均气温在近十年间增速远远大于前面的年份，于是在使用模型预测时，决定基于全体数据及近十年数据分别预测，得出结果并对比。

具体数据文件见附件：

5.2.2.2 模型建立

1. ARIMA Model

(a) Autoregressive Model(简称为AR)

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

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

然而自回归模型存在着一些不利的限制，包括：   
1、自回归模型完全是在自身数据的基础上进行的预测 。  
2、必须具有平稳性和相关性，也就是说自相关系数小于0.5的情况，并不合适使用自回归模型。

(b) Moving Average Model(简称为MA)

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

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

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

1. ARIMA Model

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

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

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

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

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

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

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

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

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

5.2.2.2 模型实现

一、确定参数值

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

季节性ARIMA的参数组合表示示例如下：

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

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

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

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

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

关于确定AIC值和季节性ARIMA的参数组合的代码运行结果如下：

Machine precision = 2.220D-16

N = 9 M = 10

At X0 0 variables are exactly at the bounds

At iterate 0 f= 6.97141D-01 |proj g|= 2.21564D-01

This problem is unconstrained.

At iterate 5 f= 6.35361D-01 |proj g|= 1.15777D-01

At iterate 10 f= 5.85957D-01 |proj g|= 3.43119D-02

At iterate 15 f= 5.84794D-01 |proj g|= 2.00194D-04

At iterate 20 f= 5.84793D-01 |proj g|= 3.09161D-05

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

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

F = 0.62897435703533366

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

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

二、模型拟合和检验

接下来进行拟合模型：

Machine precision = 2.220D-16

N = 9 M = 10

At X0 0 variables are exactly at the bounds

At iterate 0 f= 6.97141D-01 |proj g|= 2.21564D-01

At iterate 5 f= 6.35361D-01 |proj g|= 1.15777D-01

At iterate 10 f= 5.85957D-01 |proj g|= 3.43119D-02

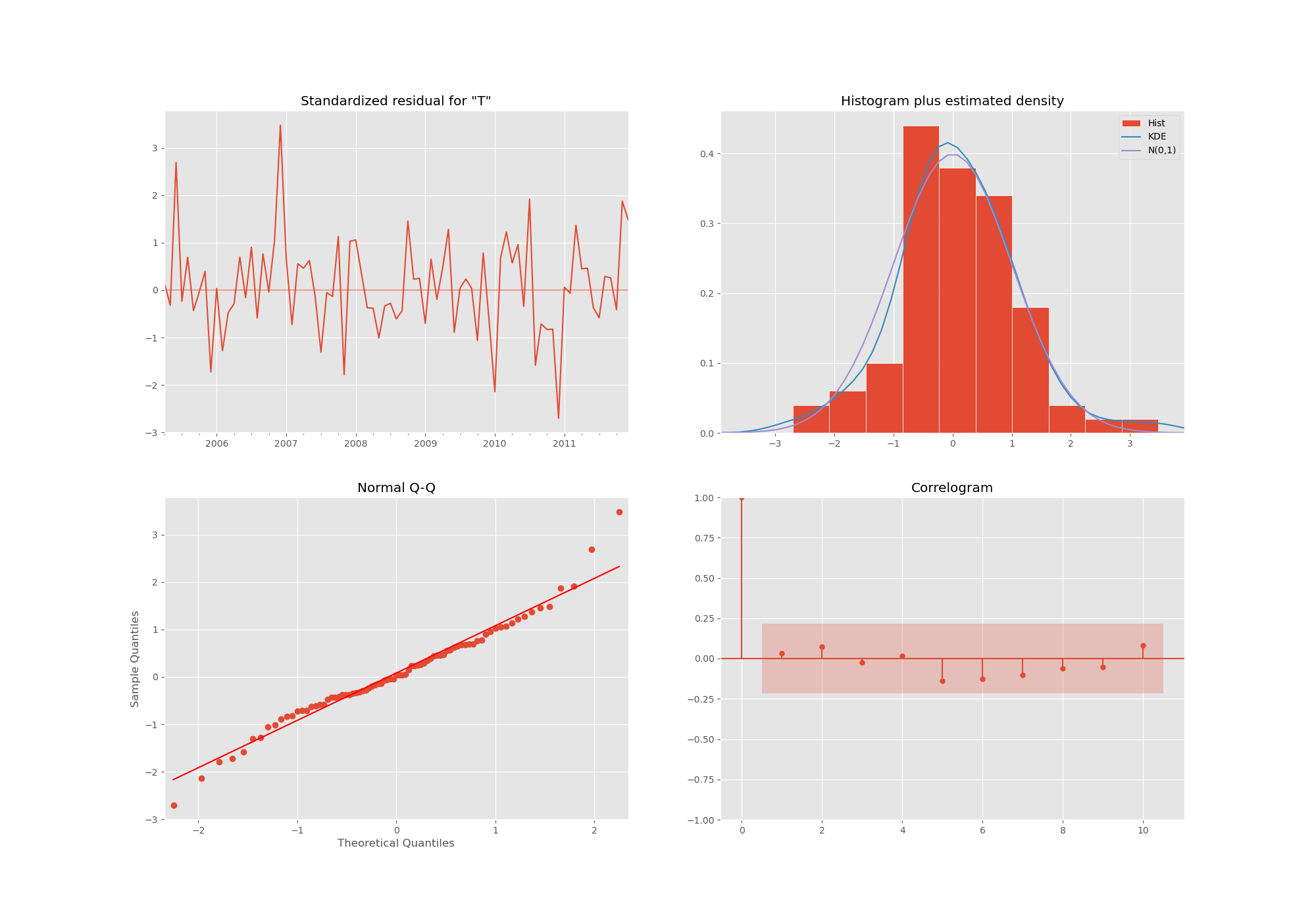
At iterate 15 f= 5.84794D-01 |proj g|= 2.00194D-04

At iterate 20 f= 5.84793D-01 |proj g|= 3.09161D-05

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

F = 0.58479347086354305

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



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

的取值范围为[-1，1]，分子代表了yt和yt-1到yt和yt-k的相关系数 ，分母代表了方差。

三、预测结果

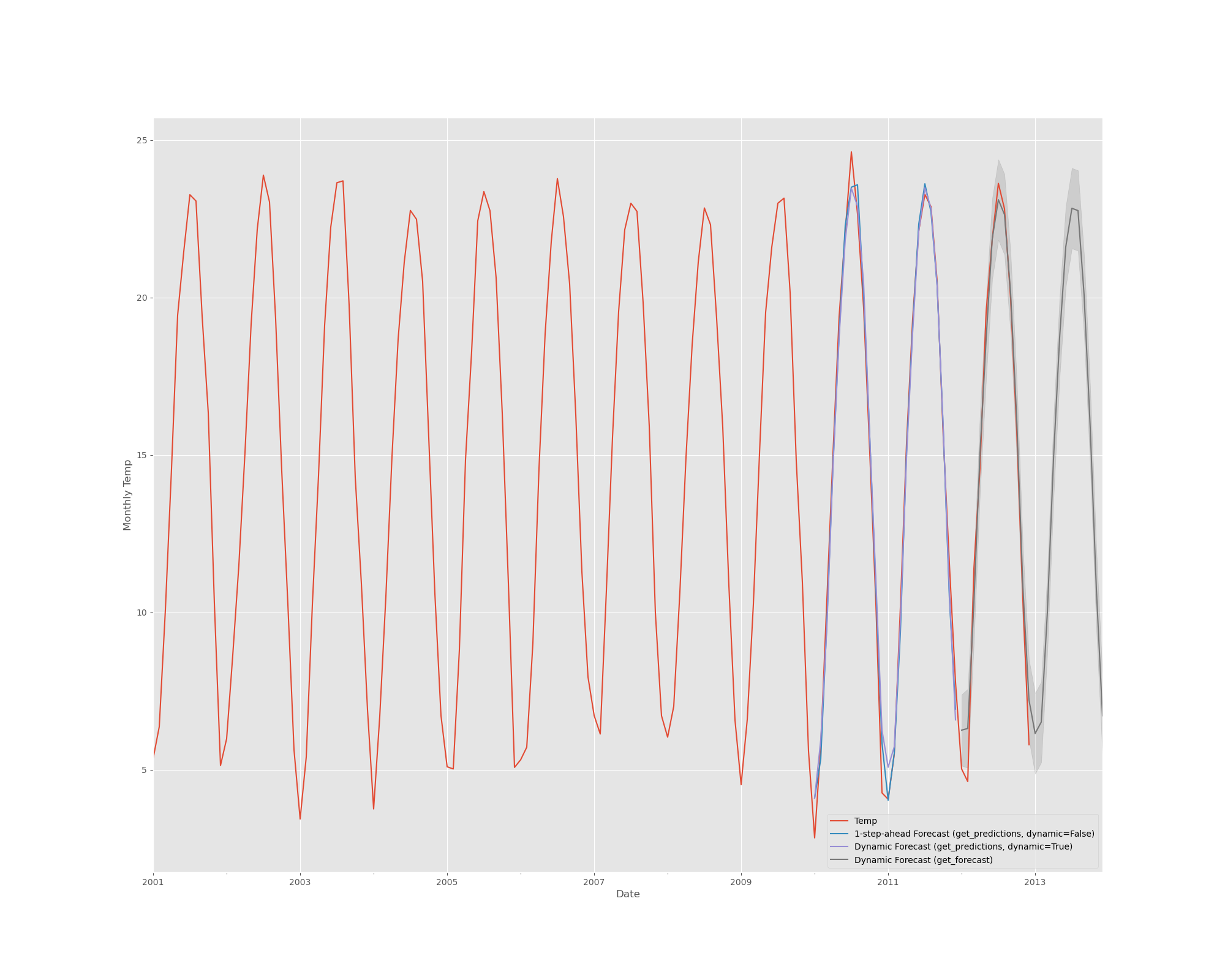
现在就可以进行对未来天气情况的预测工作了。我们将使用下述三种方法实现结果的预测:

1. 在样本预测中，提前一步预测上一年度(2011年)，具体来说，提前一步预测就是指用预测的每个点预测下一个点，预测结果在代码中计作pred0。
2. 在过去一年(2011年)的动态预测样本预测中。同样，该模型用于预测模型所基于的数据。预测结果在代码中计作pred1。
3. 样本外数据的“真实”预测。在这种情况下，模型被要求预测它以前没有输入过的数据。也是我们的目标预测结果，这部分的预测结果在代码中计作pred2。

我们成功预测出了2013年12个月的全球气温情况，见下表：

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

把预测结果全部绘制如图：



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

5.2.2.4 模型评价

1. ARIMA Model

代码算法评价指标使用的是：MAPE和MSE

a、MSE均方误差

为了量化2012年全球气温预测的准确性，我们计算均方误差指标，均方误差MSE具体来说就是利用模型预测出的参数的估计值和参数的真实值之间的差的平方的期望。MSE是一种常用的模型评价的度量--可以做到刻画预测值与被预测量之间的差异程度。因此可以说MSE是衡量我们的预测模型的平均误差的一种非常便捷的方案。MSE 可以评价数据的变化程度，注意，MSE的值越小，就意味着我们的预测模型描述实验的数据有着更加优秀的准确度。MSE公式如下：

这里N表示测量次数，observed\_t表了参数的真实值，predict\_t代表了参数的预测值。

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

b、MAPE平均绝对百分比误差

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

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