刘锐,高灿中,罗晓林. 基于统计分析和马尔可夫链的日照数据质控[J]. 陕西气象,2025(2):55-61.
文章编号:1006-4354(2025)02-0055-07

基于统计分析和马尔可夫链的日照数据质控

刘　锐[1],高灿中[1],罗晓林[2]
(1. 巴中市气象局,四川巴中　636000;2. 巴州区气象局,四川巴中　636000)

摘　要:观测数据的准确性是做好气象预报和服务的前提。特别是地面观测自动化改革后,如何提高数据质控水平,缩短发现异常数据时间,是观测业务工作的重点。利用统计分析和马尔可夫链两种方法,对巴中地区4个国家基本气象站1971—2022年的日照时数分析,探讨日照数据质控的可行性。结果表明:(1)各站的月日照时数符合正态分布,其人工观测极大值与气候异常值吻合度极高,统计分析中关于异常数据的识别方法在数据质控上有较好的适用性。(2)马尔可夫链在月日照时数的状态预测上,夏季和秋季效果较好;而冬春季的质控标准更宜用气候异常值标准和统计异常值标准。(3)业务人员应按照业务要求做好设备的日常维护,当月日照时数连续两个月以上超出马尔可夫链预测状态或气候异常值时,应加强维护,必要时可以考虑更换设备。

关键词:日照时数;正态分布;马尔可夫链;数据质控
中图分类号:P468.027　　**文献标识码**:A

2019年7月1日起,光电式数字日照计正式投入业务运行,其观测的数据作为正式记录。但是观测设备故障、维护不到位等因素,极易造成数据失真,特别是地面观测业务自动化后,日常值守班任务被取消[1],人工质控任务弱化,导致发现日照、能见度、风向等异常数据的时间更长,从而影响数据的可用性。故研究气象数据质控方法,优化数据质控区间很有必要。针对日照数据,杨小梅[2]、李金建[3]、胡慧敏[4]、苗运玲[5]、黄肖寒[6]等分别对西南地区、四川、哈密绿洲、广西罗成等地的日照特征进行分析,讨论了年日照时数的变化趋势和影响日照时数变化的气象因子;迟庆红[7]、黄维[8]、杨炳玉[9]等将自动观测日照数据和人工观测数据进行对比,针对数据差异来源和可能存在的问题进行了分析。但目前针对日照时数的质控方法研究相对较少。马尔可夫链分析是以时间序列内部概率分布结构为出发点,从时间序列中总结随机过程的概率规律,在灾害性天气现象的出现、天气形势的转变和降水预测等方面有很好的应用[10-12]。日照时数的变化随机性较大,故通过马尔可夫链对日照时数进行预测,探讨数据质控的一种新思路。

选取四川省巴中市南江、巴中、通江、平昌4个国家基本气象站(下称基本站),1971—2018年的人工观测日照时数进行分析,了解巴中地区日照特征。利用马尔可夫链预测2016—2018年的每月数据,通过实况数据进行验证,从而论证用统计方法和马尔可夫链对未来日照时数进行质控的可行性。最后再利用两种方法对2019—2022年的自动观测日照时数进行质控分析,查找异常值。

1　资料来源与研究方法

1.1　资料来源

所用气象资料来源于“气象大数据云平台·天擎”,选用巴中市各县(区)4个基本站1971—2022年逐日日照时数资料。其中1971—2018年为人工观测日照时数,2019—2022年为自动观测日照时数。

收稿日期:2023-12-22
作者简介:刘锐(1990—),男,汉族,四川巴中人,学士,高级工程师,主要从事综合气象业务工作。
基金项目:巴中市气象局科学技术研究开发课题(2023-003)

1.2 *方法*

1.2.1 K-S检验 K-S检验是检验单一样本是否来自某一特定分布的方法。通过对样本数据的累计频数分布与特定理论分布比较,在给定的显著性水平下(常取 $\alpha=0.05$),推论该样本是否取自某特定的分布族[13]。本文通过SPSS软件对4个基本站的每月数据进行正态分布检验。

1.2.2 分组法 为使时间序列的频数分布能正确地反映出观测数据的特性,本文利用序列实测的极值情况进行分组,即将实测最大值和最小值分别置于其所在组的组中值位置附近[14]。即

$$h=\frac{X_{\max}-X_{\min}}{n_k-1}, \tag{1}$$

其中 h 为组间距,$X_{\max}$为序列最大值,$X_{\min}$为序列最小值,n_k 为分组数。

1.2.3 马尔可夫链 马尔可夫过程是研究事物的状态和状态转移规律的理论,是利用某一变量现在的状态和动向去预测其未来状态的一种分析手段[15]。即已知某一随机过程"现在"的条件下,其"将来"与"过去"是独立的。其定义为:假设马尔可夫过程$\{X_n,n\in T\}$的参数集 T 是离散的时间集合,即 $T=\{0,1,2,\cdots\}$,其状态空间为 $I=\{i_0,i_1,i_2,\cdots\}$。则有随机过程$\{X_n,n\in T\}$,若对任意的整数 $n\in T$ 和任意的 $i_0,i_1,\cdots,i_{n+1}\in I$,条件概率 P 满足

$$P=\{X_{n+1}=i_{n+1}\mid X_0=i_0,X_1=i_1,\cdots,X_n=i_n\}=P\{X_{n+1}=i_{n+1}\mid X_{n=i_n}\}。 \tag{2}$$

则称$\{X_n,n\in T\}$为马尔可夫链,其转移概率是马尔可夫过程各状态之间演变的概率统计特征。在一个马尔可夫链中,系统中从一种状态转移至另一状态或与自身相同的状态,若共有 m 种状态,记为 $E_i(i=1,2,\cdots,m)$,把状态 E_i 转移到状态 $E_j(j=1,2,\cdots,m)$事件的概率称为转移概率,记为 $P_{ij}=P(E_j\mid E_i)$[16]。

由各状态的各种转移概率为元素组成的矩阵称为转移概率矩阵,记为 $\boldsymbol{P}_{ij}$ 或

$$\boldsymbol{P}_{ij}=\begin{bmatrix}P_{11} & P_{12} & \cdots & P_{1m}\\ P_{21} & P_{22} & \cdots & P_{2m}\\ \cdot & \cdot & \cdots & \cdot\\ P_{m1} & P_{m2} & \cdots & P_{mm}\end{bmatrix}。 \tag{3}$$

叠加马尔可夫链和加权马尔可夫链是常用的两种预测方法。通过对两者在2016—2018年的预测结果进行对比分析,两者的年预测准确率的平均值相同,为计算方便,选用叠加马尔可夫链对日照状态进行预测。其模型构建步骤如下。

1)建立相关序列值的分级标准。

2)对研究序列进行马氏性检验;通常用以下公式进行检验:

$$\chi^2=2\sum_{i=1}^{m}\sum_{j=1}^{m}f_{ij}\left|\ln\frac{P_{ij}}{P_J}\right|。 \tag{4}$$

其中 f_{ij} 为转移频数,表示序列从状态 i 经过一步转移到达状态 j 的频数;由 f_{ij} 组成的矩阵为转移频数矩阵。P_J 为边际概率,即将转移频数矩阵的第 j 列之和除以各行各列的总和,m 为相关序列值分级的级数。

3)计算得到不同步长的马尔可夫链状态转移矩阵。

4)分别以前面若干时段的指标值为初始状态,结合其相应的各阶转移概率矩阵,即可预测出该时段指标值的状态 $P_i^{(k)}$,k 为步长,$k=1,2,\cdots,m$。

5)将同一状态的各预测概率加权和 p_i 作为指标值处于该状态时的预测概率,即叠加马尔可夫链预测[17]。

$$p_i=\sum_{k=1}^{m}P_i^{(k)}。 \tag{5}$$

p_i 的最大值所对应的状态即为该时段指标值的预测状态。

马尔可夫链因具有时间离散、状态离散和无后效性的特点,而气象要素的演变在实际中往往是连续的,而且前面的状态对后期是有一定的影响,同时状态划分往往也影响了分析方法的成败。这是马尔可夫链分析方法存在的局限性。

2 统计分析方法在日照时数质控方面的应用

对1971—2018年巴中地区各月的人工观测日照时数按照单样本K-S检验。结果表明:南江站和巴中站各月均符合正态分布,且通过0.05的显著性检验;平昌站和通江站除2月未通过检验外,其余月均满足正态分布。根据气候术语[18]和正态分布的特征,对比分析每月人工观测的极大

值与气候异常值、统计异常值的关系。将其中每月的平均值 μ 与 2 倍标准差 σ 之和,即 $\mu+2\sigma$ 作为气候异常值;根据统计学中常用的拉伊达准则[19],每月的平均值与 3 倍标准差之和,即 $\mu+3\sigma$ 为统计异常值。

从图 1 中可分析出各站的人工观测极大值与气候异常值的变化趋势一致,且数值相近。故将人工观测的极大值作为每月日照时数质控时的可疑值,将统计异常值作为错误值。

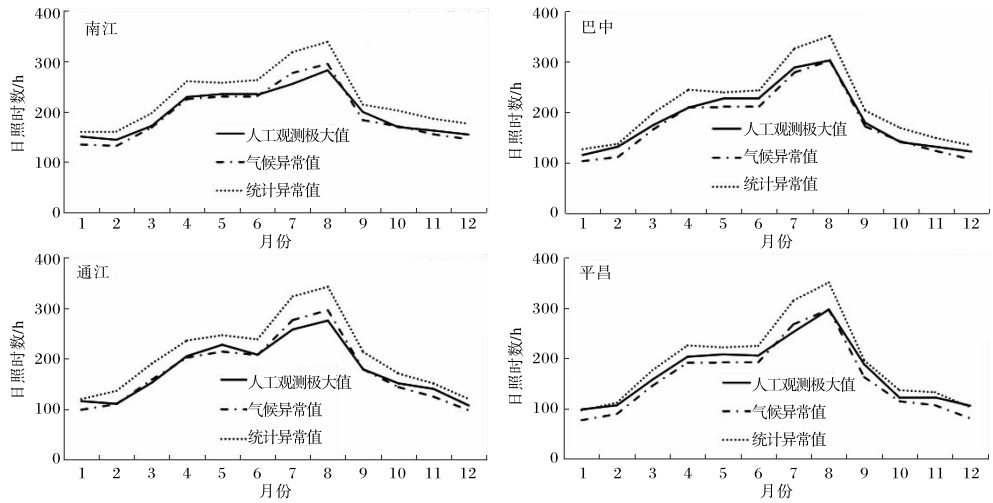

图 1　1971—2018 年巴中各基本站人工观测极值与正态分布异常值对比图

3　马尔可夫链预测月日照时数状态的可行性检验

通过对巴中地区各基本站的人工观测月日照时数进行分析,可知拉依达准则在月日照时数质控上是可行的。但在正态分布中,$\mu-3\sigma$ 到 $\mu+3\sigma$ 的随机变量分布占总数的 99.7%,其质控范围较大。为缩小日照时数质控区间,以南江站为例进行马尔可夫链分析。考虑预测结果主要用于对日照时数异常偏多的质控,故预测状态大于实际状态,也视为预测准确。

3.1　模型检验

1971—2018 年南江站 5 月序列月日照时数极大值 $X_{\max}=234.7$ h,极小值 $X_{\min}=113.4$ h,将样本数据分为 5 组,组间距为 30.0 h。因此可分为异常偏少 E_1($X<125.0$ h)、偏少 E_2(125.0 h$\leqslant X<155.0$ h)、正常 E_3(155.0 h$\leqslant X<185.0$ h)、偏多 E_4(185.0 h$\leqslant X<215.0$ h)和异常偏多 E_5($X\geqslant 215.0$ h)五个状态。根据状态分级标准,划分 1971—2018 年南江 5 月日照时数状态,计算出步长为 1 a 的一步转移频数矩阵$(\boldsymbol{f}_{ij})_{5\times5}$和转移概率矩阵$(\boldsymbol{P}_{ij})_{5\times5}$,并对其进行马氏性检验。

$$(\boldsymbol{f}_{ij})_{5\times5}=\begin{bmatrix}0&1&1&1&0\\0&2&2&2&1\\0&2&7&8&1\\3&3&6&5&0\\0&0&2&0&0\end{bmatrix}$$

$$(\boldsymbol{P}_{ij})_{5\times5}=\begin{bmatrix}0&1/3&1/3&1/3&0\\0&2/7&2/7&2/7&1/7\\0&2/18&7/18&8/18&1/18\\3/17&3/17&6/17&5/17&0\\0&0&2/2&0&0\end{bmatrix}$$

利用公式(2),可知统计量 $\chi^2=147.366$,给定显著性水平 $\alpha=0.05$ 下,通过 χ^2 分布表[15]可知,分位点 $\chi^2((m-1)^2)=\chi^2(16)=26.296$,故 5 月日照时数序列满足马氏性。其他各月的统计量 χ^2 分别为:1 月 65.017、2 月 38.253、3 月 93.891、4 月 96.278、6 月 91.198、7 月 44.901、8 月 128.396、9 月 128.489、10 月 95.966、11 月 87.268、12 月 72.269,均大于 26.296。

3.2　马尔可夫链预测

利用 2011—2015 年的月日照时数状态情况,通过叠加马尔可夫链对 2016 年 5 月的数据进行

预测。先得到步长为 1～5 a 的转移概率矩阵($\boldsymbol{P}_{ij}$)$_{5\times5}$,再对 2016 年的 5 月日照时数状态进行预测。由表 1 可知:p_i 的最大概率为 2.013,此时 $i=4$,及 5 月的预测状态为 E_4(185.0 h≤X<215.0 h),而实际状态为 167.2 h,为 E_3,较实际情况偏高。

表 1 2016 年南江站 5 月日照时数状态的最大概率预测计算表

初始年	状态	步长/a	E_1出现概率	E_2出现概率	E_3出现概率	E_4出现概率	E_5出现概率
2015	E_4	1	3/16	3/16	5/16	5/16	0
2014	E_3	2	2/15	1/15	1/3	2/5	1/15
2013	E_2	3	0	1/6	1/3	1/2	0
2012	E_2	4	1/5	0	2/5	2/5	0
2011	E_4	5	0	2/15	1/3	2/5	2/15
P_i(累加和)			0.521	0.554	1.713	2.013	0.200

注:E_1 异常偏少(X<125.0 h)、E_2 偏少(125.0 h≤X<155.0 h)、E_3 正常(155.0 h≤X<185.0 h)、E_4 偏多(185.0 h≤X<215.0 h)和 E_5 异常偏多(X≥215.0 h),下同。

同理,对 2016—2018 年其他月进行预测,其预测结果见表 2。从预测结果与实际情况分析可知(如预测状态大于实际状态,预测结果也为准确),叠加马尔可夫链预测准确率的平均值为 69.4%。2016—2018 年各年度的预测准确率分别为:75.0%、75.0%和 58.3%。四季的预测准确率:春季(3—5 月)66.7%、夏季(6—8 月)66.7%、秋季(9—11 月)88.9%、冬季(12—2 月)55.6%。由此可知,秋季预测效果较好,春季和夏季次之,冬季最差。

表 2 2016—2018 年南江站各月日照时数状态预测与实际状态对比表

月份	2016 年		2017 年		2018 年	
	预测值	实际值	预测值	实际值	预测值	实际值
1	E_2	E_3	E_3	E_2	E_3	E_3
2	E_3	E_5	E_2	E_2	E_2	E_3
3	E_3	E_3	E_2	E_2	E_4	E_5
4	E_2	E_2	E_2	E_3	E_2	E_5
5	E_4	E_3	E_3	E_3	E_3	E_2
6	E_2	E_5	E_3	E_3	E_3	E_3
7	E_3	E_3	E_3	E_4	E_3	E_3
8	E_4	E_4	E_4	E_3	E_4	E_5
9	E_3	E_3	E_3	E_2	E_3	E_2
10	E_3	E_3	E_3	E_1	E_3	E_4
11	E_3	E_2	E_3	E_3	E_3	E_2
12	E_3	E_3	E_3	E_5	E_3	E_1

4 两种质控方法在自动观测日照时数上的应用

4.1 统计分析方法在自动观测日照时数质控上的应用

根据以上分析,将 1971—2018 年的人工观测极大值作为数据的可疑值,将 $\mu+3\sigma$ 作为错误值。对巴中地区 4 个基本站 2019—2022 年的各月自动观测日照时数进行分析。

从图 2 可知:在分析的 48 个月的月日照时数

中,4 个基本站均出现了超过可疑值和错误值的数据,且主要分布在 2020 年 2 月—2021 年 5 月期间,而其中超过错误值的情况:南江站有 7 个月(2020 年 2—6 月,2021 年 1—2 月),巴中站有 11 个月(2020 年 3—6 月,9、11 月;2021 年 1—3 月,5—6 月),通江站有 9 个月(2020 年 2—6 月,9 月;2021 年 1—3 月),平昌站有 9 个月(2020 年 1—3 月,5—6 月;2021 年 1—2 月,5—6 月)。在超过错误值的月份中冬春季(12 月—次年 5 月)月份的占比分别为 85.7%(南江)、63.6%(巴中)、77.8%(通江)和 77.8%(平昌)。在设备更换后(南江、通江两站于 2021 年 5 月 25 日更换,巴中、平昌两站于 2021 年 8 月 17 日更换),各站的数据均未超过可疑值。

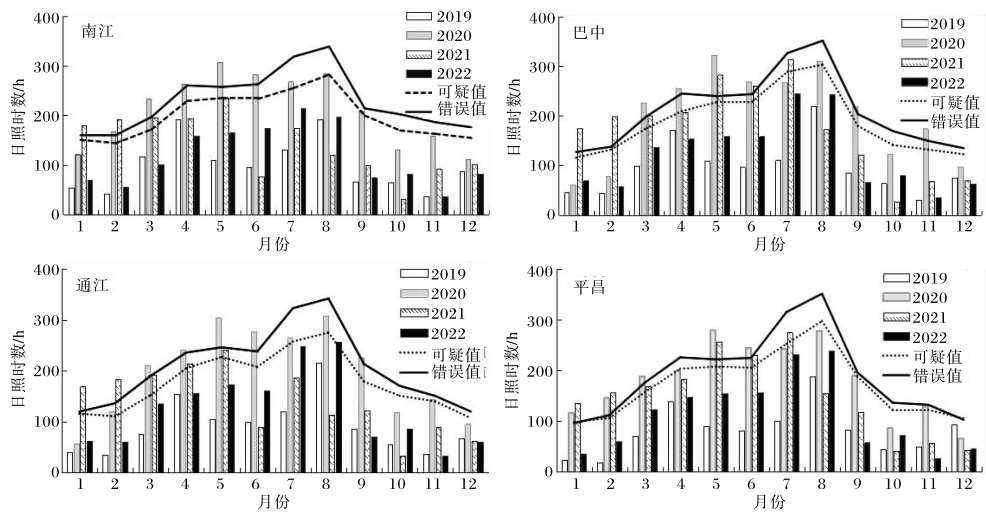

图 2　2019—2022 年巴中各基本站自动观测月日照时数特征分布图

综上分析可知,各台站在 2020—2021 年期间的各月日照数据出现了明显异常,结合日照传感器的安装时间(2018 年 11 月)、正式业务运行时间(2019 年 7 月)和第一次更换时间(2021 年 5 月和 8 月)分析,导致该时段数据失真的可能原因:仪器日常维护不到位、传感器未及时标校等[20-23]。如果日照时数连续两个月以上超出预测值或气候异常值时,业务人员应加强日照传感器日常维护,必要时可对设备做更换处理。

4.2　马尔可夫链预测自动日照的应用

利用叠加马尔可夫链对 2019—2022 年各月日照时数的状态情况进行预测。通过表 3 分析,2019—2022 年的各年度预测准确率分别为 75.0%、16.7%、50.0% 和 83.3%。2019 年和 2022 年的预测效果较 2016—2018 年人工观测数据预测的平均效果更好,但 2020 年和 2021 年的预测效果较差。从季节来看,2019、2021 和 2022 年的夏季和秋季预测准确率均为 100%,2020 年的夏季预测准确率为 0%,秋季为 66.7%。春季和冬季各月的预测准确率分别为 41.7%、16.7%,预测效果较差;排除 2020 年的预测状态,冬季和春季预测准确率也仅为 55.6%、22.2%。结合前文分析,2020 年各月的日照数据可能存在严重失真,同时也说明利用马尔可夫链预测夏季和秋季各月的日照时数状态进行数据质控是可行的。统计分析方法在冬季和春季识别异常数据效果优于马尔可夫链预测。

5　结论

(1)巴中地区月日照分布特征符合正态分布,各月的人工观测极值与气候异常值 μ(月平均值)$+2\sigma$(标准差)高度吻合,可将气候异常值 $\mu+2\sigma$ 作为质控的可疑值,统计异常值 $\mu+3\sigma$ 作为质控的错误值。

(2)马尔可夫链在夏季和秋季对自动观测日照时数的状态预测效果较好,对冬季和春季的预测状态效果较差。而统计分析方法恰好相反,其在冬季和春季识别异常数据比马尔可夫链预测效果更好。故将统计分析和马尔可夫链预测相结合

表 3 2019—2022 年南江站各月日照时数状态预测与实际状态对比表

月份	2019 年		2020 年		2021 年		2022 年	
	预测值	实际值	预测值	实际值	预测值	实际值	预测值	实际值
1	E_3	E_2	E_3	E_4	E_3	E_5	E_3	E_2
2	E_1	E_2	E_2	E_5	E_2	E_5	E_2	E_2
3	E_2	E_3	E_3	E_5	E_3	E_5	E_2	E_3
4	E_3	E_4	E_2	E_5	E_2	E_4	E_2	E_3
5	E_3	E_1	E_3	E_5	E_3	E_5	E_3	E_3
6	E_2	E_1	E_2	E_5	E_2	E_1	E_3	E_3
7	E_3	E_2	E_4	E_5	E_4	E_3	E_4	E_4
8	E_4	E_3	E_4	E_5	E_4	E_2	E_3	E_3
9	E_3	E_2	E_3	E_5	E_4	E_2	E_3	E_2
10	E_3	E_2	E_4	E_4	E_3	E_1	E_3	E_3
11	E_2	E_1	E_5	E_5	E_3	E_3	E_4	E_1
12	E_3	E_3	E_3	E_4	E_3	E_4	E_3	E_3
年预测准确率	75.0%		16.7%		50.0%		83.3%	

使用，可大大提高自动观测日照时数质控的准确率；即在冬季和春季，将 $\mu+2\sigma$ 作为可疑值，$\mu+3\sigma$ 作为错误值；在夏季和秋季，利用马尔可夫链预测状态的区间上限值作为可疑值，将 $\mu+3\sigma$ 作为错误值。

(3)当日照时数连续两个月以上超出预测值或气候异常值时，业务人员需关注日照传感器的工作状态，并加强日常维护，必要时需更换设备。

参考文献：

[1] 李进虎，韩辉福，徐泽东．青海省地面气象观测自动化改革工作思考[J]．青海气象，2020(2)：44-47.

[2] 杨小梅，安文玲，张薇，等．中国西南地区日照时数变化及影响因素[J]．兰州大学学报(自然科学版)，2012，48(5)：52-60.

[3] 李金建，秦宁生，孙善磊，等．基于均一性检验的1961年至2006年四川省日照变化规律研究[J]．资源科学，2011，33(5)：1002-1009.

[4] 胡慧敏，吴薇，杜冰．四川省1969—2018年日照时数变化规律及未来趋势分析[J]．成都信息工程大学学报，2023，38(5)：610-614.

[5] 苗运玲，张林梅，卓世新．哈密绿洲近55年日照和风速变化特征[J]．陕西气象，2017(1)：14-19.

[6] 黄肖寒，杨睿，贺春江．近54年罗城日照变化特征及其影响因子分析[J]．陕西气象，2014(1)：13-17.

[7] 迟庆红，陆桂荣，吴炫，等．DFC2型光电式数字日照计与暗筒式日照计对比分析[J]．气象水文海洋仪器，2023，40(2)：25-28.

[8] 黄维，董德保，沈玉亮．日照传感器与人工观测数据差异分析[J]．气象水文海洋仪器，2022，39(4)：46-50.

[9] 杨炳玉，舒康宁，陈焘．光电式数字日照计在气象观测中的应用评估[J]．气象水文海洋仪器，2022，39(3)：80-82.

[10] 曲静．基于马尔科夫链的西安春季首场透雨预测方法研究[J]．安徽农业科学，2011，39(24)：14938-14939，15031.

[11] 海涛，闻科伟，周玲，等．基于气象相似度与马尔科夫链的光伏发电预测方法[J]．广西大学学报(自然科学版)，2015，40(6)：1452-1460.

[12] 刘方，胡彩虹，何鹏飞．基于数理统计方法的降水量预测模型建立及应用[J]．气象与环境科学，2014，37(2)：89-93.

[13] 薛薇．统计分析与SPSS的应用[M]．6版．北京：中国人民大学出版社，2015：108-115.

[14] 李湘阁，胡凝，黄红丽，等．实用气象统计方法[M]．北京：气象出版社，2015：32-33.

[15] 黄嘉佑．气象统计分析与预报方法[M]．北京：气

象出版社,2015:247-251.

[16] 江志红,常奋华,丁裕国. 基于马尔科夫链转移概率极限分布的降水过程持续性研究[J]. 气象学报,2013,71(2):286-294.

[17] 刘方,胡彩虹,何鹏飞. 基于数理统计方法的降水量预测模型建立及应用[J]. 气象与环境科学,2014,37(2):89-93.

[18] 气候术语:DB51/T 582—2013[S].

[19] 沙定国. 误差分析与测量不确定度评定[M]. 北京:中国计量出版社,1989:66-68.

[20] 崔晓霞. DFC2 光电式数字日照计数据偏差问题研究[J]. 气象水文海洋仪器,2022,39(1):74-76.

[21] 汪清梅,蒋文轩,徐重晔,等. 数字式日照传感器维护及数据质量控制方法探讨[J]. 气象水文海洋仪器,2021,38(2):54-56.

[22] 肖路,金之川,袁乙木,等. DFC2 光电式数字日照计的安装及维护[J]. 气象水文海洋仪器,2020,37(4):115-117.

[23] 张红娟,曾英,妙娟利,等. 陕西省日照平行观测资料评估[J]. 陕西气象,2020(2):52-54.