任芝花,赵平,张强,等. 适用于全国自动站小时降水资料的质量控制方法[J]. 气象,2010,36(7):123-132.

适用于全国自动站小时降水资料的质量控制方法*

任芝花 赵 平 张 强 张志富 曹丽娟
杨燕茹 邹凤玲 赵煜飞 赵慧敏 陈 哲
国家气象信息中心,北京 100081

提 要: 区域自动站逐小时降水资料对于气象预警、决策服务、预报验证等非常重要。资料在提供使用前其质量状况应是可知的。在制作的区域自动站逐小时降水资料数据集基础上,并结合国家级台站观测的小时降水资料,通过深入分析错误数据的存在方式,研制形成全国自动站逐小时降水资料质量控制方案。该方案适用于全国范围内区域站和国家站,实时和非实时上传的逐小时降水资料自动质量控制。最后介绍了利用该方案对 2006—2009 年实时上传的全国自动站小时降水资料的质量评估结果。

关键词: 质量控制,质量评估,小时降水数据,区域自动站,国家级自动站

Quality Control Procedures for Hourly Precipitation Data from Automatic Weather Stations in China

REN Zhihua ZHAO Ping ZHANG Qiang ZHANG Zhifu CAO Lijuan
YANG Yanru ZOU Fengling ZHAO Yufei ZHAO Huimin CHEN Zhe
National Meteorological Information Center, Beijing 100081

Abstract: Hourly precipitation data of regional automatic weather stations (AWS) is very vital to meteorological operation in disaster alert, decision-making service, forecast validation, etc. Quality status of any data should be knowable before being used. Based on all the hourly precipitation data of regional AWS and national AWS in China from 2002 to 2009 and the characters of its error data, quality control procedures have been developed for real time and non-real time hourly precipitation data from all of above AWS. In the end, it is presented the quality assessment results on real time hourly precipitation data of regional AWS and national AWS in China from 2006 to 2009 using the above quality control procedures.

Key words: quality control, quality assessment, hourly precipitation data, regional AWS, national AWS

引 言

2009 年底,气象部门在全国已建国家级自动站(简称国家站,下文同)及区域自动站(简称区域站,下文同)30000 左右,其中 92%的站为无人值守的区域站。区域站观测资料从 2006 年开始逐步实时上传,其对于气象预警、决策服务、预报验证等非常重要。众多台站观测资料的质量不可能都达到正确无误的程度,但是若其质量状况是已知的和可证实的,那么资料就可得到恰当的应用[1]。为此,首先要有一系列比较有效的质量控制方法。国内外有关地面气象资料的质量控制方法主要有气候学界限值检查、台站或区域极值检查、要素间内部一致性检查、时间一致性检查以及空间一致性检查[2-17]。一直以来,国际上针对地面日或月气候统计资料质量控制方法的研究及其应用相对比较广泛[4-8],而针对小时观测资料尤其降水资料质量控制方法的研究比较

* 中国气象局基建项目"全国自动站实时资料质量控制与综合评估系统建设"、科技部项目"气象数据共享中心 01 课题—气象科学数据质量保障体系与质量控制技术研究"资助
2009 年 11 月 23 日收稿, 2010 年 1 月 15 日收修定稿
第一作者:任芝花,主要从事气象资料质量控制与评估方面的工作. Email:rzh@cma.gov.cn

有限。目前,应用到的逐小时降水资料质量控制方法主要为气候学界限值检查,上述其他检查方法的研究与应用基本尚未开展[10-16]。

国家气象信息中心气象资料室计划分要素分批研制可用于气象数据传输、数据应用服务实时业务的全国各种自动站小时观测资料质量控制方案。为此,在制作的区域自动站逐小时降水资料数据集(2006 年 1 月—2009 年 9 月)基础上,结合国家站观测的小时降水资料,2009 年 7 月首先研制了通过专家论证的 1.0 版全国自动站逐小时降水资料质量控制方案,同月该方案用于国家级气象数据服务实时业务。通过业务运行,发现 1.0 版方案控制后的可疑数据多数为正确资料,由此在原方案基础上,通过深入数据分析,目前完善形成了 1.1 版的全国自动站逐小时降水资料质量控制方案。完善后的方案适用于全国范围内区域站和国家站,实时和非实时上传的逐小时降水资料自动质量控制。下文主要介绍 1.1 版全国自动站逐小时降水资料质量控制方案的研制,及应用该方案对 2006 年 1 月—2009 年 9 月实时上传的全国各种自动站小时降水资料(考虑到我国北方和西部,自动站雨量计冬季停止观测,因此该资料包括 5—9 月全国所有站上传的数据,而其他月份仅为 32°N 以南,104°E 以东的南方台站观测的数据)的质量控制与评估结果。本文讨论的降水仅指液态降水量。

1 质量控制方案的研制

气象资料质量控制技术发展至今,从理论上讲是比较成熟的,但是在实际应用中却往往达不到理想的效果。原因在于气象数据本身受观测仪器、观测环境、数据处理技术等方面的限制,与气象状态并非完全吻合。另外,部分错误数据与代表极端异常天气气候事件的数据表现形式类同。因此质量控制方案的研制必须深入数据内部,了解数据本身的规律以及各种原因引起的错误数据的表现形式。全国自动站小时降水资料质量控制方案的研制是基于上述认识,基于全国气象部门区域自动站及国家级自动站运行后上传的所有小时降水资料开展的。

1.1 数据质量控制码的规定

在对数据进行质量控制(QC)的过程中,随着控制进程的进行,需要不断的对被检数据设置或修改 QC 码。QC 码的规定[17]如下:

数据正确: QC 码=0
数据可疑: QC 码=1
数据错误: QC 码=2
无观测数据: QC 码=8
数据未作质量控制: QC 码=9

1.2 质量控制方法及质量控制码的设置

首先设置数据的初始 QC 码=9,当无观测数据时 QC 码=8。然后按下列先后顺序进行检查。每步检查均针对 QC 码=9 的数据进行,并根据检查结果决定是否修改 QC 码。

1.2.1 界限值检查

界限值检查包括气候学界限值检查和区域界限值检查。

(1) 气候学界限值检查

超越气候学界限值范围的数据为错误,QC 码=2。

气候学界限值范围为:0~150 mm/h

(2) 区域界限值检查

根据经纬度和降水量空间分布,全国划分 6 个区域。对各区域分别制定界限值范围,范围的下限为 0 mm,上限为图 1 所示数值。超越该界限值范围的数据为错误,QC 码=2。

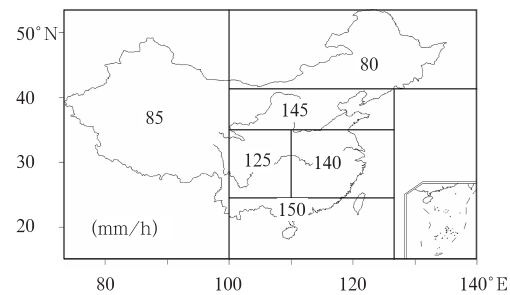

图 1 各区域小时降水量界限值

Fig. 1 Maximum allowable hourly precipitation amount in various regions

从上述 2 种界限值范围的设置可见,区域界限值检查实际上包含了气候学界限值检查。之所以仍保留了气候学界限值检查,主要是从质量控制标准[17]以及方案今后调整的角度考虑。

界限值范围的上限参数设置是否合理直接决定了界限值检查的效果。设置太低,则个别极端大的正确数据可能会被作为错误数据处理,而这部分数据对于气象预警、极端天气研究非常关键;设置太高,则错

误数据又可能被当作正确数据处理,这种大的错误数据在应用中起明显的误导作用。通过统计可得,界限值检查出的错误数据若用后续的空间一致性检查(简称空检,下文同)方法筛查,81.2%的数据可被检出,18.6%的数据因无邻近站而无法进行空检,剩下的0.2%错误数据因其邻近站为同样级别的错误数据也无法检出。因此,18.8%的明显错误数据无法通过空间一致性检查被检出,这就要求这类错误数据应尽可能在界限值检查中得到有效的控制。由此,该方案中的上限参数设置相对比较严格。

在本方案中,气候界限值、区域界限值范围的上限参数均是根据自动气象站运行后实时上传的每小时降水资料和月报文件中的小时降水资料来确定的,资料时间为 2002 年 1 月—2009 年 9 月。因此,2002 年 1 月—2009 年 9 月国家站和区域站测量的小时降水量若超越本方案设置的界限值范围则确定为错误数据。由于超越上述 2 种范围的数据并不多,全国所有站每个时次平均不足 1 个,因此建议实时质量控制业务中,输出超越上述 2 种范围的数据进行查证。若数据为正确,则适当修改上限参数,若数据错误也可及时通知台站。

界限值检出的错误数据均为虚假的降水信号,主要是由于台站雨量计故障、降水感应信号遭受干扰、数据表达方式及传输中的不规范引起,普遍以大值形式表现,个别数据表现为负值。我们知道,国家级自动站仪器由中国气象局统一规划、安装,而各省(市、区)区域自动站仪器选型、安装则各省自定。不同厂家对于仪器故障、仪器不观测等的数据表达方式可能不同,导致界限值检查出的错误数据表达形式有其地域性。比如:国家站中,95%的超界限范围的错误数据为 160 mm;F3901 站(26°30′25″N、120°06′18″E)2006 年 3 月—2007 年 5 月 3400 多个时次错误降水量为 363 mm 左右。

2006 年 1 月—2009 年 9 月实时上传到国家气象信息中心的全国所有自动站小时降水数据为 262 兆。其中,未通过界限值检查的错误数据有 15663 个,占被检数据的 0.06‰。上述错误数据中,99%的数据超越气候学界限值范围,仅 1%的数据为区域界限值检查所得。另外,界限值检出的错误数据中,27.4%的数据属于国家站,72.6%属于区域站。在 2006 年 1 月—2009 年 9 月间,有 596 个国家站、347 个区域站小时降水数据不同程度地发生超界限值现象,其中有 7 个站的错误数据量在 350～3500 个之间,占总错误数据的 46%。

图 2 给出了 2006 年 1 月—2009 年 9 月全国自动站实时上传的小时降水数据量及其界限值检查出的各月错误数据情况。从图中可见,实时上传的数据量呈明显的上升趋势。汛期(5—9 月)平均每月的数据量从 2006 年的 0.5 兆增加到 2009 年的14.3兆,非汛期数据量也相应稳步上升。界限值检查的错误数据相对于数据量来讲呈降低趋势,尤其自 2007 年 5 月开始,检出的数据错误率稳定走低,平均为 0.04‰,而在此之前的数据错误率平均为0.4‰。

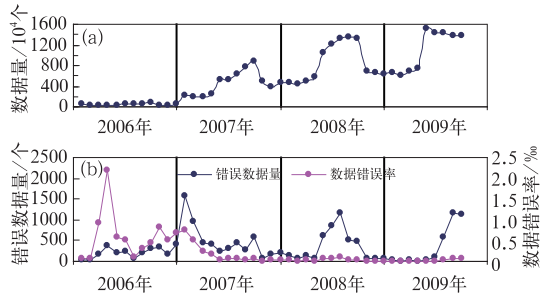

图 2　全国自动站历年各月实时上传的小时降水数据量(a)以及界限值检出的错误数据情况(b)

Fig. 2　Monthly quantity of hourly precipitation data in real-time operations (a) and the number of error data detected by climatic range check (b)

1.2.2　时间一致性检查——数据连续无变化检查

与气温要素不同,降水的时间非连续性很强,因而降水数据在进行了上述界限值检查后,再进行数据间的时间变率检查,则无实际作用。而在实时上传的小时降水数据中,有些错误数据以连续无变化的形式出现。如:53764 站(37°30′N、111°06′E)由于雨量计故障,2008 年 9 月 9 日 17 时至 10 日 9 时连续 17 个时次小时降水数据均为 0.1 mm;图 3a 中 N1009 站(23°12′18″N、108°09′48″E)2009 年 9 月份全月收到的 699 个小时降水数据中,692 个数据均为 0.1 mm;图 3b 中 H5004 站(46°40′59″N、131°10′56″E)则连续多日、多个时次上传的数据分别为 6.0 mm 或 2.8 mm。经过核实,上述个例中连续无变化的数据均为错误数据。造成该类数据错误的主要原因为:①测量仪器故障;②雨量计漏斗部分堵塞,承水器收集的降水以匀速渗漏的方式进入翻斗计量;③报文上传重复。

基于上述分析,本方案考虑了连续多个时次降水量数据无变化检查。但是在实际观测中,绵绵细雨式的降水数据也会以连续无变化的形式出现,这

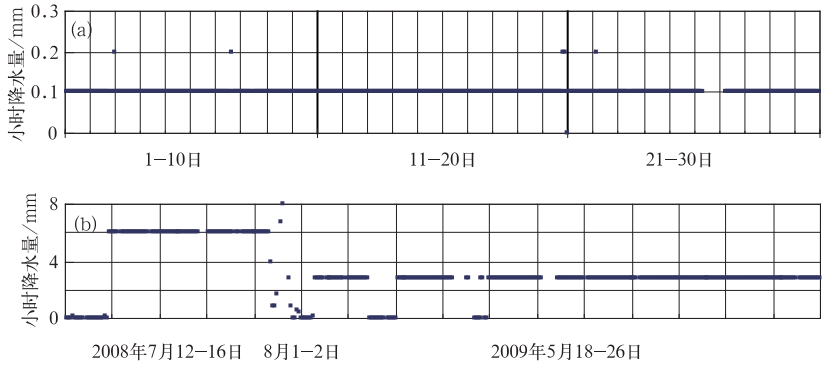

图 3 (a) N1009 站 2009 年 9 月小时降水量;(b) H5004 站不同时间段小时降水量

Fig. 3 Hourly precipitation data of Station N1009 in Sep, 2009 (a) and Station H5004 at different periods (b)

就要求给出合理的数据判错原则。基于 2002 年 1 月—2009 年 9 月全国自动气象站实时上传的每小时降水资料和月报文件中的小时降水资料,通过逐一分析该资料中连续 6 个及以上时次降水量(大于 0 mm) 无变化数据正确与否,总结出了下列 3 类实时数据判错原则:从当前时次开始前推,

(1) 连续 N1 个及以上小时降水量(简称 R,下同)相等且 $R\in(0,0.5)$ mm 时,则相应数据均判为错误,QC 码=2;

(2) 连续 N2 个及以上小时降水量相等且 $R\in[0.5,1.0)$ mm 时,则相应数据均判为错误,QC 码=2;

(3) 连续 N3 个及以上小时降水量相等且 $R\geq 1.0$ mm 时,则相应数据均判为错误,QC 码=2。

其中 N1>N2>N3。

连续无变化检查发现的错误数据介于 0.1~52.5 mm 之间,其中,44%的数据为 0.1 mm。由于从 2007 年 1 月开始,才陆续有站逐小时连续上报降水资料,因此上述时间一致性检查仅对 2007 年以来上报的数据有效。2007 年 1 月—2009 年 9 月实时上报到国家气象信息中心的全国所有自动站小时降水数据中,对通过了界限值检查后的数据再进行上述时间一致性检查,又发现 32300 个数据错误,为被检数据的 0.12‰。上述错误数据中,有 32 个数据属于 2 个国家级站,其余属于 574 个区域站。

图 4 给出了 2007 年 1 月—2009 年 9 月全国自动站实时上传的小时降水数据中,时间一致性检查出的各月错误数据情况。图中,2009 年 8 月检出的错误数据量异常偏多,这与 8 月中上旬部分站不同程度地发重复报有关。这些站的重复报有如下特点:对于同一个站来讲,每个时次重复报在 25 次左右,且每次报降水量不同,但令人意外的是不同时次的第 1 份报数据均相同、不同时次的第 2 份报数据也均相同、……。这导致无论选用哪一次报文,数据均不可用。但作为实时库来讲,后面的重复报若无更正标志的话(事实上目前的重复报普遍无更正标志),每个时次均采用第 1 份报的数据。

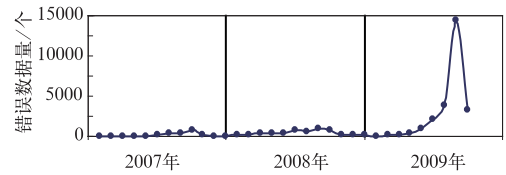

图 4 全国自动站实时上传的小时降水数据中时间一致性检查出的各月错误数据量

Fig. 4 Monthly quantity of error hourly precipitation data detected in real-time operations by temporal consistency check

连续无变化检查发现的错误数据,其降水量普遍较小。除非周围站有很大的降水过程,否则很难通过后续的空间一致性检查被检出,因此该类时间一致性检查过程也至关重要。

1.2.3 空间一致性检查

1.2.3.1 邻近站的确定

(1) 台站分布密度

区域站建设的主要目的是为了地方预警、预报实况验证和灾害评估,更多的是为了灾害预警服务。因此,区域站的空间分布主要从人口密度、面雨量以及行政位置等角度考虑,其主要安装在乡镇、邻近水域以及风景区等位置,这就导致区域站空间分布不可能均匀。在人烟稀疏的区域很少会考虑区域站建

设,而人口密度较高的区域台站密度则会加大,这也是图 5 显示的全国区域站分布表现为东部'扎堆'现象 ,即我国人口密度高的东部区域台站密度大,而人口密度低的西部以及东北三省部分区域台站密度小。另外,从图 5 可见,在台站密度较低的西部和东北,区域站建设也表现为'扎堆'现象,这同样与当地的人口密度有关。因此西部和东北区域站的建设可改善部分区域台站稀疏状况,而黑龙江北部、四川西部、西藏、新疆、青海、甘肃、内蒙古等大部分地区,区域站目前的建设不足以改变原有的台站稀疏状况。

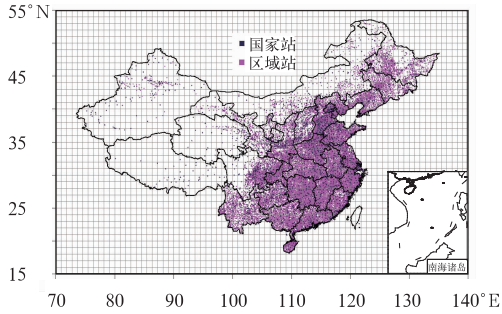

图 5　全国自动站分布

Fig. 5　Distribution of AWS in China

空间一致性检查的基本条件是台站密度达到一定的程度。当台站密度低到某种程度时,则不适合进行空检,因此,对于台站稀疏的区域不应强求进行该检查。本方案空间一致性检查主要针对我国东部区域以及西部、东北台站比较密集的区域,当被检台站周围方圆 50 km 范围内无邻近站时,则规定不再进行空间一致性检查。

(2) 邻近站分布密度

空间一致性检查需要与邻近站的数据比较,因此了解被检站的邻近站分布密度是空间一致性检查的前提。全国区域站所属的邻近站分布密度见图 6。当邻近站包括区域站和国家站时,距离被检站 20 km 范围内,除了内蒙古和新疆邻近站平均为 1 个外,其他省(市、区)邻近站均在 2 个以上,全国平均为每个站有 11 个邻近站;距离被检站 30 km 范围内,除了内蒙古、西藏和新疆邻近站平均为 3 个外,其他省(市、区)邻近站均在 5 个以上,全国平均为每个站有 25 个邻近站;距离被检站 50 km 范围内,除了西藏邻近站平均为 3 个外,其他省(市、区)邻近站均在 7 个以上,全国平均为每个站有 69 个邻近站。当邻近站只包括国家站时,距离被检站 50 km 范围内,除了内蒙古、西藏邻近站平均为 1 个外,其他地区邻近站均在 2 个以上,全国平均为每个站有 4 个邻近站。

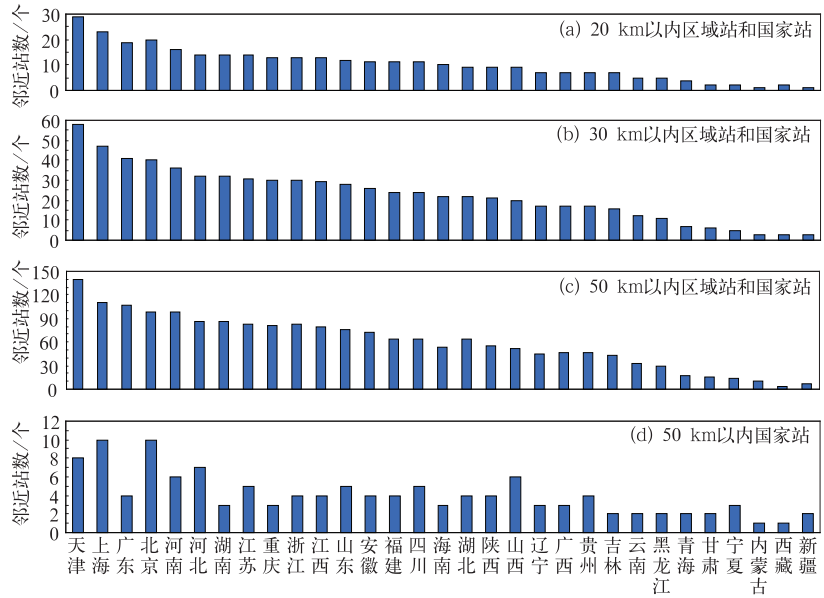

图 6　邻近站分布密度

Fig. 6　Distribution density of neighboring AWS

另外,有 5.6‰的站在其 50 km 范围内无任何邻近站,这些站中,国家站和区域站分别占 72%、28%。全国 18.8%的站 10 km 范围内、2.5%的站 20 km 范围内、1.2%的站 30 km 范围内不存在邻近站。

(3) 邻近站表的制定

进行空间一致性检查前,首先要确定可能用于与被检站比较的邻近站。为便于质量控制时邻近站的选用,本方案制定了距离被检站 50 km 范围内的

邻近站表，表中的邻近站既包括区域站也包括国家站。考虑到降水的局地性以及空间分布的方向性，邻近站表的制定原则如下：

以被检站为中心点，把距离被检站 50 km 范围内的区域划分成图 7 所示的 4 个区。邻近站表中，每个站的邻近站均按照先方位后距离的原则排列，顺序为：Ⅰ区距离被检站最近的站、Ⅱ区距离被检站最近的站、Ⅲ区距离被检站最近的站、Ⅳ区距离被检站最近的站、Ⅰ区距离被检站次近的站、Ⅱ区距离被检站次近的站、…… 。

进行空间一致性检查时，按照上述排列的先后顺序选用邻近站。

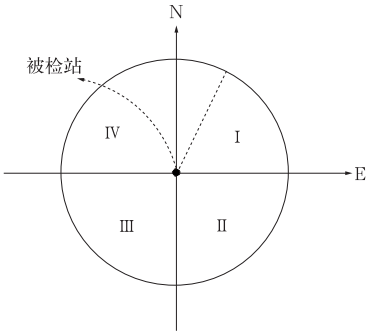

图 7 邻近站选择示意图
Fig. 7 Sketch chart of neighboring AWS choice

1.2.3.2 空间一致性检查方法

经过上述界限值检查和时间一致性检查之后，再对质量控制码仍为‘9’的数据进行空间一致性检查。界限值检查和时间一致性检查均为数据的错误性判断，对未通过检查的数据，其质量控制码直接改为错误码‘2’。而空间一致性检查则与上述不同，它包括数据的正确性判断和可疑性判断，因此经过空间一致性检查后，数据的质量控制码可能改为正确码‘0’或可疑码‘1’。基于 2002 年 1 月—2009 年 9 月全国自动气象站实时上传的每小时降水资料和月报文件中的小时降水资料，通过反复试验与调试，制定了下列 3 项空间一致性检查方法。

空间一致性检查所用的邻近站须为有效邻近站，这里有效邻近站是指：与被检站比较时，邻近站数据的 QC 码≠2 和 8。

空检① 对可能错误的连续性强降水资料进行筛查

从当前时次开始前推，连续 4 个时次降水量均在 10 mm 及以上时，进行累计降水量统计。当该站 4 个时次累计量与邻近站表中排在前面的 5 个邻近站 4～5 个连续时次累计降水量比较，均满足下列条件时，认为被检站各时次数据可疑，QC 码＝1。

被检站累计降水量＞2.5×邻近站累计降水量

空检② 当被检站降水量 R＞5 mm 且 QC 码＝9 时，与邻近的国家级站数据比较

被检站数据与邻近站表中排在前面的 5 个以内的国家站数据比较，当被检站数据与任何一个邻近的国家站数据存在如下关系时，停止比较，认为被检站数据及与其比较的国家站数据均正确，QC码＝0。

0.8×邻近站降水量≤被检站降水量≤1.2×邻近站降水量

空检③ 当被检站数据 QC 码＝9 时，继续与邻近站数据比较

在邻近站表中按前后顺序依次选择 N 个邻近站，由上述 N 个邻近站当前时次及上个时次的降水量值形成数据序列，计算序列的最大值(VALUE_MAX)和最小值(VALUE_MIN)。其中，N＝16 或 N＝NMAX（NMAX＞0，NMAX 指邻近站不足 16 个时，该站所有的邻近站数量）。

当被检站小时降水量 R 与 VALUE_MAX、VALUE_MIN 比较，若符合下列判断原则，则数据正确，QC 码＝0；若不符合下列原则且 N≥3，则数据可疑，QC 码＝1：

a1×VALUE_MAX≥R≥a2×VALUE_MIN (a1＞1，a2＜1)

在该检查中，把降水量划分成下列 5 段进行检查，对于不同的降水量段，a1 及 a2 的取值不同：

R＝0 mm、0 mm＜R≤5 mm、5 mm＜R≤10 mm、10 mm＜R≤50 mm、R＞50 mm

至此，质量控制结束，数据的质量控制码有 5 种：0、1、2、8、9。此时，质量控制码在本方案中的含义更加明确，具体如表 1 所示。

表 1 本方案中质量控制码的最终含义
Table 1 Meanings of quality control flags after above checks in the project

质量控制码	含义
0	数据通过了界限值检查、时间一致性检查和空间一致性检查。
1	数据通过了界限值检查和时间一致性检查，但未通过空间一致性检查。
2	数据未通过界限值检查或未通过时间一致性检查。
8	数据缺测或数据未上传。
9	数据通过了界限值检查和时间一致性检查，但由于不存在有效邻近站而未进行空间一致性检查，或其虽有 1～2 个有效邻近站，却仍未通过空间一致性检查。

表 1 中,质量控制码'9'也表示数据质量状况不确定。这样的状况 99.7% 是因为方圆 50 km 内不存在有效邻近站而无法进行空检造成的。用上述方案对 2006 年 1 月—2009 年 9 月实时上传的全国自动站小时降水数据进行质量控制后,质量控制码为'9'的数据中,20%属于图 6 中邻近站相对较少的新疆、西藏、青海、甘肃、宁夏和内蒙古,80%属于其余的省(市、区)。造成有效邻近站不存在的主要原因有①台站稀疏;②台站数据上传不完整。

在空间一致性检查中,起主要作用的为空检③,它包容了各降水量段数据的检查。在进行该项检查时,借助了 1～16 个邻近站的观测数据进行比较。之所以选用 16 个邻近站极限,主要是为了尽可能保证选用的邻近站在空间分布上涉及到各个方位,又至少包含 4 个最近的邻近站。由表 2 可见,2006 年 1 月—2009 年 9 月通过空检③检查的全国自动站实时上传的小时降水数据中,99.685%的数据由选用的第一个邻近站检查通过。而未通过检查的数据中,79.469% 的数据与选用的 16 个邻近站比较才确定为可疑。由表 2 进一步可知,对于数据的正确性判断,99.884%的数据由最近的 1～2 个邻近站比较即可完成;而对于数据的可疑性判断,则要求比较严格,需要与更多的站比较,在邻近站数量充足的情况下,需与周围 16 个邻近站比较才判为可疑。

表 2　空检③中所用的邻近站数与通过及未通过检查的数据比关系

Table 2　Relation between number of neighboring AWS in use and pass or no-pass check ratio of hourly precipitation data in spatial consistency check ③

邻近站/个	数据通过比[1]/%	数据未通过比[2]/%	数据质量确定比[3]/%
1	99.685	/	99.648
2	99.884	/	99.847
3	99.931	1.413	99.895
4	99.953	3.336	99.917
5	99.965	5.025	99.930
6	99.973	7.053	99.939
7	99.979	8.510	99.951
8	99.984	10.044	99.946
9	99.988	11.518	99.955
10	99.991	12.976	99.959
11	99.993	14.314	99.961
12	99.995	15.753	99.964
13	99.997	17.315	99.966
14	99.998	18.884	99.968
15	99.999	20.531	99.970
16	100	100	100

注[1] 与相应数量的邻近站数据比较,通过空检③检查的数据量占总通过量的比率;注[2] 与相应数量的邻近站数据比较,未通过空检③检查的数据量占总未通过量的比率;注[3] 与相应数量的邻近站数据比较,数据质量得到确定的数据量占总确定量的比率。数据质量确定指原质量控制码为'9'的数据,经过空检③的检查后,质量控制码改为'0'或'1'。

由空检③检查明确质量状况的数据中,99.648%的数据由 1 个邻近站检查确定。由上文可知,全国各省市区方圆 20 km 范围内平均至少有 1 个邻近站,由此可以认为,上述 99.648%的数据是由 20 km 以内的邻近站确定其质量状况的。

2006 年 1 月—2009 年 9 月全国自动站实时上传的小时降水数据中,空检①共检测到 1252 个可疑数据,占 10 mm 以上数据的 1.7‰。检测的汛期与非汛期数据量相当,但是如图 8a 所示,检测的非汛期 10 mm 以上数据可疑率远高于汛期。空检①检测的可疑数据中,97%属于区域站,82%同样也可被空间③检为可疑。

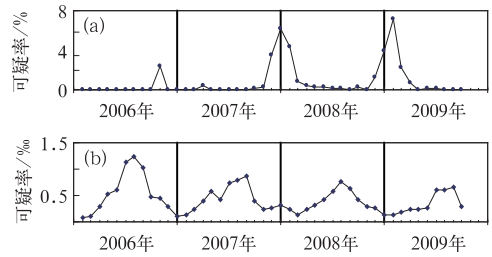

图 8　(a) 空检①检测的 10 mm 以上小时降水数据可疑率;(b) 参入空检③检测的小时降水数据可疑率

Fig. 8　Suspected data ratio of hourly precipitation above 10 mm detected by spatial consistency check ①(a), and that of hourly precipitation by spatial consistency check ③(b)

参入空检②检测的 5 mm 以上的数据中,31%的数据通过该检查。

参入空检③检测的数据中,98.641%的数据通过检测,0.44‰的数据未通过检测,1.315%的数据普遍由于无有效邻近站而质量状况不确定。参入空检③检测的小时降水数据可疑率如图 8b 所示,检测的汛期数据可疑率明显高于非汛期。

2　小时降水数据质量评估结果

2.1　不同途径上传的台站小时降水数据的差异

目前,国家级自动站观测的小时降水数据主要通过下列 2 种文件形式上传:①以 Z 文件[18] 的形式

实时上传；②以月报 A 文件[18]的形式滞后 2 个月上传 1 次。Z 文件中的观测数据由台站实时观测实时自动上传，上传到国家气象信息中心前未经任何的质量控制。而 A 文件中的数据则是在观测过程中经过了台站数据维护、与其他测量仪器观测值比较的基础上形成的。全月的观测数据进入 A 文件后，又经过了台站、省级以及国家级资料部门的三级质量控制[11]，数据的质量明显优于 Z 文件。

全国 2400 多国家站上传的 2006 年 1 月—2009 年 7 月 Z 文件和 A 文件中，在两文件同时次小时降水数据均存在的情况下，7.8‰的数据不同。因此，若以 A 文件中的观测数据为标准的话，可以认为 Z 文件中小时降水数据的错误率为 7.8‰。而利用上文介绍的方案进行质量控制，发现 2006 年 1 月—2009 年 7 月来自 Z 文件中的国家站小时降水数据疑误率却只有 0.55‰。原因分析如下：

7.8‰的不同数据中，其差异分布见表 3。由表可见，65.4%的差异为 0.1 mm，93.4%的差异在 3 mm 以内，6.6%的差异在 3 mm 以上，甚至有 1.2%的差异在 100 mm 以上。当观测值与实际降水量相差 0.1 mm，甚至几个毫米时，这种错误数据除了在观测现场可直接辨认修改外，针对单要素数据的远程质量控制则很难辨别。我们假设观测数据与实际降水量相差 3 mm 以上时，通过质量控制软件可辨别出其错误（实际上很难达到该效果），那么可分辨出的 Z 文件小时降水数据错误率为 0.51‰，这与上述方案控制结果基本一致。

表 3 Z 文件与 A 文件中有差异的数据分布

Table 3 Data distribution in various differences between File Z and File A

数据差/mm	数据量/%
0.1	65.416
0.2～0.5	14.985
0.6～1	5.8541
1.1～2	4.9183
2.1～3	2.257
3.1～4	1.2627
4.1～5	0.8717
5.1～10	1.6242
10.1～20	0.9572
20.1～30	0.3167
30.1～40	0.1319
40.1～50	0.0754
50.1～60	0.0417
60.1～70	0.0311
70.1～80	0.0183
80.1～90	0.0092
90.1～100	0.0137
＞100	1.2158

2.2 实时小时降水数据的质量状况

应用上述质量控制方案对 2006 年 1 月至 2009 年 9 月实时上传的全国自动站（包括国家站和区域站）小时降水资料进行了质量控制，发现质量控制码为‘9’的数据率（质量不确定率）只有 1.312%，而数据的正确率、可疑率、错误率分别为 98.626%、0.44‰、0.18‰。汛期（5—9 月）数据的可疑率、错误率分别为 0.54‰、0.23‰，平均每个时次疑误数据 10 个左右。

我们知道，降水现象相对于无降水来讲是小概率事件。由表 4 可见，实时上传的所有自动站小时降水数据中，仅 10.3%的数据有降水（降水量大于 0 mm），而质量控制主要是针对有降水现象的数据进行的。因此，从有降水现象的数据角度讲，2006 年 1 月—2009 年 9 月实时上传的自动站小时降水数据中可疑率、错误率分别为 4.3‰、1.8‰；汛期数据的可疑率、错误率分别为 5.2‰、2.2‰。

表 4 小时降水资料中各数据段数据量及数据疑误率

Table 4 Data quantity and suspected data ratio in various precipitation ranges

数据段/mm	数据/%	疑误率/%
0	89.7484	0.0030
0.1	2.8592	0.1881
0.2～0.5	2.8989	0.0877
0.6～1	1.3863	0.0413
1.1～2	1.2453	0.0619
2.1～3	0.5835	0.0890
3.1～4	0.3357	0.0791
4.1～5	0.2134	0.0957
5.1～10	0.4405	1.5374
10.1～20	0.2024	10.6607
20.1～30	0.0509	14.5847
30.1～40	0.0174	19.1892
40.1～50	0.0063	23.3367
50.1～60	0.0024	43.8066
60.1～70	0.0011	53.3821
70.1～80	0.0005	61.2613
80.1～90	0.0003	70.2432
90.1～100	0.0002	77.6952
＞100	0.0063	98.9713

另外，由表 3 可见，随着降水值（与实际降水量不完全相同）的增大，数据量快速减少，而数据疑误率则快速升高。这与实际观测中，许多明显的错误数据以大值的形式表现有关。

2.3　国家站与区域站评估结果的区别

2006 年 1 月至 2009 年 9 月小时降水数据中,区域站的正确率、质量不确定率、可疑率、错误率分别为 99.142%、0.796%、0.44‰和 0.18‰,而国家站则分别为 93.837%、6.102%、0.44‰和 0.17‰。前文已经讲过,实时上传的降水数据未经任何质量控制,若仅从数据的疑误率角度考虑,区域站与国家站的数据质量相当。另外,由于国家站的布局除了考虑气象服务外,更多的还要考虑天气气候的代表性和空间分布上的广泛性,而区域站建设主要是为地方预警服务考虑的。因此,国家站在全国范围内的空间分布要比区域站均匀,区域站则多以'扎堆'的形式分布在人口相对密集的地区。无论是国家站还是区域站,其邻近站多为区域站,而上述国家站和区域站的不同分布状况导致国家站中邻近站不足的比例要比区域站高,因此国家站质量不确定,即未进行空间一致性检查的数据比例比区域站明显偏高。

区域站和国家站历年各月数据质量情况见图 9。从图中可见,无论是区域站还是国家站,数据的质量从 2008 年开始得到明显改善。2009 年 7—9 月数据质量略有下降,主要与 1.2.2 节介绍的重复报问题有关。图中,质量不确定率与相应的正确率曲线走势相反。由于冬季资料只来自非结冰区的南方密集台站,而夏季资料则来自全国包括台站分布稀疏的区域,这导致国家站资料夏季无法进行空检的比例比冬季明显偏高,从而引起图中国家站的质量不确定率明显地表现为夏季高,冬季低的特点。

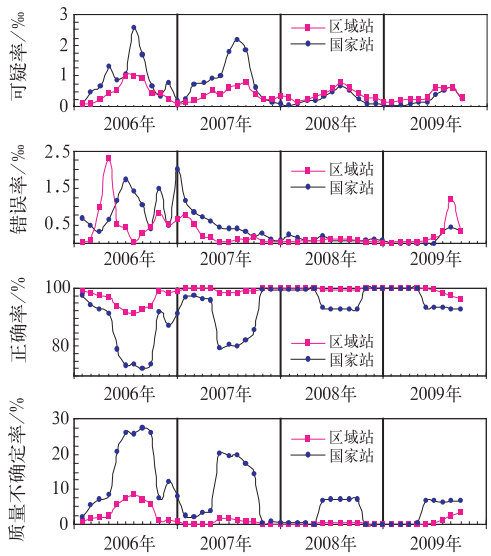

图 9　历年各月自动站小时降水数据质量状况

Fig. 9　Monthly quality of hourly precipitation data

3　结论和讨论

在了解数据本身的规律以及各种原因引起的错误数据表现形式的基础上,针对全国区域自动站和国家级自动站实时上传的逐小时降水资料,首次研制了较为全面、有效的质量控制方案。方案中包括了气候学界限值检查、区域界限值检查、时间一致性检查和空间一致性检查方法。该方案可适用于全国范围内区域站和国家站,实时和非实时上传的逐小时降水资料自动质量控制。省级资料审核部门在应用该方案对省内小时降水资料进行质量控制时,可根据本省历史资料状况,仅对区域界限值控制参数进行适当调整。利用该方案对 2006—2009 年实时上传的全国自动站小时降水资料进行了质量评估。数据的正确率、可疑率、错误率分别为 98.626%、0.44‰、0.18‰。对于降水量大于 0 mm 的数据而言,其可疑率、错误率分别为 4.3‰、1.8‰。检测的区域站和国家站数据的疑误率相当。

提供给用户的数据,其质量状况应尽可能是明确的。这就要求对上述质量控制后表示可疑的数据作进一步的判断,表明数据正确还是错误。作为用户或非实时数据服务业务可考虑结合卫星云图、雷达拼图等人工分析判断。但作为实时质量控制及其后的实时数据服务业务,人工介入分析判断数据的质量状况,显然影响服务的时效性,因此控制过程应尽可能自动化。下一步将考虑利用卫星资料或雷达基数据,定性得到可疑数据所属的时间段台站周围的降雨程度。从而实时、自动地进一步判断可疑降水量数据正确与否。

致谢:气象资料室周自江主任对方案的设计提供了许多宝贵的建议与支持;李泽椿院士等专家组成员以及信息中心高华云研究员对原方案的进一步完善提出了许多合理性建议,在此一并表示感谢。

参考文献

[1]　WMO. Guide to Meteorological Instrument and Methods of Observation[R]. WMO-No. 8,2008.

[2]　WMO. Manual on GDPS[R]. WMO-No. 485,1992,Volume 1:II. 1-4.

[3]　刘小宁,任芝花. 地面气象资料质量控制方法研究概述[J]. 气象科技,2005,33(3):199-203.

[4]　Peterson T C,Vose R S,Schmoyer R, et al. 'Global historical

climatology network (GHCN) quality control of monthly temperature data'[J]. Int J Climatol,1998, 18:1169-1179.

[5] Song Feng, Qi Hu, Weihong Qian. Quality control of daily meteorological data in China,1951-2000:A new dataset[J]. Int J Climatol,2004,24:853-870.

[6] Sciuto G, Bonaccorso B, Cancelliere A, et al. Quality control of daily rainfall data with neural networks[J]. J Hydro, 2009,364:13-22.

[7] 任芝花, 刘小宁, 杨文霞. 极端异常气象资料的综合性质量控制与分析[J]. 气象学报,2005,63(4):526-533.

[8] 任芝花, 熊安元, 邹风玲. 中国地面月气候资料质量控制方法的研究[J]. 应用气象学报,2007,18(4):516-523.

[9] Fillipov V V. Quality Control of Meteorological Data[R]. World Weather Watch Planning Report, WMO-No. 26,1968.

[10] 窦以文, 屈玉贵, 陶士伟,等. 北京自动气象站实时数据质量控制应用[J]. 气象, 2008, 34(8):77-81.

[11] 任芝花, 熊安元. 地面自动站观测资料三级质量控制业务系统的研制[J]. 气象, 2007, 33(1): 19-24.

[12] 王新华,罗四维,刘小宁,等. 国家级地面自动站 A 文件质量控制方法及软件开发[J]. 气象,2006,32(3):56-63.

[13] 任芝花, 许松, 孙化南,等. 全球地面天气报历史资料质量检查与分析[J]. 应用气象学报,2006, 17(4):412-420.

[14] Igor Zahumensk (Slovakia). Guidelines on Quality Control Procedures for Data from Automatic Weather Stations[R]. Expert Team on Requirements for Data from Automatic Weather Stations, Third Session, WMO, 2004.

[15] 熊安元. 北欧气象观测资料的质量控制[J]. 气象科技,2003, 31(5):314-320.

[16] 王海军,杨志彪,杨代才,等. 自动气象站实时资料自动质量控制方法及其应用[J]. 气象, 2007, 33(10): 102-106.

[17] 气象行业标准. 地面气象观测资料质量控制[D],2009.

[18] 中国气象局. 地面气象观测数据文件和记录薄表格式[M]. 北京:气象出版社,2005:18-65.