

文章编号: 1673-8411(2019)02-0090-05

基于三次样条插值的探空气温质量控制研究

潘霄¹, 叶小岭^{1,2}, 熊雄¹, 王佐鹏¹, 陈昕¹

(1. 南京信息工程大学自动化学院, 南京 210044;

2. 南京信息工程大学气象灾害预报预警与评估协同创新中心, 南京 210044)

摘要: 利用上海地区 2016 年 11 月份探空气温观测资料, 把观测资料分为五个高度段, 采用平均绝对误差 (MAE) 和均方根误差 (RMSE) 对气温观测资料进行质量控制分析。对每个高度段气温观测资料进行三次样条插值, 结合交叉检验的思想对探空质量控制进行研究。结果表明: 三次样条插值的质量控制算法在低空天气状况良好的条件下, 对探空气温观测资料的质量控制效果显著, 能更有效地标记出气温观测数据中的可疑值。

关键词: 气温观测资料; 三次样条插值; 交叉检验; 探空质量控制

中图分类号: P412.11

文献标识码: A

Research on Quality Control of Air Temperature Based on Cubic Spline Interpolation

Pan Xiao¹, Ye xiaoling^{1,2}, Xiong xiong¹, Wang zuopeng¹, Chen xin¹

(1. School of Automation, Nanjing University of Information Science and Technology, Nanjing Jiangsu 210044; 2. Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters, Nanjing University of Information Science and Technology, Nanjing Jiangsu 210044)

Abstract: Using the air temperature observation data of a region in Shanghai in November 2016, the observation data was divided into five height segments, and the average absolute error (MAE) and root mean square error (RMSE) were used to analyze the temperature observation data. The spline interpolation was performed on the temperature observation data of each height section, and the quality control of sounding was studied by the idea of cross-checking. The results show that the quality control algorithm of cubic spline interpolation has a significant effect on the quality control of the air temperature observation data under low and sunny weather conditions, and can more effectively mark the suspicious value in the temperature observation data.

Keywords: temperature observation data; cubic spline interpolation; cross test; quality control of sounding

引言

探测高空气象观测数据, 它作为气象基础探测数据的重要组成部分, 有着其他气象数据不可替代的作用。高空气象观测主要观测大气层中从地面到三万米左右高度之间不同高度上的风速、风向、温度、气压、高度、湿度等重要气象数据^[1]。正是由于其观测数据的特殊性以及不确定性, 导致探空

数据的异常值或者波动幅度大于地面气象观测数据^[2-5]。

就目前国内外发展来看, 探空质量控制处于一个初级阶段^[6-10]。早在 1997 年, 翟盘茂^[11]对中国历史探空资料进行了整理, 提出了综合静力学法的质量控制方法。2008 年, 郭燕君^[12]对高空温度变化趋势不确定性展开了进一步研究, 得出中国近 50a 来, 全球对流层温度趋于升高, 平流

收稿日期: 2019-03-09

基金项目: 国家自然科学基金项目 (41675156) 和南京信息工程大学人才启动经费 (2243141701053)

作者简介: 潘霄 (1994-), 男, 江苏宜兴人, 硕士研究生, 主要从事气象资料质量控制研究。E-mail: 1527507480@qq.com

层趋于下降。2010a, Paule. Ciesielski^[13] 等人对台湾地区探空数据做了质量控制, 发现湿度误差原因来源于白天干偏差、基线偏差、船甲板加热偏差和缓慢上升探测偏差等四种偏差, 并对这四种偏差展开了一系列的研究。2014 年, Paule. Ciesielski^[14] 团队又对北欧的校正数据集进行了探空质量控制, 阐明了数据集的分类, 以及偏差校正的具体方法。在 2006 年, 许小勇^[15] 等人对三次样条插值函数的构造进行了论述, 表明利用样条插值, 既可以保持分段低次插值多项式, 又可提高插值函数光滑性。2017 年, 朱亚玉^[16] 等人提出了基于分段三次样条函数逐时气象资料模拟方法研究。2018 年, 周冰、李玉立^[17] 对 GPS 掩星大气探测数据进行平滑处理分析。

本文以气象要素中的气温为例, 将气球上升过程中气温变化趋势考虑在内, 着眼于不同气压高度段内的温度变化不一致的特点, 提出了一种基于三次样条插值的探空气温质量控制方法, 对上海地区 2016 年 11 月份气温观测资料进行质量控制。

1 数据与方法

1.1 数据选取

选用上海地区采集的 2016 年 11 月探空气温观测资料。探空资料的观测间隔大多为 12h, 观测时间为 07:00 和 19:00 两个时间。探空资料的观测方式较为特殊, 用放气球的方式使观测资料通过卫星传输到地面中心, 更为特殊的是每个观测时间段只放一次气球, 气球大约经过 80min 的时间升至 3×10^5 m 高空附近自动炸毁, 所以整个观测时间段不具有连续性特征, 而在气球上升过程中具有连续性。

为检验所构建质量控制模型的适用性, 在原始数据中种入随机产生的误差值来模拟可能产生的错误观测值, 以此作为被检数据, 误差值 W_x 通过式 (1) 产生。

$$W_x = \sigma_x \cdot p_x \quad (1)$$

式 (1) 中, 是均匀分布的随机数, 服从区间为 $[-1, 1]$, 符合均值为零的均匀分布; 为原始观测值标准差, 为误差所种入位置; 为待插入原始数据的误差。

1.2 三次样条插值算法

三次样条插值算法是在 Hermite 插值的基础上引入了二阶导数, 且要求插值函数, 一阶导数, 二阶导数在区间上都连续, 这使得插值函数更接近原函数曲线, 提高了插值的精度以及稳定性。

三次样条插值用不超过三阶的函数连接两个相邻的节点, 要求在节点处 $S(x_i) = Y_i$, 且插值函

数 $S(x)$ 、一阶导数 $S'(x)$ 、二阶导数 $S''(x)$ 在区间上都连续。在区间 $[x_{i-1}, x_i]$ ($i=1, 2, \dots, n-1$) 上, $S(x)$ 为:

$$S(x) = S'(x_{i-1}) \frac{(x_i - x)^3}{6m_{i-1}} + S'(x_i) \frac{(x - x_{i-1})^3}{6m_{i-1}} + \left(Y_{i-1} - \frac{S'(x_{i-1})m_{i-1}^2}{6} \right) \frac{x_i - x}{m_{i-1}} + \left(Y_i - \frac{S'(x_i)m_{i-1}^2}{6} \right) \frac{(x - x_{i-1})}{m_{i-1}} \quad (2)$$

其中 $m_{i-1} = x_i - x_{i-1}$ 。由于函数 $S(x)$ 在样点 x_i 处具有连续二阶导数, 增加自由边界条件:

$$\begin{aligned} S''(x_0) &= Y''_0 = 0 \\ S''(x_n) &= Y''_n = 0 \end{aligned} \quad (3)$$

得到 $S'(x)$ 满足的方程:

$$\mu_i S'(x_{i-1}) + 2S'(x_i) + \lambda_i S'(x_{i+1}) = \delta_i \quad (4)$$

其中:

$$\begin{cases} \mu_i = \frac{m_{i-1}}{m_{i-1} + m_i} \\ \lambda_i = 1 - \mu_i \\ \delta_i = 6 \left(\frac{Y_{i+1} - Y_i}{m_i} - \frac{Y_i - Y_{i-1}}{m_{i-1}} \right) (m_i + m_{i-1})^{-1} \end{cases} \quad (5)$$

式 (3) ~ (5) 中 $\mu_0 = 0$, $\delta_0 = \delta_n = 0$, $\lambda_n = 0$ 。求式 (3) ~ (5) 并将结果代入公式 (2) 即可得出每个子区间的三次样条函数。

1.3 基于三次样条插值的质量控制算法

基于上述的三次样条插值方程, 将公式 (2) 应用到探空气温资料的质量控制当中, 其过程分为以下几个步骤:

第一步, 选取上海地区 2016 年 11 月某一次探测的探空气温观测资料整段气压高度序列以及每个气压高度序列所对应的气温序列。

第二步, 将整段气压高度序列按照气温的平滑程度分成四到五个高度段。选取其中一段气压高度序列 $\{Y_h, h=1, 2, 3, \dots, H\}$ 以及对应的气温序列 $\{T_h, h=1, 2, 3, \dots, H\}$ 。

第三步, 把选取的一段气压高度序列分为训练集和测试集两个序列, 对训练集用三次样条插值算法进行拟合, 得到一组序列维度与测试集相同的模拟序列, 把模拟序列与测试集序列做对比分析。

第四步, 根据交叉验证的原则, 对上述训练集和测试集的序列进行变换, 保证一段气压高度序列都做过测试集并且没有重复。上述过程为一段气压高度序列的实验, 再对其他高度段做重复实验。

第五步, 将预测值与观测值进行比较, 若其差值满足式 (6) 则认为数据通过检验, 若不满足

则认为数据可疑,对可疑数据进行标记。

$$|Y_{est} - \bar{Y}| \leq f \cdot \sigma \quad (6)$$

式(6)中,为质量控制参数,为观测序列标准误差。上述质量控制方法记为基于三次样条插值的探空气温质量控制算法。

2 试验结果分析

2.1 预测结果及分析

随机抽取 6 组上海地区 2016 年 11 月探空气温观测资料。由于不同高度段的气温波动程度不一样,把每一组气温观测资料分成 1000 ~ 550hPa; 550 ~ 250hPa; 250 ~ 110hPa; 110 ~ 45hPa; 45 ~ 10hPa 五个高度段,对每一个高度段使用三次样条插值算法,模拟出气温随着高度的变化趋

势。通过“交叉验证”对三次样条插值法的模拟效果以及稳定性进行比较,模拟效果的指标 RMSE 和 MAE 平均值分别如图 1a 和图 1b 所示。从图 1a 看出,除了 11 月 3 号的 100hPa 以上高空的 RMSE 值大于 1 以外,其余值都小于 1,有些低空区域甚至小于 0.01。图 1b 显示,11 月 3 号的 MAE 值普遍大于其他时间段,但总体 MAE 值都小于 1,在正常误差范围之内。这说明三次样条插值算法对于探空气温观测资料的模拟精度效果良好,适用于探空气温质量控制。由于 11 月 3 号在每个高度段上的 RMSE 值都远高于其他日期,对此进行了深入分析发现:当天降水量远超其他日期,导致相对湿度很大,以至于气温曲线的波动幅度比较明显。从高度方面,几乎所有日期都是 RMSE 值随着高度的升高而升高,这说明了高度越低,三次样条插值法的模拟效果越明显,精度越高。

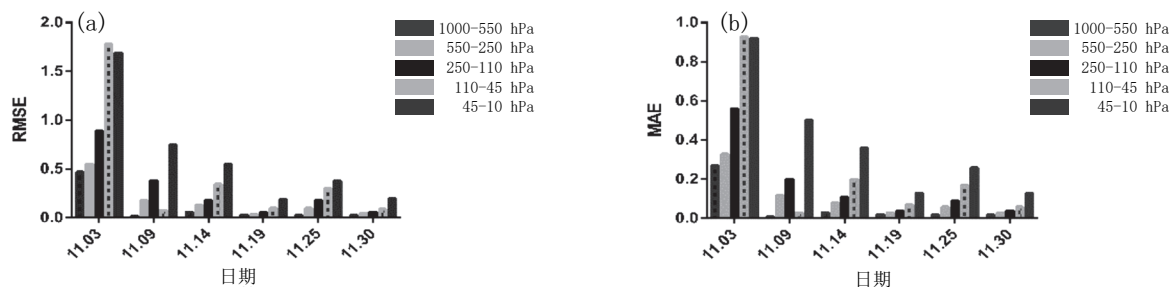


图 1 2016 年 11 月各时间段不同高度的误差指标 (a:RMSE 值; b:MAE 值)

采用“交叉检验”的方式来进行验证,从六组样本中随机挑选 11 月 30 号气温值模拟误差,如表 1 所示。五个高度段的误差指标最小值、最大值和平均值都非常接近,表明了三次样条插值算法的稳定性在各个高度段都比较好。如图 2 所

示,为 11 月 30 号整个高度段的模拟效果误差分析,可以直观看出 RMSE 值和 MAE 值随着高度升高而升高,且在同一高度段的三次实验误差都比较接近。这也再次证明了三次样条插值算法的稳定性和普适性都非常好。

表 1 11 月 30 号各高度段 RMSE 和 MAE 值

数值	1000 ~ 550 hPa		550 ~ 250 hPa		250 ~ 110 hPa		110 ~ 45 hPa		45 ~ 10 hPa	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
最小值	0.0274	0.0209	0.0353	0.0286	0.0500	0.0442	0.0893	0.0643	0.2044	0.1310
最大值	0.0280	0.0215	0.0706	0.0345	0.0813	0.0379	0.0890	0.0630	0.1960	0.1266
平均值	0.0278	0.0213	0.0478	0.0310	0.0617	0.0408	0.0891	0.0637	0.2011	0.1294

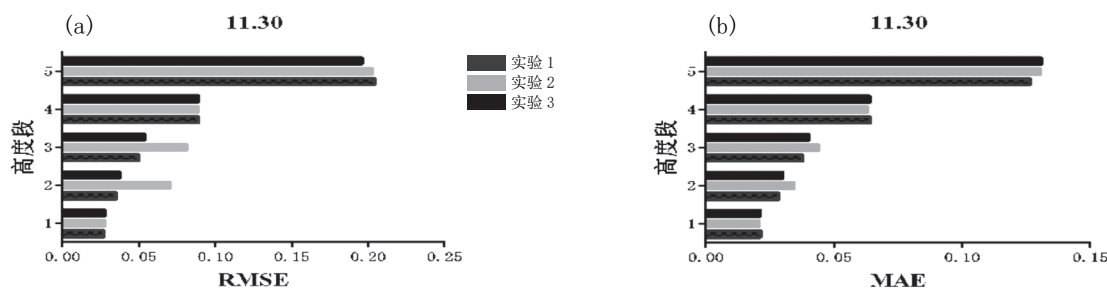


图 2 2016 年 11 月 30 日各高度段通过交叉检验的误差指标
(a:RMSE 值; b:MAE 值。纵坐标数 1 表示 1000 ~ 550hPa 高度段; 2 表示 550 ~ 250hPa 高度段; 3 表示 250 ~ 110hPa 高度段; 4 表示 110 ~ 45hPa 高度段; 5 表示 45 ~ 10hPa 高度段)

影响算法模拟效果的另一个主要因素为天气状况。李平^[18]等认为探空气球漂移量随高度增加而增大, 导致气温变化的稳定性随着高度的升高而降低。在出现极端天气比如下雨或者台风时, 气温探测非常不稳定, 经常出现数据缺测或者气球偏离严重情况, 导致算法拟合变得困难。中文随机抽取的六组样本数据所对应的低空相对湿度均在 50% 以下, 只有几次相对湿度接近 80%, 主要是由于气球在上升过程中穿过云层导致。这表明这几组样本数据当时对应的天气状况良好。

2.2 检验效果及分析

在每个高度段中种入随机错误值, 根据式 (6) 检验数据, 计算各样本检错率, 图 3 是六组样本的平均检错率。从图 3 看, 所有检错率均在 75% ~ 95% 之间, 大多集中在 85% 左右, 检错率处于相对较高的水平, 这说明了使用三次样条插值算法对于探空气温质量控制的效果比较明显。与图 1 结合比较, 相对于误差越明显的样本, 检错率越低, 检错率也普遍随着高度的升高而下降, 这也再次验证了三次样条插值算法对于低空气温模拟更有效。

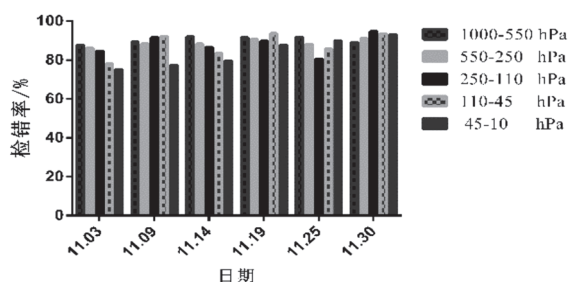


图 3 11 月各高度段的平均检错率

为了判别 f 值对检验效果的影响, 在人为添加错误值不变的情况下, f 取不同值时, 某台站的检验效果如图 4 所示。当 f 为 0.41 时, 对应检错率为 84%, 此时, 第一类错误最大, 第二类错误最小; 随着 f 值增大, 第一类错误减小, 第二类错误增大, 检错率逐渐减小。根据此原则, 选择第一类错误与第二类错误插值最小时的检错率

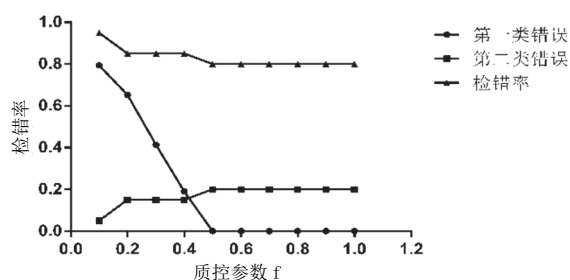


图 4 不同 f 值下的第一类错误、第二类错误以及检错率

作为最佳检错率。从这可以看出, 质控参数表示对气象数据的质量要求, 具体值需要根据具体应用环境来确定。

3 结论与讨论

以上提出了一种三次样条插值算法对探空气温进行质量控制, 并对上海地区 2016 年 11 月气温观测资料进行检验, 通过误差指标分析算法的模拟效果和稳定性, 可以得到以下结论:

(1) 三次样条插值算法适用于天气状况良好的探空质量控制, 此时模拟精度相对较好, 对于改善气温资料的缺测率和突变率有很好的效果。该方法能够有效检验出不同时间观测的不同气温观测资料中植入的人为误差, 说明这种质量控制算法具有较强的适应性和较高的误差识别能力。

(2) 通过三次样条插值算法结合“交叉检验”思想, 分析出对于不同高度段的质量控制效果是不一样的, 高度越低, 质量控制效果越明显, 所以该质量控制算法适用于低空气温观测资料, 而对于高空虽然没有低空效果显著, 但也有着良好的效果。

(3) 对基于三次样条插值质量控制算法的质量控制参数进行分析, 不同高度段最佳质量控制参数的选取存在微弱的差异, 相应的, 对错误检出率也有着较大的影响。可根据不同高度段的具体要求, 选择合适的质量控制参数, 达到最佳质量控制效果。

本文所述的是三次样条插值算法在探空质量控制上简单地应用, 仅针对单一的气象要素进行的研究, 并且在地区上具有一定的局限性。为了证明地区普遍性, 可以把其他地区的数据加入作比较, 也可以研究极端天气条件的气温如何模拟。未来也可以深入研究温度与湿度的时空一致性, 这对于湿度的探空质量控制会是一个突破口。

参考文献:

- [1] 何文旺, 谢东, 雷红萍. 基层台站实时气象观测资料质量控制初探 [J]. 气象研究与应用, 2012, 33(3): 60-63.
- [2] 丁丽佳, 蔡丽青, 凌良新, 等. 潮州热量资源空间分布模型及插值方法 [J]. 广东气象, 2015, 37(5): 62-64.
- [3] 李茂, 谢近东, 马佩强, 等. 东源探空站秒级数据在探空异常情况下的应用 [J]. 广东气象, 2014, 36(5): 79-80.
- [4] 张明, 杜裕, 廖雪萍. 基于探空秒级数据的鄂西南高空风特征分析 [J]. 气象研究与应用, 2018, 39(1): 86-89.

程, 提高传输的可靠性和时效性。引入分布式流式计算引擎, 构建自动气象站资料实时处理组件, 实现自动气象站资料解码、变温变压统计、小时累计降水量统计、分钟累计降水量统计和日值统计等功能, 通过数据处理状态跟踪机制和关联更新机制, 保障处理数据的完整性和一致性。测试结果表明, 系统的处理效率可以满足气象业务时效性要求。

系统完成了基本功能的开发, 但是也存在一定的问题, 例如系统的部署和维护不够方便、对平台整体运行状态的监控不足、数据处理效率还存在进一步提升的空间等等, 离业务化应用还有距离。在后续的研究和应用中, 可以利用容器技术对平台进行改造, 将平台的各模块进行容器化改造并部署到容器管理平台, 提高应用部署效率, 方便管理维护, 利用容器平台提供的接口实现系统资源的监控; 在处理组件的重要业务逻辑中设置埋点, 采集资料处理的性能监视信息; 与基于消息传输的标准格式 (BUFR) 自动气象站资料进行对接, 进一步提高自动气象站资料业务的时效性。

参考文献:

[1] 孙超, 霍庆, 任芝花, 等. 地面气象资料统计处理

系统设计与实现 [J]. 应用气象学报, 2018, 29(5): 630-640.

- [2] 张蕾, 王明洁, 李辉. 短时强降水临近预报相对准确率的探讨 [J]. 广东气象, 2015, 37(2): 1-6.
- [3] 李磊, 张立杰, 力梅. 深圳降水资料信息挖掘及在气候服务中的应用 [J]. 广东气象, 2015, 37(2): 48-51.
- [4] 熊文兵, 叶海宁, 吴凤莹, 等. 基于移动互联的智慧气象为农服务系统研究 [J]. 气象研究与应用, 2018, 39(3): 63-65+91+132.
- [5] 蒙绍臻, 林奕桐, 李仕强, 等. 自动站温度、雨量数据的质量控制方法和应用研究 [J]. 气象研究与应用, 2014, 35(1): 99-103.
- [6] 何健, 王潜梅, 钱光明, 等. 广东省区域自动气象站资料的质量控制与评估 [J]. 广东气象, 2011, 33(3): 37-40.
- [7] 詹利群, 黄玮萱, 陈德诚. 自动气象站中心站资料传输流程优化实践 [J]. 气象研究与应用, 2013, 34(3): 68-71.
- [8] 张来恩, 王鹏, 韩鑫强. CTS2.0 消息封装及交换控制策略设计及实践 [J]. 气象科技进展, 2018, 8(1): 271-273.
- [9] 赵文芳, 刘旭林. Spark Streaming 框架下的气象自动站数据实时处理系统 [J]. 计算机应用, 2018, 38(1): 38-43.
- [10] 张恩红, 李高洁, 乔文文, 等. 广东省气象数据通信系统的架构优化及应用分析 [J]. 广东气象, 2017, 39(4): 73-76.
- [11] 寇媛媛, 王晓明, 杨玉红, 等. 分区技术在气象数据库优化中的应用 [J]. 广东气象, 2018, 40(5): 73-76.

(上接第 93 页)

- [5] 李文娟, 郇敏杰. 基于探空资料的雷暴潜势预报方法 [J]. 广东气象, 2011, 33(3): 28-30.
- [6] 覃晓玲, 黎洁波, 韦丽英. 如何正确使用探空高度代替斜距的方法 [J]. 气象研究与应用, 2007, 28(1): 85-86.
- [7] 马佩强, 张运林, 李茂等. L 波段探空雷达跟踪异常成因及应对措施 [J]. 广东气象, 2008, 30(2): 89-90.
- [8] 翁锦辉, 罗建平, 王凡, 等. 气象探空数据动态比对中误差计算方法研究 [J]. 气象研究与应用, 2010, 31(3): 74-76.
- [9] 姚日升, 曹艳艳, 涂小萍. 插值方法在提高热带气旋路径预报时效分辨率中的应用 [J]. 广东气象, 2011, 33(1): 13-15.
- [10] 王超球, 许嘉玲. 区域自动气象站质量控制参数值高度订正的方法 [J]. 气象研究与应用, 2011, 32(3): 64-66.
- [11] 翟盘茂. 中国历史探空资料中的一些过失误差及偏差问题 [J]. 气象学报, 1997, 55(5): 563-572.
- [12] 郭艳君. 高空大气温度变化趋势不确定性的研究进展 [J]. 地球科学研究进展, 2008, 23(1): 24-29.
- [13] Paule. Ciesielski, WEN-Ming, SHAO-Chin, et al. Quality-Controlled Upper-Air Sounding Dataset for TiMREX/SoWMEX: Development and Corrections. [J]. American Meteorological Society, 2010, 46(2): 330-351.
- [14] Paule. Ciesielski, WEN-Ming, SHAO-Chin, et al. Quality-Controlled Upper-Air Sounding Dataset for DYNAMO/CINDY/AMIE: Development and Corrections [J]. American Meteorological Society, 2014, 78(3): 443-462.
- [15] 许小勇, 钟太勇. 三次样条插值函数的构造与 Matlab 实现 [J]. 自动测量与控制, 2006, 25(11): 1006-1576.
- [16] 朱亚玉, 宋丽莉, 姬兴杰. 基于分段三次样条函数逐时气象资料模拟方法研究 [J]. 气象与环境学报, 2017, 33(2): 44-52.
- [17] 周冰, 李玉立. GPS 掩星大气探测中数据的平滑处理方法分析 [J]. 城市勘测, 2018, 4(3): 88-96.
- [18] 李平, 徐枝芳, 范广洲, 等. 探空温度资料质量控制技术研究 [J]. 气象, 2013, 39(12): 1626-1634.