한국의 기온자료 품질관리 알고리즘의 검증

박창용 · 최영은*

건국대학교 지리학과

(접수: 2012년 2월 24일, 수정: 2012년 4월 7일, 게재확정일: 2012년 4월 10일)

Validation of Quality Control Algorithms for Temperature Data of the Republic of Korea

Changyong Park and Youngeun Choi*

Department of Geography, Konkuk University

(Received: 24 February 2012, Revised: 7 April 2012, Accepted: 10 April 2012)

Abstract This study is aimed to validate errors for detected suspicious temperature data using various quality control procedures for 61 weather stations in the Republic of Korea. The quality control algorithms for temperature data consist of four main procedures (high-low extreme check, internal consistency check, temporal outlier check, and spatial outlier check). Errors of detected suspicious temperature data are judged by examining temperature data of nearby stations, surface weather charts, hourly temperature data, daily precipitation, and daily maximum wind direction. The number of detected errors in internal consistency check and spatial outlier check showed 4 days (3 stations) and 7 days (5 stations), respectively. Effective and objective methods for validation errors through this study will help to reduce manpower and time for conduct of quality management for temperature data.

Key words: Temperature data, Quality management, Validation, Errors

1. 서 론

기후자료는 기후특성 및 기후변화 규모를 파악하거나 미래 기후전망을 위한 시나리오의 기초자료로 사용된다. 신뢰할 수 있는 장기간의 고품질 기후자료는 기후변화에 대한 정확한 이해와 탐지, 향상된 기후예측 모델의 생산뿐만 아니라 기후변화로 인한 부정적 영향의 적응 및 완화를 위한 대응 체계 마련에도 필수적이다. 따라서 기후변화 문제와 관련하여 과학적 평가의 기반이 되는 기후자료의 품질에 대한 관심이 높아지고 있다 (Peterson *et al.*, 1998a; Karl *et al.*, 1999; Nicholls and Murray, 1999; Vincent *et al.*, 2002; WMO, 2003; 2008; Brunet *et al.*, 2006). 그러나 오류

가 있는 기후자료가 통계나 분석에 사용된 경우에는 결과에 결함을 야기시키기도 한다 (이정덕 등, 2010). 따라서 기후자료의 오류 의심자료를 검출하는 알고리즘의 개발과 검출된 오류 의심자료의 오류 여부의 판단을 위한 검증 연구가 필요하다.

기후자료의 품질관리를 위해서는 전수검사가 필요하다. 하지만 전수검사는 많은 시간과 인력이 소요된다. 따라서 통계적 방법에 기초한, 보다 효율적인 품질관리 알고리즘이 개발되어야 한다. 이러한 연구는 미국, 유럽, 중국 등에서 활발하게 수행되었으며 주요 연구방법을 Table 1에 제시하였다 (Peterson and Vose, 1997; Peterson *et al.*, 1998b; Klein Tank *et al.*, 2002; Feng and Qian, 2004; Brunet *et al.*, 2008). 우리나라에서도 2005년부터 관측자료에 대한 품질관리가 실시간·준 실시간 및 비실시간으로 수행되고 있으나 기후자료는 매 분 획득되는 기상자료와는 달리 시간적인 통계값에 해당하므로 다른 형태의 품질관리

*Corresponding Author: Youngeun Choi, Department of Geography, Konkuk University, Seoul 143-701, Korea.
Phone : 82-2-450-3447, Fax : +82-2-3436-5433
E-mail : yechoi@konkuk.ac.kr

Table 1. Summary for quality control of climate data of key references.

Reference	Method
Peterson and Vose(1997) and Peterson <i>et al.</i> ,(1998b)	source data sets evaluation, station time series evaluation and individual data points evaluation
Klein Tank <i>et al.</i> ,(2002)	check for the occurrence of miscoding
Feng and Qian(2004)	high-low extreme check, internal consistency check, temporal outlier check and spatial outlier check
Brunet <i>et al.</i> ,(2008)	gross error checks, tolerance tests, internal consistency and spatial coherency

가 요구된다 (이정덕 등, 2010). 또한 기후자료의 동질성 검사 및 보정과정 이전에 품질관리 과정이 선행되어야 한다 (박창용과 최영은, 2011). 이정덕 등 (2010)은 우리나라 기온자료의 품질관리를 위한 알고리즘을 개발하고 부산을 사례로 이를 적용하였으며, 박창용과 최영은 (2011)은 이정덕 등 (2010)에서 개발된 알고리즘을 이용하여 우리나라 61개 지점의 기온자료 품질관리를 실시하고 동질성 검사와 보정을 수행하였다. Brunet *et al.* (2008)에 의하면 통계적 방법에 기초한 기후자료 품질관리를 수행한 후 검출된 오류 의심자료는 주변 관측소의 값과 서로 비교하거나 강수 또는 기압과 같은 다른 기후요소를 서로 비교하여 연구자의 관점에서 검출된 오류 의심자료의 이용 여부를 판단하도록 제시하였다. 박창용과 최영은 (2011)의 연구에서는 기온자료 품질관리를 통해 검출된 오류 의심자료를 단순히 주변 관측소의 일자료와 비교하여 오류 여부를 검증하였을 뿐 아직까지 기온자료 품질관리를 통해 검출된 오류 의심자료의 오류 여부에 관하여 체계적이고 구체적인 검증 연구는 아직까지 이루어지지 않았다. 따라서 본 연구는 박창용과 최영은 (2011)의 연구에서 검출된 오류 의심자료의 오류 여부를 주변 관측소 일자료와의 비교뿐만 아니라 종관 분석과 시간별 기온자료, 일강수량, 일최대풍향 등의 자료를 이용하여 더욱 체계적이고 구체적으로 검증하고자 하였다.

2. 연구자료 및 방법

본 연구에서는 장기간의 자료를 보유하고 있는 기상청 61개 관측지점 (Fig. 1)의 38년 (1973~2010년) 동안 일평균기온, 일최고기온, 일최저기온을 이용하였다. 이들 기온의 오류 의심자료에 대한 오류 여부를 판단하기 위하여 00 UTC 일기도와 시간별 기온, 일강수량, 일최대풍향, 일최고기온 및 일최저기온의 발생시간 자료를 이용하였다.

품질관리는 Feng and Qian (2004), 이정덕 등 (2010), 박창용과 최영은 (2011)이 제시한 방법을 이용하였다. 품질관리는 극한값 검사 (high-low extreme check), 내

Fig. 1. The location of 61 stations selected for the study.

적일치성 검사 (internal consistency check), 시간 이례치 검사 (temporal outlier check), 공간 이례치 검사 (spatial outlier check)의 절차로 구성된다 (Fig. 2). 본 연구에서 적용한 품질관리 방법과 절차는 자료 자체에 간단한 통계적 방법을 적용하여 의심 및 오류값을 먼저 검출한 후, 자료의 시·공간적 비교를 통해 선행과정에서 검출되지 않는 의심 및 오류값을 효과적으로 검출할 수 있다는 장점이 있다.

극한값 검사는 극값과의 비교를 통해 각 지점의 일 기온자료에서 극값을 초과하는 값의 존재 여부를 파

Fig. 2. Flowchart of quality control for daily temperature data.

약하는 검사이다. NCDC (2002)는 기온의 상한값 및 하한값을 각각 57.8°C , -89.4°C 로 설정하였는데 이는 전 세계 근대기상관측 이래 가장 높고 낮은 기온으로서, 57.8°C 는 1922년 9월 13일에 리비아 엘 아지자 (El Azizia)에서 관측되었으며, -89.4°C 는 1983년 7월 21일에 남극 보스토크 (Vostok)지에서 관측되었다. 본 연구에서는 이 범위를 벗어나는 기온을 오류로 분류하였다.

내적일치성 검사는 자료가 구조적인 조건에 부합하는지를 판별하는 검사이다. 내적일치성을 벗어나는 오류는 관측원부 및 문서 전자화 과정에서의 입력 오류로 발생한다 (Reek *et al.*, 1992). 내적일치성 검사는 내적불일치성 검사 (internal inconsistency check), 일교차범위 검사 (diurnal temperature range check), 지속성 검사 (flat line check)로 구성되어 있다. 내적불일치성 검사는 관련 변수 간의 관계가 서로 일치성을 가지는지를 판별하는 검사이다. 예를 들어, 일최저기온이 일최고기온보다 높은 경우 일평균기온이 일최저기온보다 낮거나 일최고기온보다 높은 경우를 검출하는 방법으로 검출된 자료는 명백한 오류에 해당한다. 일교차범위 검사는 극한값 검사에서 오류범위에 속하지 않고 내적불일치성 검사에서도 검출되지 않지만 비정상적으로 큰 일교차를 가지는 경우를 검출하는 검사이다. 대상 검사지점의 일교차에 대해 월평균과 표준편차 (σ)를 구하고 검사 대상일의 일교차와 월평균 일교차의 차이가 5σ 이상인 경우 오류 의심자료로 분류한다. 지속성 검사는 연속적으로 같은 값을 가지는 경우를 검사한다. 연속으로 7일 이상 같은 값이 나타

났을 때 첫 번째 값을 제외한 나머지 값을 오류 의심자료로 분류한다. 일교차범위 검사와 지속성 검사를 통해 검출된 오류 의심자료는 검사 대상일의 주변 관측지점의 값과 비교하여 오류 여부를 최종 판단한다.

시간 이레치 검사는 검사 대상일에 대하여 전날과 다음날, 과거의 동일한 날과 그 전날 및 다음날과 큰 차이를 나타내는 값들을 검출하는 과정으로서 현재 미국 국립기후자료센터의 기후자료 품질관리에 적용되고 있다. 시간 이레치 검사는 이평균 (biweight mean)과 이표준편차 (biweight standard deviation)를 이용하는데, 이는 자료 집합의 평균값과 표준편차를 보다 잘 대표할 수 있도록 자료에 대한 정규분포의 중심으로부터 멀어질수록 통계적 가중치를 작게하여 소수의 이레치에 의해 평균값과 표준편차가 커지는 영향을 최소화하였다 (Lanzante, 1996). 먼저 검사대상 지점 기온자료의 날짜와 그 전날 및 다음날의 모든 연도에 해당하는 자료의 집합 (X_i)을 구한다. 이 집합에서 검사대상일 (X_0)은 제외한다. 예를 들면, 1977년 7월 29일 (X_0)의 자료를 검사하기 위해서 해당 지점의 모든 연도의 7월 28일 자료 및 7월 30일 자료와 X_0 (1977년 7월 29일)를 제외한 모든 연도의 7월 29일 자료가 시간 이레치 검사를 위한 자료 집합 X_i 가 된다 (Table 2). 이렇게 얻은 X_i 로부터의 중간값 M 을 구하고, M 과 X_i 편차의 절대값 ($|X_i - M|$) 시계열의 중간값인 MAD (absolute deviation from the median)를 구한다. 가중치 u_i 는 식 (1)을 이용하여 계산한다. 산출된 $|u_i| \geq 1.0$ 인 경우 $u_i = 1.0$ 으로 정의한다. 식 (1)에서 c 는 가중치가 0으로 전환되는 정규분포의 중심으로부터의 거리를 결정하는 상수이다. Lanzante (1996), Feng and Qian (2004), 이정덕 등 (2010)은 5σ 에서 0으로 전환되는 7.5를 사용하였으며, 본 연구에서도 이를 적용하였다.

$$u_i = \frac{X_i - M}{c \cdot MAD} \quad (1)$$

Table 2. Example of X_i for temporal outlier check on 29 Jul 1977, patterned cell showed a section of X_i , X_0 is the value for 29 Jul 1977.

		day				
		...	28 Jul	29 Jul	30 Jul	...
year						
1973						
...						
1976						
1977				X_0		
1978						
...						
2010						

식 (1)에서 산출된 u_i 값과 X_i 로 이평균 (식 (2)) 및 이표준편차 (식 (3))를 산출한다.

$$\bar{X}_{bi} = M + \frac{\sum_{i=1}^n (X_i - M)(1 - u_i^2)^2}{\sum_{i=1}^n (1 - u_i^2)^2} \quad (2)$$

$$s_{bi} = \frac{\left[n \sum_{i=1}^n (X_i - M)^2 (1 - u_i^2)^4 \right]^{1/2}}{\left| \sum_{i=1}^n (1 - u_i^2)(1 - 5u_i^2) \right|} \quad (3)$$

식 (3)과 (4)에서 이평균 ($\bar{X}_{bi}$) 및 이표준편차 (s_{bi})는 정규분포로 가정된 자료의 양 끝보다 중심 쪽으로 조금 더 크게 가중되므로, 분포의 처음부터 끝까지 동일한 가중치를 적용하는 일반적인 방법보다 이례치에 의한 영향을 덜 받게 된다. 이렇게 계산된 이평균 및 이표준편차를 이용하여 Z 점수 (Z-score)를 산출하며 점수가 5를 초과하면 오류 의심자료로 검출한다 (식 (4)).

$$Z = \frac{|X_o - \bar{X}_{bi}|}{s_{bi}} \quad (4)$$

Lanzante (1996)에 의하면 X_i 의 크기가 11개 이상인 경우 Z 점수가 3, 18개 이상인 경우 Z 점수가 4를 초과하는 자료를 오류 의심자료로 분류했다. 하지만 Z 점수의 기준을 3 또는 4로 결정할 경우 많은 수의 자료가 오류 의심자료로 분류되었다. 따라서 오류 의심자료로 분류하는 기준은 Feng and Qian (2004)에서 제시한 5를 사용하였다. Z 점수의 기준 값을 5로 하는 경우 의심으로 분류되는 값들은 평균에 비해 10~20°C 정도 높거나 낮은 값에 해당하였다. 시간 이례치 검사로 검출된 오류 의심자료의 오류 여부를 판정하기 위하여 오류 의심자료 발생일의 값과 주변 관측 지점 값 및 전날의 관측값을 비교하고, 00 UTC 일기도를 통해 종관분석을 실시하였다.

공간 이례치 검사는 주변 지점의 자료와 비교하였을 때 특이하게 나타나는 이례치를 검출하는 검사이다. 먼저 공간 이례치 검사 대상지점에서 직선거리로 가장 가까운 10개의 지점을 비교를 위한 후보지점으로 선정한다. 검사대상 지점과 선정된 후보지점의 상관계수 R (correlation coefficient)을 월별로 산출하고, 선정된 후보지점과 검사대상 지점의 월별 선형회귀식을 구한다. 10개의 가장 가까운 후보지점 중에서 검사대상 지점과 $R \geq 0.75$ 의 상관관계를 보이는 후보지점을 비교지점으로 선별하고 그 수가 5개를 초과할 경우에

는 검사대상 지점과 후보지점과의 선형회귀에서 제공된 평균오차가 가장 작은 비교지점 5개를 선별한다. N 개의 비교지점 ($N \leq 5$)에서 선형회귀관계를 이용해 비교지점의 관측값을 가지고 검사대상 지점의 예상 관측값 (VF_{ij})을 구한다. 검사대상지점의 관측값이 V_i 일 때 N 개의 쌍에 대해서 검사대상 지점의 관측값이 식 (5)에서 제시한 범위에 존재하는지 검사한다 (Hubbard and Sivakumar, 2001).

$$VF_{ij} - F \times RMSE_j < V_i < VF_{ij} + F \times RMSE_j \quad (5)$$

$j = 1, \dots, N$ 는 비교지점이고 $i = 1, \dots, m$ 는 특정일의 관측값을 의미한다. VF_{ij} 는 선형회귀관계를 통해 j 지점의 i 일의 관측값으로 구한 검사대상 지점 일기온의 예상 값을 의미한다. 여기서 F 는 신뢰한계를 결정짓는 값이다. V_i 의 자료 집합이 정규분포를 이룬다고 가정하였을 때, 식 (5)에서 $F = 2$ 이면 V_i 는 94.45% 신뢰구간에 포함되며, $F = 3$ 이면 V_i 는 99.73% 신뢰구간에 포함된다. 공간 이례치 검사에서는 검사대상 지점에 따라 월별로 다른 비교지점을 가지게 된다. 식 (5)를 이용하여 검사대상 지점의 값을 검사하고 모든 비교지점에서 범위를 벗어나는 경우 오류 의심자료로 검출한다. 본 연구에서 $F = 2$ 를 적용할 경우 많은 수의 자료가 오류 의심자료로 분류되므로 Feng and Qian (2004)이 제시한 $F = 3$ 을 적용하였다. 이후 공간 이례치 검사로 검출된 오류 의심자료의 발생일의 시간별 기온, 일강수량, 일최대풍향, 일최고기온 및 일최저기온 발생 시간 분석을 통하여 오류 여부를 판단하였다.

3. 연구결과

3.1. 극한값 및 내적일치성 검사

극한값 검사를 수행한 결과 원기온자료의 일평균 기온, 일최고기온, 일최저기온에서 상한 및 하한값을 벗어나는 오류는 검출되지 않았다. 내적일치성 검사 중 내적 불일치성 검사를 통해 강릉, 부여, 밀양 등 3개 지점에서 구조적으로 명백한 4개 (4일)의 오류가 검출되었다 (Table 3). 강릉에서 검출된 오류는 일평균기온이 일최고기온보다 높았고, 부여, 밀양에서 검출된 오류는 일평균기온이 일최저기온보다 낮았다. 이러한 오류는 1973년과 1974년에 발생한 것으로서 일기온자료를 잘못 기입했거나 데이터베이스 구축과정에서 발생한 오류로 판단된다.

3.2. 시간 이례치 검사

시간 이례치 검사에서는 총 80회의 오류 의심자료가 검출되었다. 일평균기온은 17개 지점에서 19회로 23.8%, 일최고기온은 1개 지점에서 1회로 1.3%, 일최저기온은 35개 지점에서 60회로 75.0%를 차지하여 일

Table 3. Results for internal consistency check.

Station	Date	Mean (°C)	Max. (°C)	Min. (°C)	Reasons
Gangneung	3 June 1973	23.2	19.2	9.9	mean temp. > maximum temp.
	6 June 1973	23.3	19.6	14.6	mean > maximum temp.
Buyeo	10 March 1973	4.2	16.9	4.8	mean < minimum temp.
Miryang	7 February 1974	-0.5	3.8	-0.2	mean < minimum temp.

최저기온에서 가장 많은 오류 의심자료가 검출되었다. 일평균기온과 일최저기온이 함께 검출된 횟수는 6개 지점의 6일로 나타났다. 지점들이 중복되어 나타난 것을 1일로 보았을 때 21일에서 오류 의심자료들이 검출되었다. 2009년 3월 18일에 13개 지점에서 일평균기온 6회, 일최저기온 10회가 검출되어 이 날에 오류 의심자료가 가장 많이 검출되었다. 1988년 7월 30일에 10개 지점에서 일최저기온 10회, 그 다음날인 1988년 7월 31일에 7개 지점에서 일최저기온 7회 등의 순으로 오류 의심자료가 검출되었다.

시간 이레치 검사로 오류 의심자료가 검출된 지점들이 동일한 일에 나타난 경우가 다수 발견되었다. 1988년 7월 30일과 7월 31일은 각각 10개 지점과 8개 지점에서 일최저기온이 오류 의심자료로 검출되었다. 1988년 7월 30일과 7월 31일은 31일 광주의 약한 강수 (0.0 mm)를 제외하고는 모든 관측지점에서 강수가 없었으며 2일 모두 많은 지점에서 동풍계열이 우세하였다. Fig. 3은 1988년 7월 30일과 1988년 7월 31일 00 UTC 지상일기도이다. 1988년 7월 30일에는 일기도에 저기압으로 표시된 일본 남부에 위치한 강한 열대폭풍 (Severe Tropical Storm, STS) 아그네스 (Agnes)가 동쪽에 위치한 북태평양 고기압을 파괴하면서 서쪽으로의 진출을 저지하였다. 1988년 7월 31일에는 일본 남부에 위치한 아그네스가 태풍으로 강화되면서 일본 동부로 이동하였고, 태풍 아그네스 남북으로 또 다른 저기압 중심이 위치하여 북태평양고기압의 진출을 더욱 저지하였다. 따라서 2일간 동해에 위치한 고기압의 영향으로 많은 지점에서 동풍계열이 우세했던 것으로 판단하였다.

이로서 7월 30일은 모든 관측지점에서 일최저기온이 평년보다 최소 2.1°C (성산)에서 최대 9.6°C (영주) 낮은 값이 나타났고, 7월 31일도 유사한 일최저기온이 나타났다. 이와 같은 원인으로 1988년 7월 30일과 7월 31일은 대부분의 지점에서 평년보다 현저히 낮은 일최저기온을 보였다. 따라서 시간 이레치 검사를 통해 검출된 1988년 7월 30일과 7월 31일의 오류 의심자료는 오류자료가 아닌 정상자료로 판단하였다.

1993년 7월 21일과 7월 22일은 5개 지점에서 일최저기온이 오류 의심자료로 검출되었다. 2일 모두 동해안 지역에서 강수가 있었으며, 대부분의 지점에서

Fig. 3. Surface weather chart at 00 UTC, A: on 30 July 1988, B: on 31 July 1988.

동풍이 우세하였다. Fig. 4는 1993년 7월 21일과 7월 22일의 00 UTC 지상일기도이다. 2일 모두 오후크크해에 중심을 둔 고기압의 영향을 받고 있으며, 일본 남부에 동서로 길게 저기압이 형성되어 있다. 이는 장마전선으로 판단된다. 따라서 오후크크해 기단의 영향으로 동해안 지역에서는 강수가 발생하였고 전국의 기온이 평년보다 현저히 낮았음을 볼 때 시간 이레치 검사를 통해 검출된 1993년 7월 21일과 7월 22일의 오류 의심자료는 오류자료가 아닌 정상자료로 판단하였다.

Fig. 4. Surface weather chart at 00 UTC, A: on 21 July 1993, B: on 22 July 1993.

2002년 1월 15일은 6개 지점에서 3회의 일최고기온과 5회의 일최저기온이 오류 의심자료로 검출되었다. 기온의 평년값이 가장 낮은 시기임에 불구하고 이 날은 모든 지점에서 기온이 평년보다 크게 높았다. 일최고기온의 경우 평년보다 최소 6.1°C (거창)에서 최대 14.3°C (정읍) 높았으며, 일최저기온의 경우 평년보다 최소 6.4°C (울릉도)에서 최대 19.8°C (전주) 높았다. 중부이남 지역에서 남풍 또는 서풍이 우세하였다.

Fig. 5(A)의 2002년 1월 15일의 지상일기도를 살펴보았을 때 일기도에 표기되지 않았으나 북쪽의 시베리아 고기압과 남쪽의 고기압 사이에 형성된 전선대가 우리나라에 걸쳐있고 남쪽에 위치한 고기압에서 온난한 공기가 유입됨에 따라 전국의 기온이 평년보다 크게 높아진 것으로 판단하였다. 따라서 2002년 1월 15일에 시간 이례치 검사를 통해 검출된 오류 의심자료는 정상자료로 판단하였다.

2009년 3월 18일은 10개 지점에서 6회의 일최고기온과 10회의 일최저기온이 오류 의심자료로 검출되었

Fig. 5. Surface weather chart at 00 UTC, A: on 15 January 2002, B: on 18 March 2009.

다. 이 날은 모든 지점에서 일최고기온이 평년보다 최소 4.7°C (거제)에서 최대 15.9°C (속초) 높았으며, 일최저기온의 경우 최소 6.1°C (영주)에서 최대 17.5°C (천안) 높았다.

이 날 대부분의 지점에서 남풍 또는 서풍이 우세하였다. Fig. 5(B)의 지상일기도를 살펴보면 앞서 제시했던 2002년 1월 15일의 지상일기도와 유사한 패턴을 보이고 있다. 북쪽의 시베리아 고기압과 남쪽에 고기압 사이에 저기압이 우리나라 북부에 놓여있어서 북쪽의 시베리아 고기압의 진출을 저지하였다. 남쪽에 위치한 고기압에서 온난한 공기가 유입됨에 따라 전국의 기온이 평년보다 크게 높아졌으므로 2009년 3월 18일에 시간 이례치 검사를 통해 검출된 오류 의심자료는 정상자료로 판단하였다. 시간 이례치 검사를 통해 5개 지점 이상에서 오류 의심자료가 발생한 날 뿐만 아니라 나머지 이외의 날들도 지상일기도 또는 주변 관측값과의 비교를 통해 정상자료로 판단하였다. 시간 이례치 검사를 통해 5개 지점 이상에서 오류

Table 4. Results for temporal outlier check.

Date	The number of stations	Temperature	Decision of error
30 July 1988	10	minimum	no errors
31 July 1988	7	minimum	no errors
21 July 1993	5	minimum	no errors
22 July 1993	5	minimum	no errors
15 January 2002	6	maximum minimum	no errors
18 March 2009	13	maximum minimum	no errors

의심자료가 발생한 날, 오류 의심자료가 검출된 수, 기온요소, 오류여부를 Table 4에 나타내었다.

3.3. 공간 이레치 검사

공간 이레치 검사를 통하여 일최고기온은 35개 지점에서 69일로 가장 많은 오류 의심자료가 검출되었고 일최저기온은 20개 지점에서 31일, 일평균기온은 15개 지점에서 17일이 검출되었다. 공간 이레치 검사에서 다음 세 가지 지리적 요인에 의해 발생하는 오류 의심자료는 정상자료로 판단하였다. 첫 번째는 검사 대상지점과 비교 지점이 장벽효과를 일으키는 산맥을 중앙에 두고 위치해 있을 때 발생하는 기온 차이이고, 두 번째는 검사 대상지점이 해안에 위치해 있고 비교 지점이 내륙에 위치해 있을 때 또는 그 반대의 경우에서 발생하는 기상 현상에 의한 것이다. 마지막으로 검사 대상지점과 비교 지점 간의 고도 차이로 발생하는 기온 차이가 있다.

앞에서 언급한 지리적 요인 중 첫 번째 요인으로 인해 발생한 오류 의심자료를 정상자료로 판단한 사례를 들어보았다. Fig. 6은 속초의 1973년 6월 22일 일최고기온과 비교 지점의 일최고기온 및 비교 지점별 속초의 일최고기온 예상 값을 나타낸 것이다. 이날 5개 비교 지점별 속초의 일최고기온 예상 값은 일최고기온인 30.8°C 보다 크게 낮았다. 하지만 5개 비교 지점의 일최고기온은 속초와 같은 영동지역에 위치한 강릉이 31.1°C로 속초보다 높았고, 영서지역에 위치한 나머지 4개 지점에서는 강릉, 속초보다 3°C 이상 낮았다.

이 날은 서풍계열이 우세하여 비교 지점으로 선택된 지점 중 영동지역에 위치한 강릉은 일최고기온이 영서지역에 위치한 비교 지점보다 높았고 매우 건조한 날씨를 보였다. 따라서 강릉과 같은 영동지역에 위치한 속초의 일최고기온은 비교 지점의 예상 값의 차이가 크게 벌어져 공간 이레치 검사를 통해 오류 의심자료로 분류되었다. 속초의 1980년 8월 24일과 1988년 8월 21일에 검출된 일최고기온 오류 의심자료도

Fig. 6. Example for spatial outlier check of Sokcho; selected 5 neighbouring stations and observed (estimated) daily maximum temperature on 23 June 1973.

이와 유사한 패턴을 보였다.

공간 이레치 검사를 통해 검출된 오류 의심자료가 검사 대상지점과 비교 지점 간 지리적인 특성의 차이로 인해 발생한 것으로 판단되면 정상자료로 판단하였다. 또한 지리적인 특성의 차이뿐만 아니라 다음과 같은 경우로 검출된 오류 의심자료도 정상자료로 판단하였다. 첫 번째로, 검출된 오류 의심자료가 비교지점으로 선정되지 않았지만 거리가 가까운 지점의 자료와 비교했을 때 유사한 기온이 나타나거나, 두 번째로, 검사 대상일의 전날과 다음날의 값을 비교했을 때 급격한 기온 변화가 나타나지 않을 때, 마지막으로 검사 대상지점의 시간별 기온을 조사하여 값이 튀지 않고 점진적인 일변화를 보였다면 정상자료로 판단하였다.

공간 이레치 검사로 검출된 오류 의심자료의 대부분이 정상자료로 판정되었다. 일평균기온, 일최저기온의 오류 의심자료는 모두 정상자료였고, 5개 지점 7일의 일최고기온을 오류로 판정하였다.

Fig. 7. Example for spatial outlier check of Gwangju; Selected 5 neighbouring stations and observed (estimated) daily maximum temperature on 17 April 1973.

공간 이레치 검사로 검출된 오류 의심자료 중 오류로 판단한 사례를 들어보았다. Fig. 7은 1973년 4월 17일 광주의 일최고기온과 비교 지점의 일최고기온 및 비교 지점별 광주의 일최고기온 예상 값을 나타낸 것이다. 이 날 광주의 일최고기온은 20.7°C였고, 비교 지점의 일최고기온은 11.7~13.7°C의 분포를 보여서 광주는 비교 지점보다 7~9°C 높았다. 일교차는 15.8°C로서 8~10°C인 비교 지점의 일교차보다 현저히 컸다. 이 시기에는 매시간의 관측값이 아닌 3시간 간격의 관측값이 존재하므로 이를 살펴보면 광주는 오전 9시에 9.9°C, 오후 12시에 13.1°C, 오후 3시에 12.0°C를 보였다. 일최고기온인 20.7°C는 11시 56분에 관측되었다. 4분 후인 오후 12시에 13.1°C로 급격하게 기온이 하강하기란 불가능하다고 판단되므로 1973년 4월 17일 광주의 일최고기온은 오류로 판정하였다. 1982년 9월 21일 영주에서도 이와 유사한 경우가 발생하여 이 날 영주의 일최고기온도 오류로 판정하였다. 이와 같이 공간 이레치 검사는 먼저 주변 관측소 값의 비교를 통해 오류 의심자료를 먼저 검출한 후 시간별 기상 요소 분석 등을 수행하여 극한값 검사, 내적일치성 검사, 시간 이레치 검사를 통해서 검출되지 않는 오류를 효과적으로 검출할 수 있는 장점이 있다고 판단된다.

4. 요약 및 결론

본 연구에서는 우리나라 61개 기상관측지점을 대상으로 기온자료 품질관리를 통해 검출된 오류 의심자

료에 대한 오류 여부를 객관적으로 검증하였다. 극한값 검사에서는 상한 및 하한값을 벗어나는 오류는 검출되지 않았다. 내적일치성 검사 중 내적 불일치성 검사를 통해 강릉, 부여, 밀양 등 3개 지점에서 구조적으로 명백한 4일의 오류가 검출되었다. 이러한 오류는 1973년과 1974년에 발생한 것으로서 일기온자료를 잘못 기입했거나 데이터베이스 구축과정에서 발생한 오류로 판단된다.

시간 이레치 검사로 검출된 오류 의심자료의 오류 여부를 판정하기 위하여 오류 의심자료 발생일의 값과 주변 관측지점값 및 전날의 관측값을 비교하고, 00 UTC 일기도를 통해 종관분석을 실시하였다. 이를 통해 시간 이레치 검사로 검출된 오류 의심자료는 정상자료로 판정되었다. 공간 이레치 검사로 검출된 오류 의심자료의 발생일의 시간별 기온, 일강수량, 일최대풍향, 일최고기온 및 일최저기온 발생 시간 분석을 통하여 오류 여부를 판정하였다. 이를 통해 공간 이레치 검사로 검출된 오류 의심자료의 대부분이 정상자료로 판정되었는데, 일평균기온, 일최저기온의 오류 의심자료는 모두 정상자료였고, 5개 지점 7일의 일최고기온이 오류로 판정되었다.

본 연구에서는 선행 연구의 기온자료 품질관리 알고리즘으로 검출된 오류 의심자료에 대한 오류 여부를 주변 관측소의 일자료와 비교 뿐만 아니라 종관 분석과 시간별 기온자료, 일강수량, 일최대풍향 등의 자료를 이용하여 더욱 체계적이고 객관적으로 검증하였다. 본 연구 결과는 고품질 기온자료 생산을 위해 수행되어야 하는 기온자료 동질성 검사 (homogeneity test) 및 보정 (adjustment) 과정에 핵심적인 기초자료로 활용될 것이다. 또한 본 연구를 통해 기온자료의 오류 의심자료를 검출하고 오류 자료 판정을 위해 효율적이고 객관적으로 검출할 수 있는 방법을 제시함에 따라 기온자료 품질관리를 수행함에 있어서 인력과 시간의 절감에 크게 기여할 것으로 생각한다.

감사의 글

이 연구는 기상청 기후변화 감시·예측 및 국가정책지원 강화사업 (RACS 2010-4013)의 지원으로 수행되었음.

참고문헌

- 이정덕, 최영은, 박창용, 2010: 한국의 기온자료 알고리즘 개발. 기후연구, 5(2), 162-174.
- 박창용, 최영은, 2011: 고품질의 기온자료를 이용한 연·계절 평균기온의 변화에 관한 연구, 대한지리학회지, 46(1), 20-35.

- Brunet, M., Saladié, O., Jones, P. D., Sigró, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., Lopez, D., and C. Almarza, 2006: The development of a new dataset of Spanish Daily Adjusted Temperature Series (SDATS) (1850–2003), *Int. J. Climatol.*, **26**, 1777–1802.
- Brunet, M., Saladié, O., Jones, P., Sigró, J., Aguilar, E., Moberg, A., Lister, D., Walther, A., and C. Almarza, 2008: A case-study/guidance on the development of long-term daily adjusted temperature datasets, *Climate Data and Monitoring WCDMP-No. 66*, World Meteorological Organization WMO-TD 1425, WMO, Geneva.
- Feng, S., and W. Qian, 2004: Quality control of daily meteorological data in China, 1951–2000: A new dataset, *Int. J. Climatol.*, **24**, 853–870.
- Karl, T. R., Nicholls, N., and A. Ghazi, 1999: CLIVAR/GCOS/WMO Workshop on indices and indicators for climate extremes, *Climatic Change*, **42**, 3–7.
- Kenneth G. Hubbard and M.V.K. Sivakumar(Eds.). 2001. Automated Weather Stations for Applications in Agriculture and Water Resources Management: Current Use and Future Perspectives. Proceedings of an International Workshop[held in Lincoln, Nebraska, USA, 6–10 March 2000. Lincoln, Nebraska: High Plains Climate Center and Geneva, Switzerland: World Meteorological Organization.
- Klein-Tank, A. M. G., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Milet, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Salá, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., Van Encelen, A. F. V., Forland, E., Milet, M., Coelho, F., Mares, C., Razuvaev, V., Niepova, E., Cegnar, T., López, J. A., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and P. Petrovic, 2002: Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment, *Int. J. Climatol.*, **22**, 1441–1453.
- Lanzante, J. R., 1996: Resistant, robust and nonparametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data, *Int. J. Climatol.*, **16**, 1197–1226.
- NCDC, 2002: United States climate normals, 1971–2000 inhomogeneity adjustment methodology, Asheville, NC, USA.
- Nicholls, N., and W. Murray, 1999: Workshop on indices and indicators for climate extremes, *Climatic Change*, **42**, 23–29.
- Peterson, T. C., and R. S. Vose, 1997: An overview of the global historical climatology network temperature database, *Bull. Amer. Meteor. Soc.*, **78**, 2837–2849.
- Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Førland, E. J., Hanssen-Bauer, I., Alexandersson, H., Jones, P., and D. Parker, 1998a: Homogeneity adjustments of IN SITU atmospheric climate data: A Review, *Int. J. Climatol.*, **18**, 1493–1517.
- Peterson, T. C., Vose, R., Schmoyer, R., and V. Razuvaev, 1998b: Global historical climatology network (GHCN) quality control of monthly temperature data, *Int. J. Climatol.*, **18**, 1169–1179.
- Reek, T., Doty, S. T., and T. W. Owen, 1992: A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network, *Bull. Amer. Meteor. Soc.*, **73**, 753–762.
- Vincent, L. A., Zhang, X., Bonsal, B. R., and W. D. Hogg, 2002: Homogenization of daily temperatures over Canada, *J. Climate*, **15**, 1322–1334.
- WMO, 2003, Guidance on metadata and homogenization, WMO-TD 1186, WMO, Geneva, Switzerland.
- WMO, 2008, A case-study/guidance on the development of long-term daily adjusted temperature datasets, *Climate Data and Monitoring WCDMP-No. 66*, World Meteorological Organization WMO-TD 1425, WMO, Geneva, Switzerland.