INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. (in press)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/joc.1524

RMetS
ROYAL METEOROLOGICAL SOCIETY

Thermometer screen intercomparison in De Bilt (the Netherlands) – Part II: Description and modeling of mean temperature differences and extremes

T. Brandsma* and J. P. van der Meulen
Royal Netherlands Meteorological Institute (KNMI), De Bilt, the Netherlands

ABSTRACT: Temperatures of nine thermometer screens during a 6-year field experiment in De Bilt (the Netherlands) have been compared. The screens are currently in use throughout the world and comprise the following types: an aspirated Young screen, four naturally ventilated round-shaped multi-plate screens (KNMI, Vaisala, Young, Socrima), a slightly aspirated version of the KNMI screen, a synthetic Stevenson screen (both aspirated and naturally ventilated) and a naturally ventilated wooden Stevenson screen. The multi-plate KNMI screen served as a reference. For the daily minimum, maximum, and mean air temperatures (T_n, T_x, T_{mean}), the absolute seasonal mean differences with the reference were almost all $\leq 0.1\,^{\circ}$C. An exception is the aspirated Young screen for which differences in mean T_x in summer (JJA) are most notable and amount to $-0.43\,^{\circ}$C. The differences of the individual T_n, T_x, T_{mean} values may be much larger than their seasonal averages. For the aspirated Young and the naturally ventilated Stevenson screens they are most notable, where the Young is generally cooler (T_x and T_{mean}) and the Stevensons warmer than the reference. The absolute temperature differences between the screens and the reference are shown to increase with decreasing cloud cover and windspeed. Furthermore, using the original 15-s samples it is shown that random variations cause fast-responding screens to have more extreme T_n, T_x values than slow-responding screens. For a supposed transition of the natural ventilated synthetic Stevenson screen to the reference screen, transfer functions are successfully developed for the 10-min temperature data and the daily T_n and T_x data. It is argued that the explained variance of the models could have further improved when high-accuracy (especially in the range 0–3 m/s) windspeed measurements were available (at screen level and position). Copyright © 2007 Royal Meteorological Society

KEY WORDS air temperature; thermometer screens; field experiment; transfer functions; the Netherlands

Received 16 June 2006; Revised 18 January 2007; Accepted 15 February 2007

1. Introduction

Increasing interest in air temperature trends and extremes (Klein Tank and Können, 2003; Alexander *et al*., 2006) necessitates an adequate assessment of inhomogeneities in temperature measurements on a daily to sub-daily level. A potentially important source of these inhomogeneities is the change in thermometer (or radiation) screen design that currently takes place in many countries around the World. Because of ongoing automation, smaller and economically more attractive screens replace traditional screens like the Stevenson screen. An unwanted spin-off of these new, often automated, systems may be the introduction of inhomogeneities in the temperature time series. These inhomogeneities may be of the same order of magnitude as the long-term temperature trends (e.g. Quayle *et al*., 1991) and may therefore seriously restrict the usefulness of the data.

Screen changes will most likely result in measurements that are closer to the real air temperature, which were defined in Part I as 'the temperature of the air at the position of the sensor if no measurement equipment would be installed' (Van der Meulen and Brandsma, 2007). It is well known that the differences between the temperatures measured in a screen and the real air temperature are largest (up to several degrees Celsius) for clear-sky and windless conditions and that they reduce to zero for overcast and windy conditions (see, e.g. Brandsma, 2004; Parker, 1994, 2004). It logically follows that the magnitude of the inhomogeneities in time series of air temperature, resulting from changes in thermometer screens, strongly depends on the weather conditions. To homogenize the series afterwards on a daily or sub-daily level, weather-dependent corrections have to be applied. However, the development of homogenization methods for that purpose is just starting (Brandsma and Können, 2006; Della-Marta and Wanner, 2006).

Instead of statistically homogenizing temperature time series afterwards, it may be better to derive and apply transfer functions that transform the temperatures of one screen into those of the other. A logical way to proceed may then be to perform a field intercomparison of the old and the new screen. The data of the intercomparison can then be used to develop the transfer function. It is

* Correspondence to: T. Brandsma, KNMI, PO Box 201, 3730 AE De Bilt, The Netherlands. E-mail: theo.brandsma@knmi.nl

not yet clear how long such an intercomparison should take, but sometimes a period of about 1–2 years is suggested (WMO, 2003). The transfer functions should be derived from an analysis of the inter-screen temperature differences in combination with simultaneously measured other weather variables of interest. This labor-intensive work should only be done for a selected set of climate monitoring stations.

Thermometer screen intercomparisons have been reviewed by several authors (Mawley, 1897; Sparks, 1972; Parker, 1994; Barnett *et al*., 1998). Most of the reported studies focus on the thermometer screens which were in use in the past and the emphasis is on mean monthly temperature differences between screens. National studies of the last 15 years focus on the thermometer screens currently in use (e.g. Andersson and Mattisson, 1991; Lefebvre, 1998; Hatton, 2002; Larre and Hegg, 2002). In these latter papers attempts sometimes have been made to study the temperature differences between screens for varying weather situations. The magnitude of the inter-screen temperature differences in the above studies largely depends on the screens compared. Comparison of Stevenson screens (or similar ones) with outdated screens may reveal mean monthly differences in daily, maximum, and minimum temperatures between a few tenths of degree Celsius in winter to about one degree in summer. On the other hand, comparison of Stevenson screens with modern small round-plated screens shows much smaller mean temperature differences, ranging between zero and about two tenths of a degree Celsius.

Although there have been a number of intercomparisons of thermometer screens around the world, only a few have been published in the international literature and little is quantitatively known about the dependence of the mutual differences between screens on other weather variables. Also a balanced treatment of short-term effects (e.g. for a particular day with specific weather conditions) and mean effects (averages for a large number of days) is lacking. In this study, we present the results of a 6-year field experiment in De Bilt (the Netherlands) that originally compared ten screens that are currently in use around the world. In Part I (Van der Meulen and Brandsma, 2007), we focused on understanding inter-screen temperature differences for typical days. As a result of the calibration in that study one of the screens had to be omitted, and the present analysis is therefore restricted to nine screens. We focus here on the mean temperature differences between the screens and extremes in relation to other weather variables. For the screens and climate conditions considered, the study enables the climate-research community to assess whether modern-day changes of thermometer shelters have caused (or will cause) inhomogeneities in average and extreme temperatures. If there are no significant changes, other countries may save the effort and costs of making comparisons themselves. For a case where the changes are significant, we derive transfer functions that can be used to transform the temperature data of one screen into the other.

In the remainder of the paper, Section 2 first describes the data and methods used. Section 3 presents the results of the field experiment. Section 4 concludes the paper with a discussion and conclusions.

2. Data and Methods

2.1. Data

The data used in the screen comparison were collected during a 6-year field experiment (9 January, 1989–11 February, 1995) at the testing-site of KNMI in De Bilt (52°06′N, 05°11′E), situated in the center of the Netherlands. A detailed presentation of the setup of the experiment and the calibration of the temperature data is presented in Part I (Van der Meulen and Brandsma, 2007). This section only presents the most essential information.

All temperature measurements were performed at a height of 1.5 m above a flat terrain with short cut-grass cover and sufficiently far removed from major obstacles. Nine screens are compared in this study. Details of the screens are presented in Table I. The KNMI multi-plate screen (Knmi.ref) operated during the whole 6-year period and is taken as the reference screen. The other screens operated for periods of at least two years. Three major types of screens can be distinguished from Table I: Stevenson screens, multi-plate screens, and the Young aspirated screen. The latter screen has often been suggested as an ideal reference screen.

The operational measurement uncertainty of the sensors is 0.1 °C. Owing to the calibration described in part I, we were able to obtain inter-sensor accuracies of about 0.03 °C. Temperature is sampled every 15 s. Unless stated otherwise, we use the 10-min mean temperature values of these samples. Besides temperature, we use the following operationally measured elements at the KNMI-terrain: horizontal wind speed u at 10 m until 26 June 1993 and thereafter at 20 m, global radiation or total short-wave radiation received at the surface $K\downarrow$, cloud cover N, and relative humidity rh.

The terrain roughness is large, especially for directions between southeast and north, where 15–30 m high trees are found at distances ranging between 80 and 220 m from the measurement site. Operational wind measurements at the terrain of KNMI are therefore taken about 200 m east from the experimental site and the other operationally measured elements. As a result, the wind measurements may not always be representative of the actual wind speed at screen height near the experimental site.

2.2. Methodology

For most climatological applications the main interest is in the differences in daily minimum (T_n), daily maximum (T_x), and daily mean (T_{mean}) temperature. These differences will be presented first in Section 3. For each screen, we calculated the temperature differences ΔT (screen – Knmi.ref) for T_n, T_x, *and* T_{mean} for days with complete data for both the screen and Knmi.ref. T_n

DOI: 10.1002/joc

Table I. Details of screens and sensors. The Stevenson screens are of KNMI design.

Screen	Abbreviation	Start date	End date	Diameter (m)	Ventilation	Sensor
KNMI multi-plate	Knmi.ref	89/01/09	95/02/01	0.30	Natural	Pt500
KNMI multi-plate aspirated	Knmi.asp	90/12/13	93/02/20	0.30	Aspirated (1 dm^3/min)	Pt500
Vaisala multi-plate DTR11	Vaisala	89/01/09	93/02/20	0.30	Natural	Pt500
Young Gill multi-plate 41002	Young	89/01/09	93/02/20	0.12	Natural	Pt500
Young aspirated type II 43408	Young.aspII	92/08/18	95/02/01	0.15/0.025	Aspirated (0.1 dm^3/s)	Pt1000[1]
Socrima multi-plate BMO 1167A	Socrima	91/03/08	95/02/01	0.20	Natural	Pt500
Stevenson PVC	Stev.pvc	89/01/09	91/03/06	0.70	Natural	Pt500
Stevenson PVC aspirated	Stev.pvc.asp	91/03/07	95/02/01	0.70	Aspirated 1 dm^3/s)	Pt500
Stevenson wood	Stev.wood	89/01/09	93/02/20	0.70	Natural	Pt500

[1] For the Young aspirated screen, the sensor is an integral part of the measuring device. The first value for the diameter of these screens refers to the diameter of the radiation shield, the second to the tube that houses the sensor.

and T_x are calculated for each screen separately and for each day as the minimum and maximum, respectively, of the 144 non-sliding 10-min temperatures. T_{mean} is calculated as the mean of these 10-min temperatures. A day runs from 0–24 UTC.

After the presentation of results for T_n, T_x, *and* T_{mean}, the differences in the daily cycle for each season were studied for four combinations of windspeed and cloudiness: I: $u \leq 3.5$ m/s $\cap$ $N \leq 4/8$; II: $u > 3.5$ m/s $\cap$ $N \leq 4/8$; III: $u \leq 3.5$ m/s $\cap$ $N \geq 5/8$; and IV: $u > 3.5$ m/s $\cap$ $N \geq 5/8$, where N equals the fraction of cloud cover (originally in octas).

A fact that is often overlooked is the effect of random temperature variations on the calculation of T_n and T_x. The original temperature samples are influenced by turbulent eddies with a wide range of space and length scales. Those eddies are advected along the screens causing deviations from the mean temperature. It is obvious that fast-responding screens follow better these turbulence driven temperature deviations than slower ones. In practice, T_n and T_x are operationally calculated from sliding L-min averages of the temperature samples. The turbulence driven temperature deviations cause the L-min averages of fast-responding screens to fluctuate more around the mean temperature than those of slow-responding screens. As a result, fast-responding screens have a higher probability of obtaining more extreme T_n and T_x than slow-responding screens (disregarding other causes of temperature differences, like radiation errors). This effect was studied by comparing the screens for several lengths of the averaging interval using the original 15-s data.

Parallel measurements like those presented here can be used in climatology for two main reasons: (1) to assess the magnitude of the bias for the transition from an old screen to a new screen, and (2) to derive a transfer function that transforms the temperatures measured with the old screen into a series measured with the new screen, or the other way around. Here we derive such a transfer function. In Part I, we demonstrated the importance of lag-time differences, windspeed, and cloudiness in explaining temperature differences between screens. Here we use that information and the results in the present paper for the derivation of the transfer function. Stev.pvc and Knmi.ref are used here as an example.

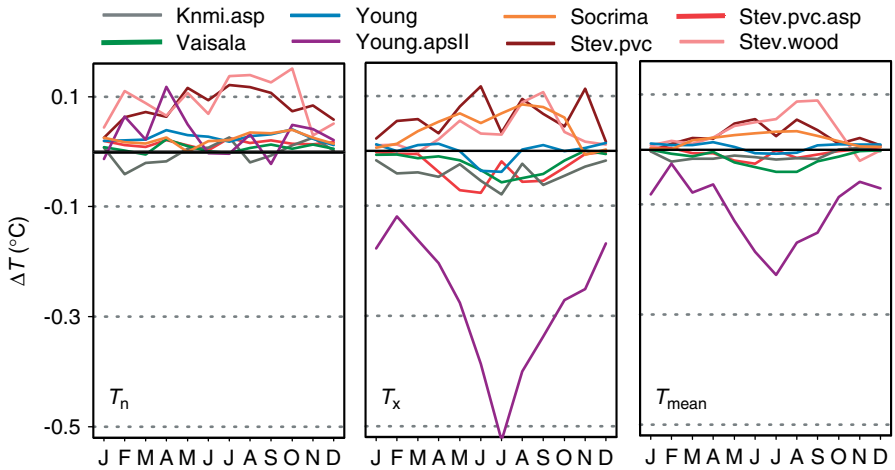

Figure 1. Monthly mean temperature differences ΔT (screen – Knmi.ref) for daily minimum temperature T_n, maximum temperature T_x, and mean temperature T_{mean}. The screens and their period of overlap with Knmi.ref are defined in Table I.

DOI: 10.1002/joc

Table II. Mean temperature differences ΔT (screen – Knmi.ref) for winter (DJF), spring (MAM), summer (JJA), and autumn (SON) for daily minimum temperature T_n, maximum temperature T_x, and mean temperature T_{mean}. The screens and their period of overlap with Knmi.ref are defined in Table I. The values in brackets give the standard error.

	ΔT_n (°C)				ΔT_x (°C)				ΔT_{mean} (°C)			
	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON	DJF	MAM	JJA	SON
Knmi.asp	*−0.009* (0.007)	*−0.010* (0.008)	*0.004* (0.006)	*0.006* (0.008)	−0.024 (0.004)	−0.036 (0.006)	−0.052 (0.006)	−0.049 (0.006)	−0.009 (0.002)	−0.015 (0.002)	−0.016 (0.002)	−0.008 (0.002)
Vaisala	0.005 (0.002)	0.010 (0.003)	*0.002* (0.003)	0.010 (0.003)	−0.007 (0.002)	−0.013 (0.004)	−0.047 (0.005)	−0.021 (0.005)	−0.003 (0.001)	−0.014 (0.001)	−0.038 (0.001)	−0.013 (0.001)
Young	0.018 (0.003)	0.031 (0.005)	0.024 (0.005)	0.032 (0.005)	0.009 (0.004)	*0.007* (0.005)	−0.024 (0.006)	*0.005* (0.005)	0.009 (0.001)	0.009 (0.002)	−0.007 (0.002)	0.009 (0.001)
Young.aspII	*0.019* (0.015)	0.065 (0.016)	*0.010* (0.015)	*0.022* (0.012)	−0.159 (0.015)	−0.220 (0.017)	−0.433 (0.021)	−0.286 (0.019)	−0.064 (0.009)	−0.093 (0.009)	−0.192 (0.007)	−0.099 (0.008)
Socrima	0.019 (0.003)	0.013 (0.004)	0.025 (0.004)	0.033 (0.004)	*0.006* (0.004)	0.054 (0.004)	0.068 (0.005)	0.048 (0.007)	*0.002* (0.001)	0.022 (0.002)	0.031 (0.002)	0.016 (0.002)
Stev.pvc	0.046 (0.009)	0.085 (0.010)	0.111 (0.012)	0.088 (0.011)	0.032 (0.013)	0.057 (0.011)	0.081 (0.013)	0.070 (0.014)	*0.004* (0.003)	0.030 (0.004)	0.044 (0.005)	0.022 (0.004)
Stev.pvc.asp	0.017 (0.003)	0.014 (0.004)	0.014 (0.005)	0.016 (0.004)	*−0.003* (0.004)	−0.042 (0.005)	−0.051 (0.011)	−0.032 (0.005)	0.003 (0.001)	−0.012 (0.002)	−0.015 (0.004)	−0.004 (0.001)
Stev.wood	0.065 (0.012)	0.086 (0.012)	0.115 (0.013)	0.101 (0.013)	*0.011* (0.008)	0.025 (0.009)	0.049 (0.010)	0.053 (0.012)	*0.006* (0.003)	0.025 (0.003)	0.063 (0.004)	0.034 (0.006)

* Values in italics are not significantly different from zero (2 × se).

3. Results

3.1. Mean and extreme differences in daily T_n, T_x, and T_{mean}

Figure 1 shows the monthly mean ΔT values. The most striking feature is the anomalous behaviour of ΔT_x for Young.aspII (also reflected in T_{mean}), with the largest deviation from Knmi.ref in July when T_x of Young.aspII is on average 0.5 °C lower than T_x of Knmi.ref. The temperatures for the multi-plate screens are all close to Knmi.ref. Socrima, however, has a tendency to be a little warmer than the other multi-plate screens, which may be related to reduced natural ventilation within this screen compared to the other multi-plate screens. The natural ventilated Stevenson screens (Stev.pvc, Stev.wood) are warmer than Knmi.ref both during day and night. The monthly mean ΔT_n and ΔT_x are largest in the summer half year, but mostly do not exceed 0.1 °C. The results for Stev.pvc.asp show that the effect of aspiration is that the Stevenson screen more closely resembles the multi-plate screens compared to the unventilated version of the screen (Stev.pvc).

The ΔT values for each screen are presented in Table II (seasonal average values) and Table III (annual average values). From the 96 average ΔT values in Table II, 80 are significantly different from zero (because of serial correlation in the difference series within a season, the standard errors may go up to 1.5 times larger than those presented). However, only 37 of these values are larger than the inter-sensor accuracy (0.03 °C). As may already be inferred from Figure 1, the largest ΔT values occur for Young.aspII, Stev.pvc, and Stev.wood. The annual average ΔT values in Table III are small for these three screens but not always negligible. For ΔT_n they range between 0.029 °C (Young.aspII) and 0.092 °C (Stev.wood); for ΔT_x between −0.275 °C (Young.aspII) and 0.060 °C (Stev.pvc); and for ΔT_{mean} between −0.112 °C (Young.aspII) and 0.032 °C (Stev. wood).

Additional information about the distribution of the individual daily ΔT values can be obtained from the

Table III. Annual mean temperature differences ΔT (screen – Knmi.ref) for daily minimum temperature T_n, maximum temperature T_x, and mean temperature T_{mean}. The screens and their period of overlap with Knmi.ref are defined in Table I. The values in brackets give the standard error.

	ΔT_n (°C)	ΔT_x (°C)	ΔT_{mean} (°C)
Knmi.asp	*−0.002* (0.004)	−0.040 (0.003)	−0.012 (0.001)
Vaisala	0.007 (0.001)	−0.022 (0.002)	−0.017 (0.000)
Young	0.026 (0.002)	*−0.001* (0.003)	0.005 (0.001)
Young.aspII	0.029 (0.007)	−0.275 (0.009)	−0.112 (0.004)
Socrima	0.023 (0.002)	0.044 (0.003)	0.018 (0.001)
Stev.pvc	0.082 (0.005)	0.060 (0.006)	0.025 (0.002)
Stev.pvc.asp	0.015 (0.002)	−0.032 (0.003)	−0.007 (0.001)
Stev.wood	0.092 (0.006)	0.035 (0.005)	0.032 (0.002)

* Values in italics are not significantly different from zero (2 × se).

DOI: 10.1002/joc

boxplots in Figure 2. Consider for example the boxplot of ΔT_n of Stev.wood for autumn (SON). This boxplot shows that the ΔT_n distribution is positively skewed with the 90th percentile equal to 0.33 °C (in 10% of the days ΔT_n is larger than this number). A result of the positive skewness in this case is that the average ΔT_n of 0.10 °C (Table II) is 0.04 °C larger than the median (50th percentile) of 0.06 °C shown in the boxplot. The figure shows that most of the ΔT distributions of Stev.pvc and Stev.wood exhibit this positive skewness. Figure 2 also shows that the distribution of ΔT values for T_{mean} is much narrower than that for T_n, T_x. For (almost) identical screens, the boxplots may also provide information about the spread of the ΔT values that may be expected because of pure random variation alone. In this experiment, Vaisala is more or less identical to Knmi.ref and the corresponding statistics of the ΔT values between these screens could be considered as a measure for natural background noise.

3.2. Diurnal temperature cycle and the effect of wind and cloudiness

For each screen, the annual mean diurnal cycle of the temperature differences ΔT (screen – Knmi.ref) was calculated with an hourly resolution. Figure 3 shows these cycles for all screens together with the mean annual diurnal temperature cycle of Knmi.ref. For Stev.pvc and Stev.wood, it appears that the largest absolute values of ΔT are just after sunrise and just after sunset. Because of their relatively large response times, these screens have more difficulties in following the steep temperature changes at those times than the other screens (see also Part I). For Young.aspII, the largest ΔT is around the time of maximum solar radiation. Figure 4 shows the same diurnal cycle differences but now for the four combinations of u and N defined in Section 2.2 for both winter (DJF) and summer (JJA). The figure clearly shows that the temperature differences between screens depend on u and N. Small values of u and N (category I) result in large mutual differences in the diurnal cycles while large values (category IV) minimize those differences. It is interesting to see that the large absolute values of ΔT for Stev.pvc and Stev.wood, just after sunrise and just after sunset, almost disappear for $u > 3.5$ m/s. It is also noteworthy that for large values of u and N (category IV) the daytime bias of Young.aspII remains relatively large (about −0.1 °C in winter and −0.2 °C in summer).

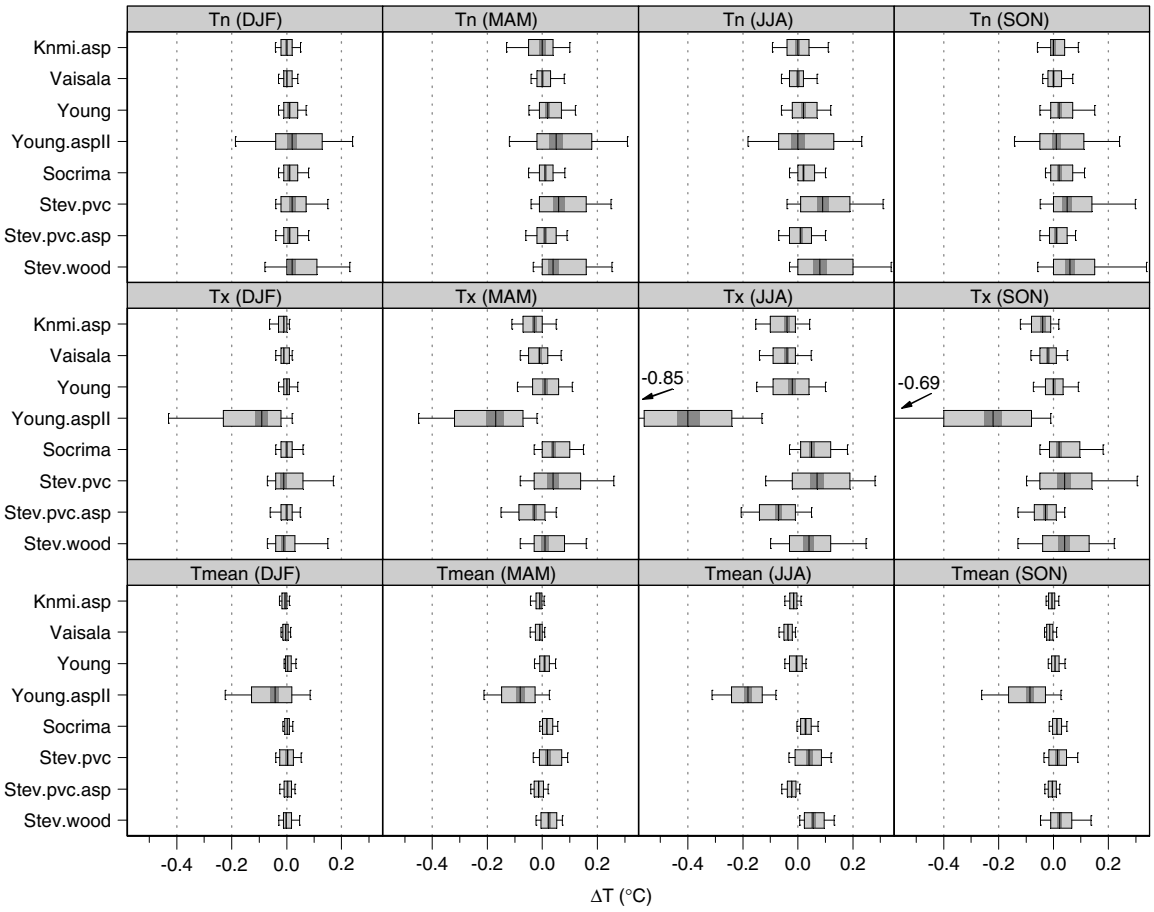

Figure 2. Boxplots of the individual temperature differences ΔT (screen – Knmi.ref) for winter (DJF), spring (MAM), summer (JJA), and autumn (SON) for daily minimum temperature T_n, maximum temperature T_x, and mean temperature T_{mean}. The screens and their period of overlap with Knmi.ref are defined in Table I. The left and right limits of the box represent the 25th/75th percentiles (quartiles); the vertical line within the box represents the 50th percentile (median) with 95% confidence interval (dark gray); and the whiskers mark the 10th/90th percentiles. Two of the Young.aspII whiskers are outside the horizontal scale; the corresponding values are shown in the figure with arrows pointing to the location of the whiskers.

DOI: 10.1002/joc

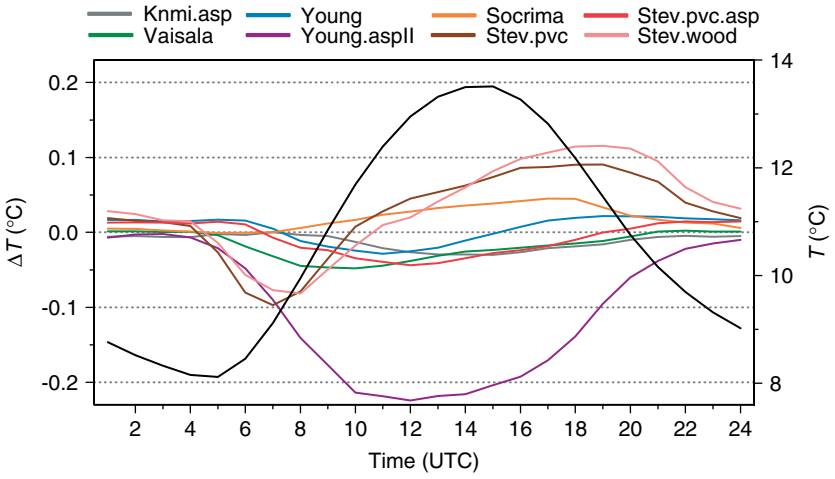

Figure 3. Annual mean diurnal cycles of the hourly temperature differences ΔT (screen – Knmi.ref) of the eight screens (left axis) and of the hourly temperatures of Knmi.ref (black line, right axis). The screens and their period of overlap with Knmi.ref are defined in Table I.

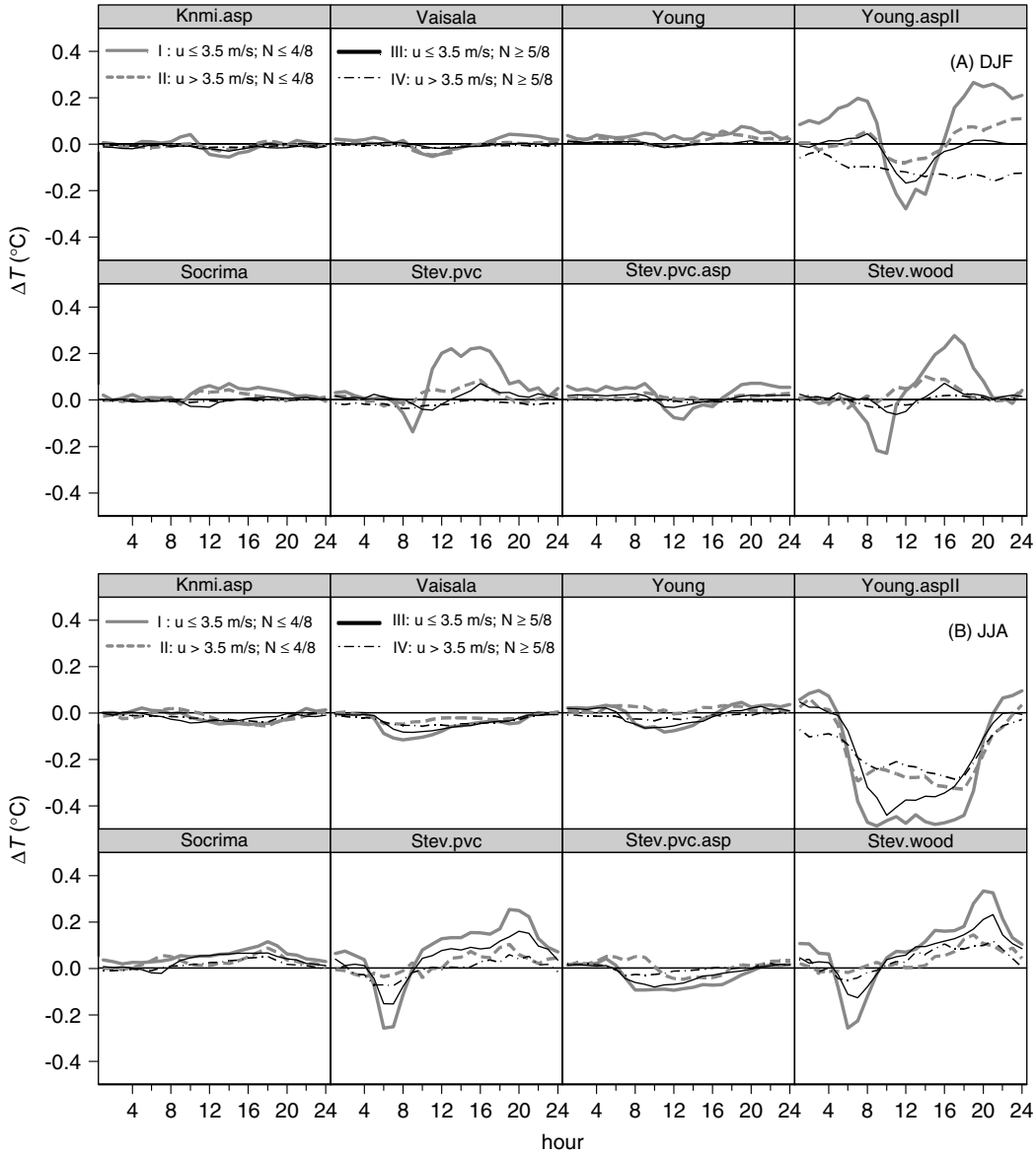

Figure 4. Mean diurnal cycles of the hourly temperature differences ΔT (screen – Knmi.ref) for four combinations of windspeed (u) and fraction of cloudiness (N). The upper two rows (A) apply to the three winter months (DJF) and the lower two rows (B) to the three summer months (JJA). The screens and their period of overlap with Knmi.ref are defined in Table I.

DOI: 10.1002/joc

3.3. Effect of averaging interval length L on T_n and T_x

The effect of the averaging interval length L on T_n and T_x is studied by calculating T_n and T_x for $L = 0.25, 1, 2, \ldots,$ 15 min, where $L = 0.25$ min corresponds to the original 15-s temperature samples (no averaging). T_n and T_x are selected as the smallest and largest values, respectively, from the sliding L-min temperature averages in a day ($24 \times 60 \times 4 = 5760$ values). For the presentation of the results, we restrict ourselves to two thirty-day summer periods, without missing values, in 1989 (July 1–30) and in 1994 (June 17–July 16). For each L, T_n and T_x are calculated as the average of the 30 daily values. To facilitate the comparison between the screens, we subtracted for each screen the mean T_n and T_x values for all L. The resulting anomalies are denoted as T_n^* and T_x^*.

Figure 5 shows the results for July 1989 for both T_n^* (left) and T_x^* (right). The figure clearly shows that the modern type round multi-plated screens are more sensitive to changes in L than the Stevenson screens and that these effects are much larger for T_x^* than for T_n^*. For instance, shifting from $L = 10$ min to $L = 1$ min results in an increase in T_x of Knmi.ref of 0.33 °C, whereas for Stev.pvc the increase amounts to only 0.13 °C. For T_n, the corresponding values are a decrease in T_n of Knmi.ref of 0.10 °C and a decrease of Stev.pvc of 0.03 °C.

Figure 6 shows the results in 1994 for the period between June 17 and July 16 . Compared to the July 1989 period, the absolute values for Knmi.ref are somewhat smaller here. Noteworthy are the large (absolute) values for Young.aspII and (to a smaller extent) of Stev.pvc.asp in comparison to Knmi.ref. In January (not shown) the effects of changing L are about a factor 3–4 smaller than in July. Note also the relatively small absolute values for Socrima compared to Knmi.ref.

The implication of the above is that fast-responding screens have by nature more extreme T_x and T_n than slow-responding screens and they are more sensitive to changes in L. This effect may result in differences of several tenths of °C between the T_x and T_n of those screens. In practice, this effect has to be considered in connection with other effects (e.g. the larger artificial heating of natural ventilated Stevenson screens compared to modern screens).

3.4. Transfer function for a transition of Stev.pvc to Knmi.ref

A supposed transition of Stev.pvc to the faster-responding Knmi.ref is used here to demonstrate the construction of transfer functions. Stev.pvc is selected because its temperatures show significant differences with those of Knmi.ref and because it was in use for an almost uninterrupted period of nearly two years without major problems becoming visible in the calibration (see also Part I). The period of simultaneous measurements is used to derive a transfer function that enables a user to transform the new measurements into the old ones or the other way around. We consider two transfer functions: (1) for the 10-min temperatures, and (2) for T_n and T_x. In both cases, we first explored the relationships between the predictant and all possible predictors for the

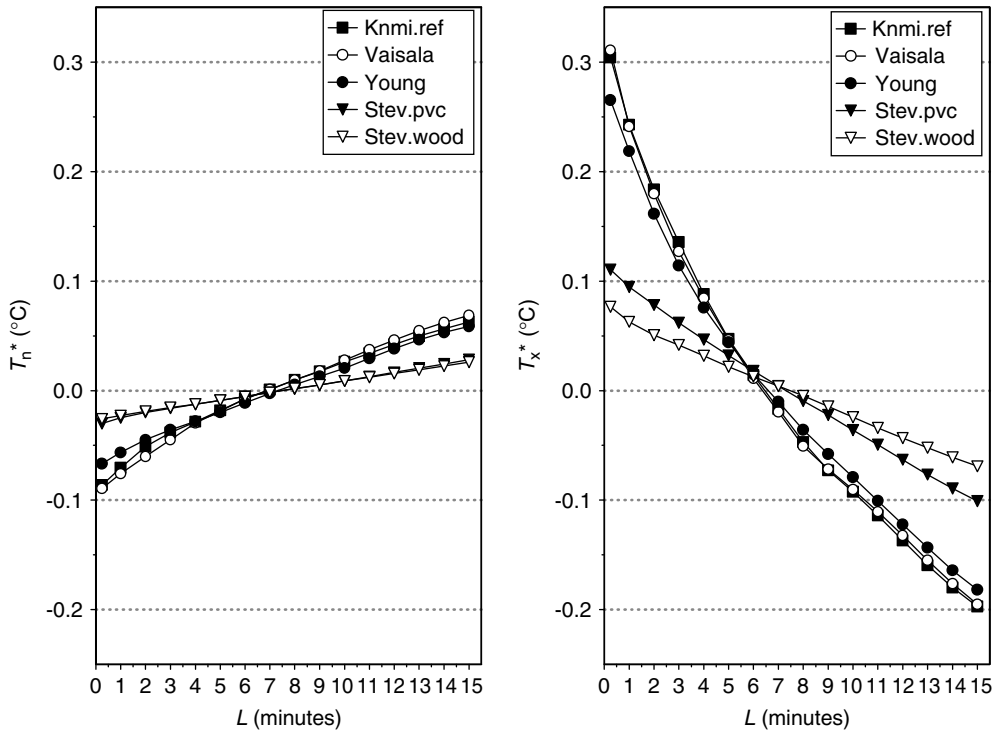

Figure 5. Mean daily T_n^* (left) and T_x^* (right) as a function of the averaging interval for July 1989. For each screen, the mean T_n and T_x values for all L are subtracted to obtain T_n^* and T_x^*. Each daily T_n and T_x is calculated from 5760 overlapping (15 s) intervals of length $L = 0.25, 1, 2, \ldots,$ 15 min.

DOI: 10.1002/joc

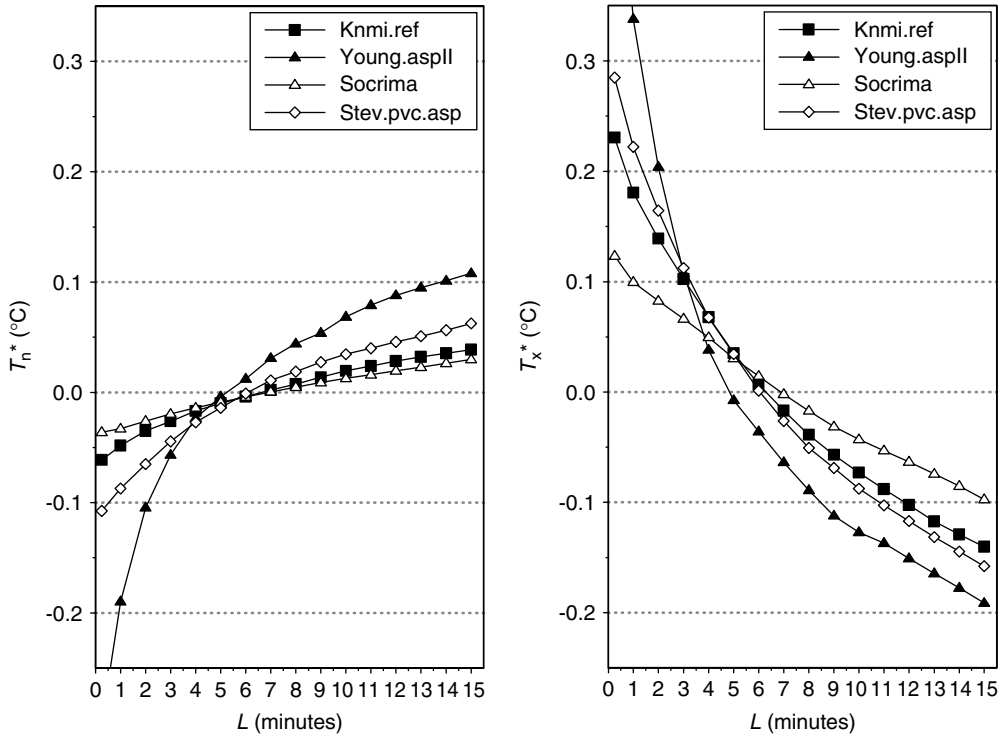

Figure 6. As in Figure 5, but showing the period ranging from June 17–July 16, 1994. The $L = 0.25$ min values for Young.aspII fall out of the paper and equal $-0.30\,^{\circ}$C in the left plot and $0.57\,^{\circ}$C in the right plot.

presence of non-linear effects and interaction using non-parametric loess functions (Cleveland and Devlin, 1988). After fitting the equations, we examined the relationships between the error series and the predictors using the same procedure.

3.4.1. *Model for 10-min data*

For the 10-min data we propose the following equation:

$$\Delta T_i = a_0 + a_1 \Delta T_{i-1} + a_2 \frac{(dT_{ref}/dt)_i}{(u_i + 1)} + a_3 \frac{K{\downarrow}_i/100}{(u_i + 1)} + a_4 \frac{\min(1 - N_i, I_i)}{(u_i + 1)} + \varepsilon_i$$

$$I_i = \begin{cases} 0 & \text{during daytime} \\ 1 & \text{during nighttime} \end{cases} \qquad (1)$$

where i is the counter of the 10-min interval considered ($i \in [1, 2, \ldots, M]$) with M the total number of 10-min intervals, ΔT_i equals the temperature difference $T(\text{Stev.pvc}) - T(\text{Knmi.ref})$ in °C, for the ith 10-min interval, dT_{ref}/dt is the derivative of the temperature curve of Knmi.ref at time t which is defined for the ith 10-min interval as $T(\text{Knmi.ref})_i - T(\text{Knmi.ref})_{i-1}$, u_i is windspeed for the ith 10-min interval in m/s, $K{\downarrow}_i$ global radiation for the ith 10-min interval in W/m^2 which equals zero during nighttime by definition, N_i is the cloudiness fraction for the ith 10-min interval ($N \in [0, 1/8, 2/8, \ldots, 8/8]$), ε_i is the residual for the ith 10-min interval in °C assumed to be independent, normally distributed and zero mean, and $a_0, \ldots, a_4$ are coefficients to be estimated. The two terms with coefficients a_1 and a_2, on the right-hand side of Equation (1), account for the slow response of Stev.pvc compared to Knmi.ref. The next two terms account for the fact that for clear-sky conditions, observed inter-screen temperature differences may be larger (in an absolute sense) than for cloudy conditions. The division by $u_t + 1$ in the terms involving T_{ref}, $K{\downarrow}$, and N_t is needed because windspeed dampens their effects on ΔT, the $+1$ m/s is added to account for the fact that anemometers have a threshold speed and to prevent division by zero. As $K{\downarrow}$ may take values as large as 1000 W/m^2, it is divided by 100 to facilitate the intercomparison of the coefficients.

The coefficients of the model in Equation (1) have been estimated for each season using ordinary least-squares regression (OLS). Table IV presents the results. For all seasons, the intercept term $\hat{a}_0$ is significant but nearly zero. The table further shows that there is a strong contribution of the autoregressive term ΔT_{t-1} for each season. The relatively large negative values for $\hat{a}_2$ demonstrate that Stev.pvc is not able to follow the fast temperature fluctuations of Knmi.ref. The clear-sky effect on ΔT is significant and seasonally dependent both during day and night. Note that the effects are opposite (relative warming of Stev.pvc during the day and relative cooling during the night) and their potential magnitudes are about an order of magnitude larger during the day (e.g. with conditions in summer with the maximum $K{\downarrow}$ about 1000 W/m^2 and $u = 0$ m/s) than during the night (for conditions with $N = 0$ and $u = 0$ m/s).

Table IV shows that the explained variance is large and ranges between 81.5% and 87.1% for spring and autumn, respectively. The results do not improve by applying

DOI: 10.1002/joc

Table IV. Estimated regression coefficients ($\hat{a}_0, \ldots, \hat{a}_4$), number of observations (n), residual standard error (RSE), and the coefficient of determination (R^2) for the model in Equation (1) for winter (DJF), spring (MAM), summer (JJA) and autumn (SON) for the 10-min temperature differences ΔT (Stev.pvc−Knmi.ref) in °C. The values within brackets are the t-statistics of the parameter estimates above.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	n	RSE	R^2
DJF	−0.0093 (−23.6)	0.5918 (208.1)	−1.0378 (−181.5)	0.1682 (80.4)	−0.0233 (−9.0)	27005	0.051	0.816
MAM	−0.0084 (−11.1)	0.5363 (182.0)	−1.0916 (−189.4)	0.0647 (64.2)	−0.0631 (−17.9)	25084	0.082	0.815
JJA	−0.0113 (−13.1)	0.5660 (200.6)	−1.0626 (−187.7)	0.0613 (61.7)	−0.0424 (−11.3)	25962	0.092	0.828
SON	−0.0079 (−12.9)	0.6133 (239.6)	−1.0410 (−199.9)	0.0831 (64.4)	−0.0393 (−13.1)	23703	0.069	0.871

a power transformation to $(u + 1)$ in Equation (1). The residual series show some autocorrelation with the first-order autocorrelation coefficient ranging between 0.17 in summer and 0.26 in autumn. As a result, the statistical significance of the coefficients in Table IV will in fact be somewhat smaller than the calculated values.

The large values for the explained variance are partly a result of using observed values of ΔT_{i-1} in Equation (1). When we use the simulated $\Delta \hat{T}_{i-1}$ instead (using the estimated regression coefficients $\hat{a}_0, \ldots, \hat{a}_4$), the explained variance decreases to values between 56.9% and 66.4% for winter and autumn, respectively. These values do not further degrade when we use the first part of the data (1989) for model calibration and the second part (1990) for model verification.

To check whether model or data deficiencies may (partly) be responsible for the decrease in explained variance, we repeated the experiment using simulated ΔT values. First, we simulated a new series of random normally distributed errors ε_i in Equation (1) using the calculated mean and standard deviation of ε_i. Second, the simulated ε_i series was used to simulate a new ΔT_i series using the estimated regression coefficients $\hat{a}_0, \ldots, \hat{a}_4$) in Equation (1), and the measured values of the predictors. Finally, Equation (1) was fitted with the new simulated ΔT_i series (and the accompanying ΔT_{i-1} series) resulting in a new set of estimated regression coefficients $\hat{a}_0, \ldots, \hat{a}_4$. The explained variance is, as it should be, almost the same as before and ranges between 80.0% and 86.7%. When again using the simulated $\Delta \hat{T}_{i-1}$ instead of ΔT_{i-1} (using the newly estimated regression coefficients $\hat{a}_0, \ldots, \hat{a}_4$), the explained variance now decreases much less than before to values between 70.8% and 79.0% for winter and autumn, respectively. These results indicate that model and/or data deficiencies may indeed (partly) be responsible for the decrease in explained variance. Examination of the relationships between the error series of all possible predictors does not reveal serious interaction or non-linear effects. We therefore suggest that the lack of adequate windspeed data at screen height at the experimental site is partly responsible for the decrease in explained variance (windspeed is strongly represented in Equation (1)).

3.4.2. *Model for T_n and T_x data*

At the times t_n of the minimum temperature and t_x of the maximum temperature, the first-order derivatives of the temperature curve equal zero and the effects of differences in response times of the screens on ΔT should be minimal. However, in Section 3.3 we have shown that the faster responding screens have a higher probability to reach more extreme T_n or T_x than the slow-responding screens. To account for this, we define a new variable reflecting the sharpness of the dale or peak in the temperature curve at t_n and t_x of the reference screen:

$$DPS_d = T_{ref,d,i} - \frac{T_{ref,d,i-1} + T_{ref,d,i+1}}{2} \qquad (2)$$

where DPS_d stands for the Dale or Peak Sharpness of Knmi.ref at t_n or t_x at day d and $T_{ref,d,i}$ is the temperature of Knmi.ref at day d in the ith 10-min time interval corresponding to t_n or t_x. The larger the absolute value of DPS, the greater the discrepancy between Knmi.ref and the slower responding Stev.pvc. For T_n the following equation is used:

$$\Delta T_{n,d} = b_0 + b_1 \frac{DPS_d}{(u_d + 1)} + b_2 \frac{(1 - N_d)}{(u_d + 1)} + \varepsilon_d \qquad (3)$$

where u_d and N_d are the 10-min windspeed and cloudiness at t_n for day d, ε_d is assumed to be an independent normally distributed zero mean residual, and b_0, b_1, b_2 are coefficients to be estimated. Although Equation (3) is in principle set up for nighttime t_n, it also works for the scarce daytime t_n values.

For T_x the following equation is used:

$$\Delta T_{x,d} = c_0 + c_1 DPS_d + c_2 \frac{K\downarrow_d /100}{(u_d + 1)} + \varepsilon_d \qquad (4)$$

where u_d and $K\downarrow_d$ are the windspeed and global radiation at t_x for day d, ε_d is assumed to be an independent normally distributed zero mean residual, and c_0, c_1, c_2 are coefficients to be estimated. Inclusion of nighttime cloudiness, for nighttime t_x, is not significant. Division of DPS_i by $(u + 1)$ is not advantageous here. Note that t_n and t_x of Knmi.ref and Stev.pvc need not be exactly the same. The weather variables in Equations (3) and (4) are at t_n and t_x of Knmi.ref.

The coefficients of the models in Equations (3) and (4) have been estimated for each season using OLS. If coefficients turned out to be not significant at the 5% level, the corresponding predictors were omitted from the

DOI: 10.1002/joc

equation, starting with the predictor with the smallest t-statistic. After the deletion of an insignificant predictor, the remaining coefficients were reestimated. Tables V and VI present the results for T_n and T_x, respectively. Instead of the coefficient of determination (R^2), the tables give the squared correlation coefficient r^2 between the predictant and the fitted values. The latter coefficient is preferred when models with and without intercept are compared (for models with intercept $R^2 = r^2$).

Tables V and VI show that the intercept term is only significant for ΔT_x in spring. The coefficients $\hat{b}_1$, $\hat{b}_2$, $\hat{c}_1$, $\hat{c}_2$ show some seasonal variation. Note that for ΔT_n the contribution of the term with cloudiness is not significant in both winter and autumn. For ΔT_n, the *DPS* term has a relatively large contribution. For instance, in summer *DPS* at t_n ranges between -0.7 and 0 resulting in a maximum contribution of $+0.48\,^{\circ}$C (with $u = 0$ m/s) to ΔT_n. This is large compared to the maximum contribution of cloudless skies (with $u = 0$ m/s) of about $+0.12\,^{\circ}$C in summer. In contrast, for ΔT_x in summer the clear-sky effect has the largest contribution of about $+1.5\,^{\circ}$C ($K\downarrow = 1000$ W/m^2 and $u = 0$ m/s). Here *DPS* at t_x ranges between 0.0 and 0.9 resulting in a maximum contribution of $-0.25\,^{\circ}$C (with $u = 0$ m/s) to ΔT_x.

The r^2 in Tables V and VI are relatively low and range between 0.27 and 0.53. We suggest that the lack of adequate windspeed data at screen height at the experimental site is even more important here than for the 10-min model of Equation (1). Another factor could be the relatively large natural variation of T, which cannot be explained by a model. To obtain an idea of the magnitude of this factor, we can take the difference between Knmi.ref and Vaisala as a measure for the natural variation (Knmi.ref and Vaisala are almost identical screens). The standard deviation of ΔT_n for these screens varies between $0.035\,^{\circ}$C in winter and $0.056\,^{\circ}$C in summer. For ΔT_x, the corresponding values are $0.037\,^{\circ}$C in winter and $0.077\,^{\circ}$C in summer. Comparing these values with the residual standard error (RSE) in Tables V and VI shows that there is still enough room for improvement.

4. Discussion and conclusions

In this paper, we compared the temperatures of nine thermometer screens that are currently in use around the world. The comparison was based on measurements taken during a 6-year field experiment in De Bilt (the Netherlands). All screens were situated close to each other on the same experimental site and faced the same weather and environmental conditions. Furthermore, the screens were all supplied with the same type of temperature sensor.

The results indicate that the seasonal differences in T_n, T_x, and T_{mean}, with respect to the reference screen Knmi.ref, are small (generally $\leq 0.1\,^{\circ}$C) for most screens. This is in agreement with other studies that intercompared similar screens (e.g. Andersson and Mattisson, 1991; Larre and Hegg, 2002). For the artificially ventilated Young.aspII, the seasonal differences in T_n, T_x, and T_{mean} are much larger than for all other screens. The value for ΔT_x is largest and varies between $-0.159\,^{\circ}$C (DJF) and $-0.433\,^{\circ}$C (JJA). Also the annual mean ΔT_{mean} is largest for Youngs.apsII and amounts to $-0.112\,^{\circ}$C whereas for all other screens the annual mean absolute ΔT_{mean} is not greater than $0.032\,^{\circ}$C (Stev.wood). The latter value may be considered small with respect to the observed global temperature rise of the last 150 years.

Table V. Estimated regression coefficients ($\hat{b}_0, \ldots, \hat{b}_2$), number of observations (n), residual standard error (RSE), and the squared correlation coefficient (r^2) for the model in Equation (3) for winter (DJF), spring (MAM), summer (JJA), and autumn (SON) for the daily minimum temperature differences ΔT_n (Stev.pvc–Knmi.ref). The values within brackets are the t-statistics of the parameter estimates above.

	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$	n	RSE	r^2
DJF	–	−1.2223 (−12.0)	–	171	0.096	0.381
MAM	–	−0.8018 (−6.9)	0.0741 (2.0)	171	0.113	0.306
JJA	–	−0.6838 (−6.7)	0.1207 (3.3)	173	0.130	0.276
SON	–	−1.1326 (−17.8)	–	163	0.099	0.533

Table VI. Estimated regression coefficients ($\hat{c}_0, \ldots, \hat{c}_2$), number of observations (n), RSE, and the squared correlation coefficient (r^2) for the model in Equation (4) for winter (DJF), spring (MAM), summer (JJA), and autumn (SON) for the daily maximum temperature differences ΔT_x (Stev.pvc–Knmi.ref). The values within brackets are the t-statistics of the parameter estimates above.

	$\hat{c}_0$	$\hat{c}_1$	$\hat{c}_2$	n	RSE	r^2
DJF	–	−0.4458 (−5.9)	0.5722(12.2)	171	0.126	0.459
MAM	0.0584 (2.6)	−0.3566 (−6.6)	0.1128 (5.5)	170	0.128	0.271
JJA	–	−0.2826 (−7.7)	0.1531 (12.4)	173	0.134	0.363
SON	–	−0.3648 (−6.9)	0.2695 (11.1)	163	0.144	0.349

DOI: 10.1002/joc

When looking at extremes, the daily values of ΔT_n, ΔT_x, and ΔT_{mean} are important. The results show that these values may be much larger than the seasonal averages and that the studied screens can be roughly divided into three major groups: (1) Knmi.asp, Vaisala, Young, Socrima, and Stev.pvc.asp; (2) Stev.pvc and Stev.wood; and (3) Young.aspII. The screens in the first group behave more or less the same as Knmi.ref, the daily differences are mainly a result of natural fluctuations. The unventilated Stevenson screens of the second group show much more variation in the daily values than the screens in the first group and their values display positive skewness. The Young.aspII that constitutes the third group shows the largest variation in the daily values with a large negative bias and negative skewness of ΔT_x in all seasons. These findings are also reflected in the diurnal cycle differences for winter and summer for four combinations of windspeed and cloudiness.

It is shown that for the calculation of T_n and T_x from the 15-s samples the length of the measurements interval L may be an important factor. Slow responding screens, like the natural ventilated Stevenson screens, have less extreme T_n and T_x values than fast-responding screens. The reason is the smaller variation of the 15-s samples in the sliding L-min intervals for the slow-responding screens compared to the fast-responding screens. For T_x this effect is a factor 3–4 larger than for T_n. In practice, when changing from an old screen to a new screen, this effect will often partly compensate for other effects. For instance, a change from Stev.pvc to Knmi.ref or Young.aspII may result in a lowering of T_x because Knmi.ref and Young.aspII are less sensitive to radiation errors than Stev.pvc. The above-mentioned effect partly compensates for this lowering (depending on the magnitude of L). The compensation effect may be even larger when the old screen is supplied with a mercury thermometer and the new screen with a fast-responding sensor like the platinum resistance sensor (e.g. the Pt500). More research is needed to estimate the magnitude of these effects in that case.

We constructed two transfer functions that transfer Knmi.ref into Stev.pvc, one for the 10-min mean temperature values and one for the daily T_n and T_x. We have shown that it is feasible to derive such functions, although it was felt that the lack of adequate windspeed data at screen height near the experimental site hampered the derivation.

Windspeed is an important factor in the transfer functions, with its relative importance increasing with decreasing windspeed. Laboratory experiments show that, e.g. the radiation error under conditions of snow cover may increase to about 8 °C for windspeeds close to 0 m/s (Gill, 1983). Brandsma (2004) showed the importance of small windspeeds on temperature differences for even slight changes in location of the screen. Measurements of high-accuracy small windspeeds at screen height will therefore be very helpful in developing improved weather-dependent transfer functions in case of future inhomogeneities. The extent of this deserves further study. In addition, long records of such windspeeds may allow for objectively monitoring the local roughness of temperature sites. We therefore recommend that temperature stations of climatological interest will be equipped with anemometers for high-accuracy (especially in the range 0–3 m/s) wind speed measurements, placed near the thermometers' screen at screen level. These measurements could be performed in addition to the current operational windspeed measurements (mostly at 10-m height as recommended by the WMO).

In contrast to the practice for homogenizing series using mean monthly or mean annual correction, weather-dependent daily or sub-daily corrections using transfer functions will almost always reduce the values of the new measurements to the old. The reason is that the predictors of interest (like global radiation or windspeed at screen level) are mostly not available in the past. On the other hand, using these variables in transfer equations necessitates that continuity of their measurements can be guaranteed towards the future.

The duration that an intercomparison should take depends on the application. For instance, when using a transfer function for 10-min temperatures like the one in Equation (1), 1 year of simultaneous measurements may be long enough to obtain statistically significant coefficients. For the daily T_n and T_x in Equations (3) and (4) a period of 2 years may be hardly sufficient. On the other hand, without considering the weather dependence of temperature differences, much longer periods than 2 years may be needed to obtain, e.g. representative monthly mean corrections. In the latter case, it should be kept in mind that the longer the duration of an intercomparison, the more difficult it will be to keep conditions constant (responsible people, environment, measuring methods, etc.).

If adequate modeling of rare events (like periods with snow cover in the Netherlands) is important, it may be advisable to perform part of the screen intercomparison in countries with a comparable climate but with a much more frequent occurrence of the event considered.

The present work focused on the effects of changes in thermometer screens on temperature measurements. The same approach applies to other variables and to other changes in the measurement infrastructure or changes in the local environment surrounding the measurement site. However, the national agencies responsible for carrying out climate observations are not yet equipped for deriving and applying weather-dependent corrections using transfer functions in these cases. Still, when we want to use the time series of the corresponding meteorological variables for studying long-term trends in, e.g. daily or sub-daily extremes, we have to deal with these problems. Because of the labor-intensive character of the related work, it may only be possible for a selected set of high-quality climate monitoring stations (e.g. a subset taken from the GCOS Surface Network; Peterson *et al*., 1997). We hope that the present work (Parts I and II) will stimulate both the

DOI: 10.1002/joc

performance of parallel measurements and the accompanying development and application of transfer functions for these stations.

Acknowledgements

We are grateful to David Murphy for proofreading the manuscript. We also thank our colleague, Adri Buishand for his valuable comments.

References

Alexander LV, Zhang X, Peterson TC, Caesar J, Gleason B, Klein Tank AMG, Haylock M, Collins D, Trewin B, Rahimzadeh F, Tagipour A, Ambenje P, Rupa Kumar K, Revadekar J, Griffiths G. 2006. Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research* **111**: D05109 (1–22).

Andersson T, Mattisson I. 1991. *A Field Test of Thermometer Screens*. Swedish Meteorological and Hydrological Institute: Report No.: RMK 62. 41.

Barnett A, Hatton DB, Jones DW. 1998. *Recent Changes in Thermometer Screen Design and their Impact*. Geneva: WWW/OSY Instruments and observing methods: Report No.: 66, WMO/TD871, 12.

Brandsma T. 2004. *Parallel Air Temperature Measurements at the KNMI-terrain in De Bilt (the Netherlands) May 2003 – April 2005 (Interim Report)*. KNMI-publication 207, De Bilt, The Netherlands, 29.

Brandsma T, Können GP. 2006. Application of nearest-neighbor resampling for homogenizing temperature records on a daily to sub-daily level. *International Journal of Climatology* **26**: 75–89.

Cleveland WS, Devlin SJ. 1988. Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**: 596–610.

Della-Marta PM, Wanner H. 2006. A method of homogenizing the extremes and mean of daily temperature measurements. *Journal of Climate* **19**: 4179–4197.

Gill GC. 1983. *Comparison Testing of Selected Naturally Ventilated Solar Radiation Shields*. Report to NOAA Data Buoy Office for Development Contract #NA-82-0A-A-266, 15, 15 figs.

Hatton DB. 2002. Results of an intercomparison of wooden and plastic thermometer screens. In papers presented at *The WMO Technical Conference on Meteorological Instruments and Methods of Observation (TECO-2002)*, Bratislava, Slovak Republic, 23–25 September 2002, P1.1 (19). Instruments and Observing Methods report No.75 (WMO/TD – No 1123, WMO, Geneva, 2002).

Klein Tank AMG, Können GP. 2003. Trends in indices of daily temperature and precipitation extremes in Europe, 1946–1999. *Journal of Climate* **16**: 3665–3680.

Larre MH, Hegg K. 2002. Norwegian national thermometer screen intercomparison. In papers presented at the *WMO Technical Conference on Meteorological Instruments and Methods of Observation (TECO-2002)*, Bratislava, Slovak Republic, 23–25 September 2002, P1.1 (eq1). Instruments and Observing Methods Report No.75 (WMO/TD – No 1123, WMO, Geneva, 2002).

Lefebvre G. 1998. *Comparison of Meteorological Screens for Temperature Measurement*, paper presented at TECO-98 (Casablanca), Instruments and Observing Methods Report No. 70, WMO/TD – No. 877, pp 315, Geneva.

Mawley E. 1897. Shade temperature. *Quarterly Journal of the Royal Meteorological Society* **23**(102): 69–87.

Parker DE. 1994. Effects of changing exposure of thermometers at land stations. *International Journal of Climatology* **14**: 1–31.

Parker DE. 2004. Large-scale warming is not urban. *Nature* **432**: 290.

Peterson TC, Daan H, Jones PD. 1997. Initial selection of a GCOS surface network. *Bulletin of the American Meteorological Society* **78**: 2145–2152.

Quayle RG, Easterling DR, Karl TR, Hughes PY. 1991. Effects of recent thermometer changes in the cooperative station network. *Bulletin of the American Meteorological Society* **72**: 1718–1723.

Sparks WR. 1972. *The Effect of Thermometer Screen Design on the Observed Temperature*. WMO – No. 315: Geneva, Switzerland, 106.

Van der Meulen J, Brandsma T. 2007. Thermometer screen intercomparison in De Bilt (the Netherlands), Part I: micro and short-term effects. *International Journal of Climatology*, this issue. DOI: 10.1002/joc.1531.

WMO 2003. *Manual on the Global Observing System (Volume I, Part III, Part. 2.9).* WMO – No. 544: Geneva, Switzerland.