Royal Netherlands
Meteorological Institute
Ministry of Infrastructure
and Water Management

Homogenization of daily temperature data of the five principal stations in the Netherlands - version 2.0

C. de Valk, T. Brandsma

De Bilt, 2026| Scientific report; WR-26-01

Homogenization of daily temperature data of the five principal stations in the Netherlands – version 2.0

Cees de Valk, Theo Brandsma

De Bilt, 2026 | Scientific Report; WR 26-01

Preface

Homogeneous time series of daily minimum (TN), maximum (TX), and mean (TG) temperatures play a key role in studying climate change and variability. This report presents version 2.0 of the homogenization of 20th-century daily temperature records from the five principal meteorological stations in the Netherlands. The homogenization adjusts the data for the effects of documented instrument changes and relocations.

Compared to version 1.0 (Brandsma, 2016a), the most important methodological updates are:

- For the four stations with parallel measurements – De Kooy, Eelde, Vlissingen, and Beek/Maastricht Airport (collectively referred to as the H4 stations), measurements of additional weather variables – such as wind, humidity, and cloud cover – are used to better account for meteorological influences on temperature inhomogeneities in TN and TX. TG is no longer homogenized stand-alone but its adjustments are derived from the adjustments in TN and TX.
- For De Bilt, where we do not have suitable parallel measurements, the homogenization is made more robust and precise by using more data from the H4 stations than in version 1.0.
- The original measurement data and both homogenized datasets (old version 1.0 and new version 2.0) are now readily accessible via the KNMI data portal, allowing users to assess the impact of homogenization on their applications.

For De Kooy, Eelde, Vlissingen, and Beek/Maastricht Airport, the updated version provides more realistic day-to-day adjustments. However, the long-term trends differ little from those of version 1.0.

For De Bilt, differences in long-term trends between the versions and between version 2.0 and the unadjusted data are substantial for climate indices indicating heat such as the annual maximum temperature and the number of heatwaves. The number of heatwaves in 1901-1950 is now estimated to be 14, twice as many as in version 1.0. However, even for these indices, the long-term trends from both versions cannot be convincingly distinguished because they are highly uncertain due to natural year-to-year fluctuations. For other climate indices, the differences between the versions are minor.

Together, the homogenized data from the five stations now provide a spatially and temporally more coherent and consistent foundation for studying the climate of the Netherlands and its change over time.

Acknowledgments

We would like to acknowledge the valuable contributions by the six reviewers of the draft of this report: Enric Aguilar, Jules Beersma, Frans Dijkstra, Jos de Laat, Peter Siegmund and Gerard van der Schrier.

In addition, we thank Frans Dijkstra for conducting an independent analysis of the results, part of which we replicated to produce Figs. C.1–C.3. We are also indebted to Jouke de Baar for showing the potential of cross-validation and forward selection of covariates and their interactions.

Summary

This report presents version 2.0 of the homogenization of records of daily minimum temperature TN, daily maximum temperature TX and daily mean temperature TG at the five principal stations in the Netherlands. These historical records – starting at the beginning of the 20^{th} century – have been influenced over time by station relocations and changes in measurement equipment. For four of the stations – Den Helder/De Kooy, Groningen/Eelde, Maastricht/Beek, and Vlissingen/Souburg (the H4 stations)—parallel (simultaneous) measurements at the old and the new locations were made over periods of 4–10 years to support accurate adjustment. At De Bilt, only the effect of a screen change in 1950 was measured in parallel; the impact of the near-simultaneous relocation was not directly recorded.

Our approach to homogenization rests on the following premises: (a) homogenization is applied cautiously and only when supported by evidence, focusing on adjusting for known and substantial changes in instrumentation or location. (b) gradual environmental changes, like urbanization, are not adjusted for due to their complexity. (c) homogenized data are versioned and stored alongside raw data for transparency and future updates. (d) robust methods are preferred, and automatic breakpoint detection is used only for quality control – not for undocumented changes – to avoid introducing new errors. We further note that daily maximum temperatures may be affected by turbulent eddies during warm sunny days. Daily minimum temperatures may be extremely sensitive to small breezes during stable nights. Even the existence of high-quality parallel measurements cannot fully account for these effects.

This updated version of the homogenization was developed for three main reasons:

1. Recent research (de Valk and Brandsma, 2023) has shown that for the H4 stations where parallel measurements are available, daily temperature adjustments to account for instrument relocation can be improved by including other weather variables – like wind, humidity and cloudiness – that influence temperature differences between sites.

2. For De Bilt, where parallel measurements cannot be used for homogenization, a more robust adjustment can be obtained by a careful choice of the reference data to be used for this purpose (Dijkstra et al., 2022); in particular, the use of homogenized data instead of raw measurements and careful selection of station(s) and calibration time-intervals.

3. There is growing demand for making both the original and adjusted datasets more transparently available, allowing users to assess how homogenization affects their analyses.

For the H4 stations, two methods are compared: an updated version of the original Quantile Delta Mapping (QDM) technique and a newer method using Generalized Additive Models (GAMs). The latter enables the use of measurements of additional weather variables to improve the accuracy of the temperature estimates for individual days.

For De Bilt, where no suitable parallel dataset exists, an improved QDM method is chosen, because we cannot show that GAMs can offer benefits. The QDM method uses longer calibration periods than in version 1.0 (15 years before and after breakpoints) and the average of the homogenized time series of two inland stations (Eelde and Maastricht) as reference. This leads to more precise and robust adjustments for the change in screen combined with the relocation in 1950-1951.

Key findings and conclusions:

1. The homogenization helps to align trends across stations.

2. For the H4 stations De Kooy, Eelde, Vlissingen, and Beek/Maastricht Airport, the updated version leads to more realistic day-to-day adjustments (reductions in mean square error between 6% and 76%). However, the long-term trends differ little from those of version 1.0. This suggests that the earlier method was already fairly reliable for these stations.

3. For De Bilt, the adjusted version 2.0 method improves precision and robustness when compared to the previous version, with the caveat that validation is limited by the absence of suitable parallel measurements.

4. For De Bilt, differences in long-term trends between the versions and between version 2.0 and the unadjusted data are substantial only for climate indices indicating heat, such as the number of heatwaves. The number of heatwaves in 1901-1950 is now estimated to be 14, twice as many as in the previous version. However, even for these indices, the long-term trends from both versions cannot be conclusively distinguished, because of their high uncertainty due to natural year-to-year fluctuations. For other climate indices, the differences are minor.

5. Homogenization can introduce errors if the models used are not calibrated with sufficient precision, for example when calibrated on a dataset that is too small. We find that the version 2.0 homogenization has been calibrated with sufficient precision; the calibration has little impact on the overall precision of long-term climate trends, which is determined mainly by the natural year-to-year variability.

Checking for inhomogeneity of measurement records and – where needed – homogenization helps to ensure the temporal and spatial consistency of the national climate records, which is essential for the analysis of the climate of the Netherlands and its change over time.

Contents

Part I

Report

Glossary

The definitions in this glossary are valid within the scope of this report.

Accuracy Accuracy expresses how close a measurement is to the true value. One way to quantify accuracy is by the root mean square (RMS) error (higher value means less accurate).

Bias Bias is systematic error in a measurement or estimate; it does not change if averages over multiple measurements/estimates are taken.

Candidate station Station to be homogenized

H4 stations The four stations with parallel measurements: De Kooy (formerly Den Helder), Eelde (formerly Groningen), Vlissingen (temporary at Souburg), and Beek/Maastricht Airport (Beek) (formerly Maastricht).

Homogenization Climate data homogenization is the adjustment of climate records, if necessary, to remove the effects of non-climatic factors, so the temporal variations in the adjusted data reflect only the variations due to climate processes.

Parallel measurements Simultaneous measurements from both the old and the new location (in case of a relocation) or (in case the instrument type is changed) from the old and new instrument at the same site.

Precision Precision expresses how close statistics derived from repeated measurements are to each other (either in a real experiment or in a "thought experiment"). The variation in repeated measurements is known as the sampling uncertainty or non-systematic error (see below).

Principal stations The five principal climatological stations monitored by KNMI: De Kooy (235), De Bilt (260), Eelde (280), Vlissingen (310) and Maastricht Airport (380) (Beek).

Reference series An independent, homogeneous time series used to assess homogeneity of a candidate station, usually derived from neighboring stations.

Robustness Robust means resistant to errors caused by fluctuations or by minor violations of its underlying assumptions.

Root mean square (RMS) error The RMS error is the square root of the mean of the squared deviations from the true value. It is a measure of accuracy; it encompasses both sampling uncertainty and bias.

Sampling uncertainty The sampling uncertainty is the error caused by observing a sample instead of the whole ensemble.

Standard deviation The standard deviation is the square root of the variance (see below).

Variance Variance is the mean of the squared deviations from the mean value. It expresses precision (lower variance means higher precision).

Chapter 1

Introduction

This report presents the latest update of the ongoing effort to improve the quality and usability of historical temperature records in the Netherlands. The focus is on the homogenization of daily temperature data from five principal weather stations, building on earlier work and incorporating new insights and methods. This chapter first provides background information on the reasons for and challenges of homogenizing meteorological time series, followed by an overview of the report structure and objectives.

1.1 Background

Meteorological time series may be inhomogeneous because of artificial changes unrelated to actual climate variations. For example, relocation of weather stations and/or instruments, changes in instruments and measurement practices, and changes in the environment of the instrument such as the construction of buildings or the growth or removal of trees (Pielke Sr et al., 2007; Venema et al., 2012; Brandsma, 2011, 2025). For climate change and variability studies, it is important to deal with potential sources of inhomogeneities and to adjust for them if the effects are serious (WMO, 2020). This adjustment is called homogenization.

The daily temperature series of the five principal stations in the Netherlands have known inhomogeneities due to relocations. In addition, at De Bilt, the thermometer was moved from a large pagoda screen—open at the bottom—to a Stevenson screen. In most cases, parallel observations have been made that facilitate the adjustment of the series for the inhomogeneities. In Brandsma (2016a,b), KNMI presented version 1.0 of homogenized daily minimum, maximum and mean temperatures (TN, TX, TG) for these five stations.

Meanwhile, further research has been published on the homogenization of these temperature data (Dijkstra et al., 2022; de Valk and Brandsma, 2023), indicating that further improvement of certain aspects of the homogenization may be possible. In particular, incorporating additional weather variables (where available) and improving the selection of reference series for De Bilt may enhance the robustness of the homogenization.

With the new version 2.0 of the homogenized temperature data, KNMI aims to

- provide an updated high-quality dataset for climate research and applications that incorporates these improvements;
- in addition, provide users of the homogenized data with information on the accuracy of the homogenization, and
- make successive versions of the homogenized data, as well as the unadjusted data, available to users, both in the interest of transparency and to enable comparison with earlier analyses based on different versions of the data.

1.2 This report

This report presents version 2.0 of the homogenization of the historical records of TN, TX and TG at the five principal stations in the Netherlands.

Chapter 2 describes the data from the five principal stations and documents known changes in location, instrumentation, and surroundings.

In Chapter 3, several methods are described, compared, and evaluated, and the uncertainties in the outcomes are assessed. This assessment is quantitative where possible. In particular, we compare nonlinear trends and aggregated climate indices—mostly related to temperature extremes—derived from the homogenized series. These comparisons inform the method selection for version 2.0.

In Chapter 4, the resulting new time series of TN, TX and TG are compared to the previous version 1.0 and to the unadjusted data.

The findings are summarized in Chapter 5, together with some guidance on the use of homogenized daily temperature data.

Chapter 2

Data

In this chapter, we outline the historical temperature records of the five stations considered in this study. The impact of station relocations and equipment changes on the measurements is discussed, as well as the presence or absence of parallel observations to aid homogenization. In addition, we briefly address the nationwide lowering of the thermometer screens around 1960. We focus on the metadata related to changes known to have resulted in major inhomogeneities dealt with in this report. A more extended description of the metadata (detailed station history and measurement conditions) is presented in appendix L.

2.1 Description of the temperature data

The data to be homogenized consist of the daily operational TN, TX and TG data of the five principal stations of KNMI (see Figure 2.1): De Kooy (235), De Bilt (260), Eelde (280), Vlissingen (310) and Maastricht Airport (380) (referred to as Beek in this report).

The operational series from De Bilt begins in 1901, while the other stations' series start in 1906. Brandsma et al. (2013) standardized the data and methods for calculating TN, TX and TG for those stations.

TN and TX represent, respectively, the minimum and maximum temperatures over the 0:00–0:00 UTC interval. TG is the arithmetic mean of 24 hourly temperatures (T1 to T24), where each index corresponds to the hour in UTC.

The following major relocations took place:

1. De Kooy continued the monitoring of the station Den Helder since 1 August 1972. Den Helder was located along the North Sea dike on the western edge of the city of Den Helder, whereas De Kooy is located at an exposed site on the airport on the SE edge of Den Helder, about 1 km from the Waddenzee. Figure 2.2 shows both locations.

2. Eelde was first situated in the city of Groningen until 1950 and was then relocated to an exposed site at the airport, at 10 km to the south of Groningen. Figure 2.3 shows both locations.

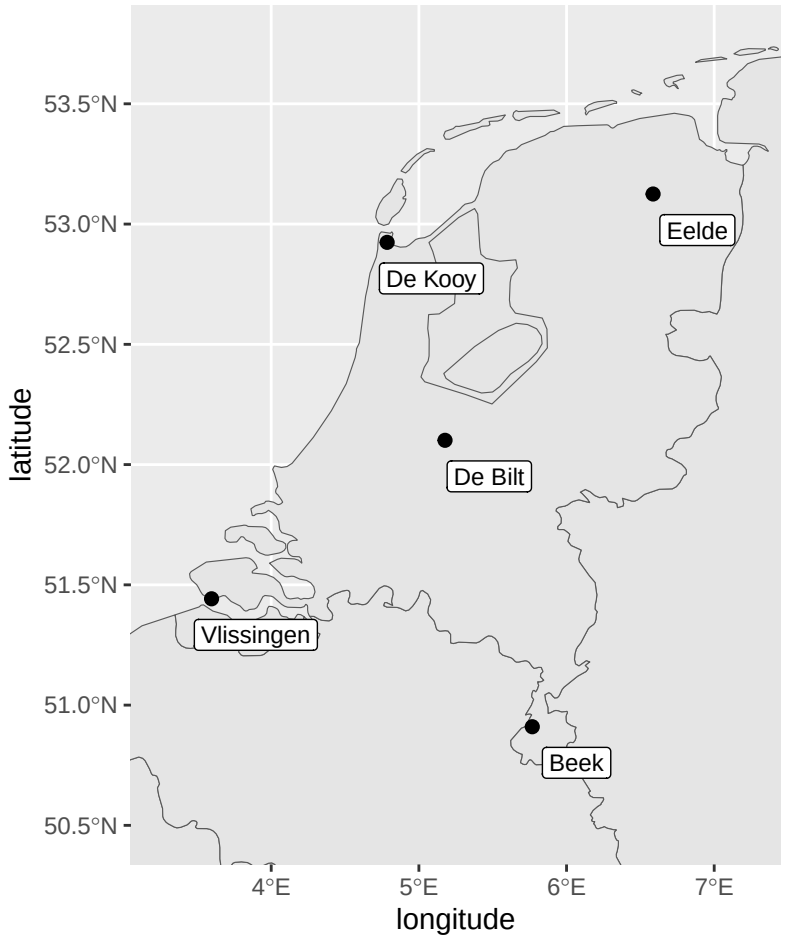

Figure 2.1: Locations of the five principal stations De Kooy, Eelde, De Bilt, Vlissingen and Beek (Maastricht Airport).

3. Vlissingen is at an exposed site in the harbour on the Westerschelde estuary. The station was temporally relocated to Souburg between 1947 and 1958. Station Souburg was located on an inland airport at 1.8 km north-northwest (NNW) of Vlissingen. Figure 2.4 shows both locations.

4. Beek was situated in the city of Maastricht until 1950 and was then relocated to an exposed location on Beek airport in 1951. This airport is about 9 km NE of Maastricht, now called Maastricht Airport. The elevation of Beek is about 65 m higher than that of Maastricht. In addition, temperature measurements in Maastricht were made at 20 m above ground level, which deviated strongly from

the standard measurement height of 2.2 m at that time. Figure 2.5 shows both locations.

5. De Bilt experienced a change from a large pagoda screen to a Stevenson screen on 16 September 1950 (Figure 2.6), a modification known to affect temperature readings. The effect of the screen change was studied in Brandsma (2019). The screen change was followed by a southward relocation over about 300 m from a sheltered site to an exposed site on 27 August 1951. Figure 2.7 shows both locations.

Figure 2.2: Map showing the current location at Airport De Kooy and the old location along the coast in the city of Den Helder.

The details of the station relocations are summarized in Table 2.1.

2.2 The collection of parallel temperature measurements

For the H4 stations, relocation involved conducting parallel measurements at the old site and the new site over periods ranging from 4 to 10 years in order to facilitate homogenization (see Table 2.1). These observations were also part of the standardization in Brandsma et al. (2013).

Figure 2.3: Map of the area around Groningen showing the current location at Groningen Airport Eelde and the old location in the city of Groningen.

Unfortunately, no parallel observations were made during De Bilt's 1951 relocation. The move was unplanned, and subsequent construction activities disturbed the former site. Therefore, data from other stations have been used for the homogenization of this series.

For the H4 stations, Figure 2.8 shows the monthly mean temperature differences (TN, TX, TG) observed between the old and new sites during the parallel measurement periods. Differences of up to 2°C in magnitude are observed; the largest differences concern TN. This is likely due to site-specific nighttime cooling effects. For instance, the difference in the transport of heat between a more maritime location (Den Helder) and a more continental location (De Kooy) and the difference in nighttime outgoing long-wave radiation for a station in a city (Groningen) and on an airport (Eelde)

Since no parallel measurements exist for De Bilt for the combined screen change and relocation, we consider the differences between the monthly means over 15-year intervals after and before the known breakpoints. Because these differences are affected by climatological change, we subtract a similar difference from a reference time series constructed by averaging the homogenized version 2.0 time series of all H4 stations. Compared to the H4 stations, the differences for De Bilt are relatively small and mainly

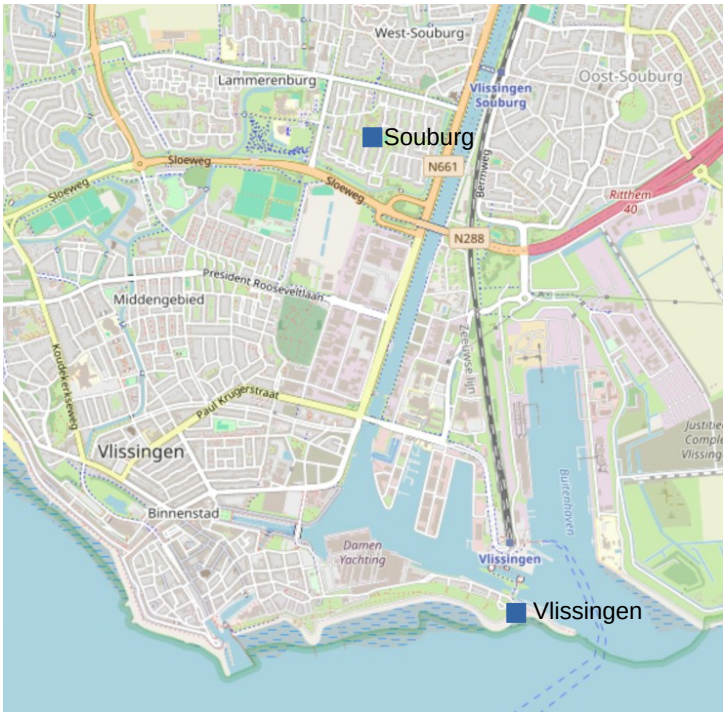

Figure 2.4: Map of Vlissingen showing the current Vlissingen location along the water and the old Souburg location. Note that in the current map the airport at Souburg is replaced by a residential area.

concern TX. The largest monthly mean TX differences are between 0.7 and 1.0°C in the period May–August. This is much larger than can be explained from the screen change alone. Brandsma (2019) shows that the main effect of the screen change is a decrease of about 0.3°C in monthly mean summer TX.

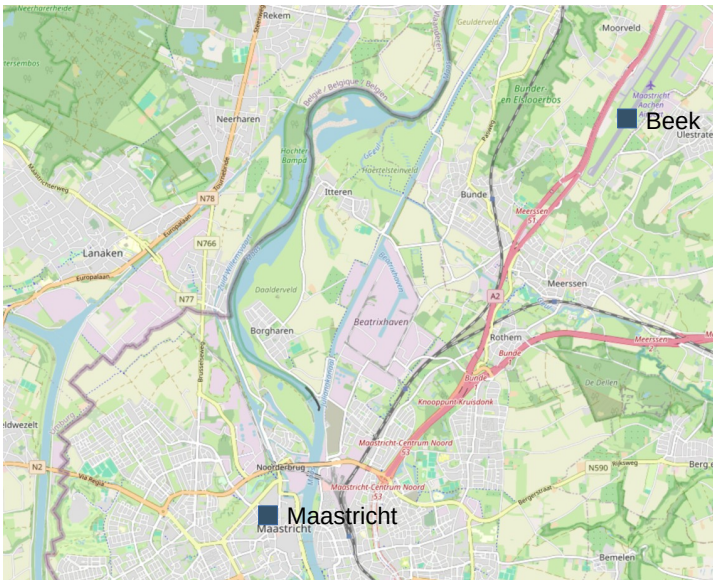

Figure 2.5: Map of the area around Maastricht showing the current location at Beek Airport and the old location in the city of Maastricht.

current name	station	lat.	lon.	elev.	period(s)	overlap
De Kooy	Den Helder*	52.967	4.750	4.4	1906-1970	1961-1970
(235)	De Kooy	52.924	4.785	0.5	1961-present	(10 yr)
Eelde	Groningen	53.217	6.550	2.1	1907-1951	1946-1951
(280)	Eelde	53.125	6.586	3.5	1946-present	(6 yr)
Vlissingen	Souburg	51.467	3.583	-0.5	1947/08/16-1962	1959–1962
(310)	Vlissingen**	51.442	3.596	8.0	1906-1947/08/14, 1958/05-present	(4 yr)
Maastricht	Maastricht	50.850	5.693	49.4	1906-1952	1946-1952
Airport (380)	Beek	50.910	5.768	114.0	1946-present	(7 yr)
De Bilt	De Bilt (a)			2.0	1901-1951	–
(260)	De Bilt (b)	52.101	5.177	2.0	1952-present	(0 yr)

Table 2.1: Station, latitude [deg], longitude [deg], elevation [m], period(s) covered by the measurements and overlap period (duration of the parallel measurements). Notes: * Data gap from September 1944 – May 1945: not filled in. ** Data gap from October 1944 – July 1945: not filled in.

Figure 2.6: Screen change in De Bilt on 16 September 1950. On the left the old pagoda screen and on the right the new Stevenson screen. Note that this Stevenson screen is still at 2.20 m above ground level and not at the current 1.50 m (see Section 2.4).

Figure 2.7: Map from the KNMI terrain in 1960 showing the old location (De Bilt-old) and the new location (De Bilt-new) from 27 August 1951 onward. Note that the building just north and west of De Bilt-old was built after 27 August 1951 and was the reason for the relocation.

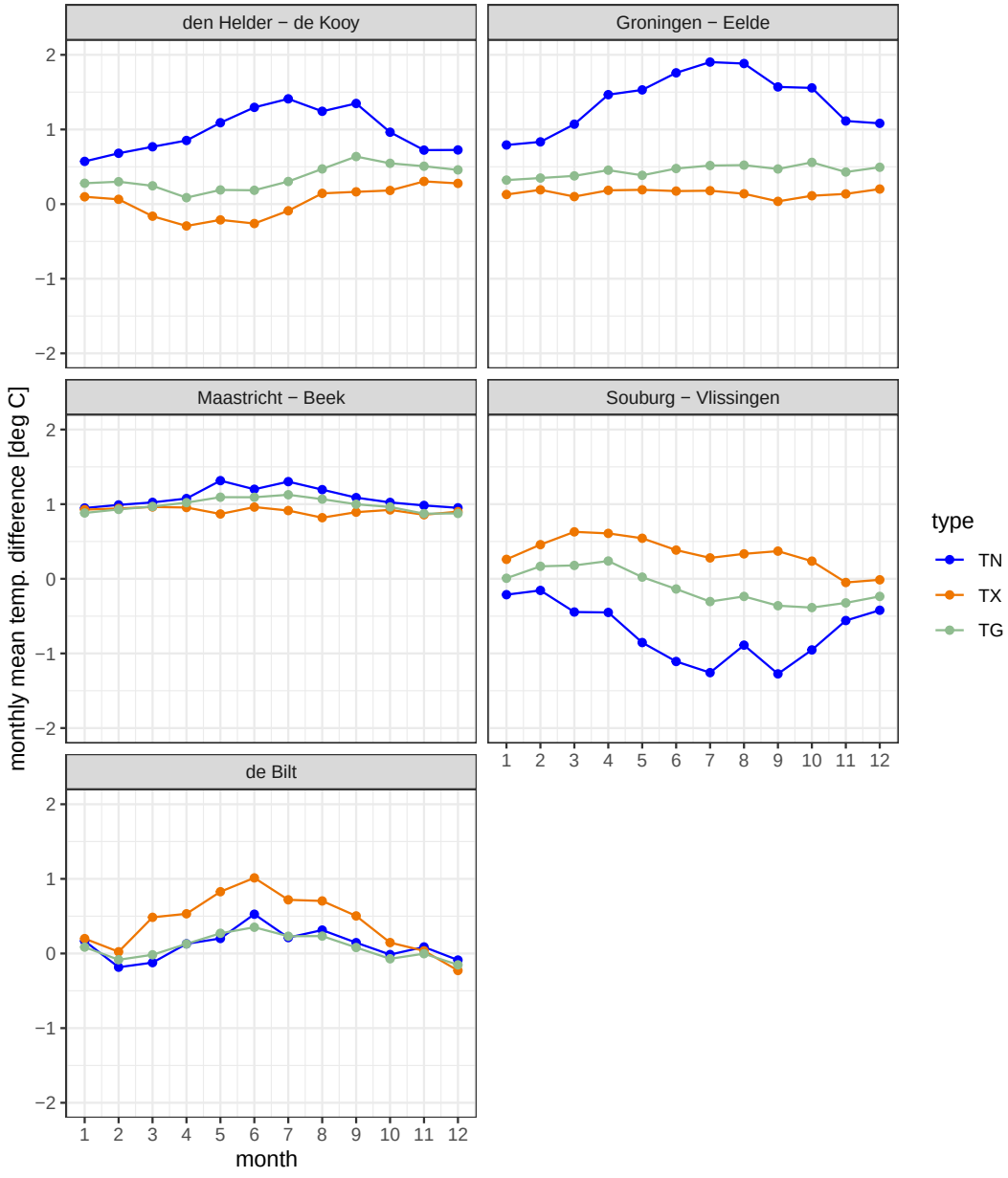

Figure 2.8: Monthly mean differences of TN, TX and TG at the old site and the new site from the parallel measurements of the H4 stations (top and middle). For De Bilt (bottom): monthly mean differences of measured TN, TX and TG of De Bilt over 15-year intervals before and after the breakpoints in 1950/51 minus the corresponding differences of the mean of homogenized (v2.0) TN, TX and TG over the four H4 stations.

2.3 Meteorological measurements other than of temperature

Version 1.0 of the homogenization was based on temperature data alone. For the H4 stations, version 2.0 introduces the use of other meteorological variables as covariates, because it is known that differences between the daily temperatures at two different sites may depend on the weather conditions and, in the case of coastal stations, also on sea surface temperatures. This will be discussed in Section 3.3.

The following data are available as potential covariates: hourly measurements of wind speed (F), wind direction (D), cloudiness (N) and relative humidity (RH). While F and D were measured hourly, N and RH are only available at 7:40, 13:40 and 18:40 UTC. As RH is strongly correlated with temperature T, we calculate specific humidity (HUM) from T and RH (WMO, 2014) as a measure of the dryness of the air. We then use HUM as a covariate instead of RH. For the homogenization of TN, we use the morning values of N and HUM, and for TX, we use the afternoon values of N and HUM. F and D are combined into a wind vector U with an easterly and northerly component. For TN, we use the value of U in the hour that TN occurs (tTN) and for TX, the value in the hour that TX occurs (tTX).

The stations Den Helder/De Kooy and Vlissingen/Souburg are located near the sea. An effect of sea surface temperature (SST) on the temperature differences is anticipated. We therefore include SST as a potential covariate for these stations. We have a long series of monthly mean values of SST near Den Helder (Van der Hoeven, 1982) with daily values over a sub-period. For the relatively smooth SST signals, daily values are estimated from the observed monthly averages using a GAM model (de Valk and Brandsma, 2023) variant which can use linear functionals as observations, with the degree of smoothing tuned on the available daily values from Den Helder.

Missing values and inhomogeneities in the covariates are dealt with as follows:

- A tiny fraction (about 0.01%) of the needed hourly values of T, F, D and RH are missing. These are filled in with the data of principal station De Bilt (Figure 1).
- From 1951/03/01 onward, D and F were no longer measured at Maastricht. Therefore, D and F of Beek are used instead for the period (1951/03/01–1952/12/31). As Maastricht and Beek have different wind speeds, we used parallel measurements at these stations over 1948/01/01–1950/12/31 to adjust Beek to Maastricht by quantile matching (see Section 3.2).
- The time series of F from Maastricht is inhomogeneous in the period 1916/07/01–1926/08/31. For this period, we used the F data of De Bilt multiplied by a factor derived from parallel measurements over 1931/01/01–1950/12/31.
- Over the period 1906–1927, there is a trend in F in Den Helder making the F series inhomogeneous. To homogenize the series, we calculated monthly mean values of F for Den Helder relative to De Bilt for each year in the period 1906–1950. We used a loess smoother (span = 0.4) to find monthly/annual increments correcting the F values before 1928 to the values thereafter.

- From 1961/01/01 onwards, N was no longer measured at Souburg. Therefore, cloud cover of Vlissingen for the period 1961/01/01–1962/12/31 is used instead.

The methods used to correct inhomogeneity of non-temperature variables listed above are somewhat crude. However, this is justified, as the effect of these non-temperature variables on temperature differences between two sites is relatively small compared to the effect of the temperature itself.

2.4 Lowering of thermometer screens around 1960

Following a new WMO regulation, KNMI lowered all thermometer screens at its stations from 2.20 m to 1.50 m above ground level around 1960. Brandsma (2022) analyzed parallel temperature measurements at both heights in De Bilt (2017–2019) and in De Bilt and Witteveen (1950s). The effect of the lowering of the screen on annual minimum temperatures (TN) is a decrease of about 0.2°C, while the effect on annual mean maximum temperature (TX) is an increase of about 0.1°C. In general, the differences in 2.20 and 1.50 m air temperature show a seasonal cycle with the largest values in summer and the smallest in winter. The impact on daily mean temperature (TG) is negligible. While the effect is small and varies with wind and season, it confirms that the screen height change had a measurable influence – though not large enough to warrant adjustments to the historical records.

Such nationwide changes require parallel measurements to detect and quantify potential inhomogeneities. Automatic homogenization methods are not sufficient for identifying or adjusting such changes.

Chapter 3

Methods

This chapter describes the methods used to homogenize the historical daily temperature series of the five principal meteorological stations in the Netherlands. Building on earlier work, new approaches have been introduced or refined to improve the accuracy and robustness of the homogenized data. The chapter begins by outlining general limitations to homogenization and the premises used in the current study. This is followed by a description of the previous method (version 1.0) and the improvements in version 2.0. Separate approaches are presented for stations with and without parallel measurements, reflecting the different data constraints at each site.

3.1 Limitations and premises

3.1.1 Limitations

In this section we describe several limitations to homogenization.

Homogenization has a number of limitations, as documented in WMO documents (Aguilar et al., 2003; World Meteorological Organization, 2004, 2017) and journal articles (Venema et al., 2012; Williams et al., 2012). In particular:

1. You can't remove what you can't detect. Very small or gradual inhomogeneities, or changes shared by many stations, may be indistinguishable from genuine climate change. They will often survive in "homogenized" data.

2. Some inhomogeneities are effectively undetectable with current techniques. This may occur when breaks are small relative to year-to-year variability, when multiple breaks are close together, or when metadata are missing and/or incorrect.

3. Adjustment is uncertain even when a break is found. Estimated step sizes and their timing – when not documented in the metadata – may have uncertainties that directly translate into uncertainty in trends and variability.

4. Reference series are not perfect. Relative homogenization assumes neighbors share the same climate but not the same non-climatic breaks. When this assumption

fails (network-wide changes, regional shifts in observing practice), residual biases or mis-assignments are unavoidable.

5. Some variables are inherently harder to homogenize than others. In general, monthly temperature is "easiest"; precipitation, wind, radiation, humidity and daily extremes (as in this report) are much harder to homogenize robustly, and benchmark studies show lower performance.

6. Results depend on network design and metadata quality. Sparse networks and poor metadata can push homogenization beyond its reliable limit. WMO stresses that good network design and meticulous metadata are as important as the statistical method.

7. Homogenization can change the inferred climate signal. This is both its point and a risk. Homogenization can alter regional trends from strongly warming to weakly warming (or vice versa), especially for poorly measured variables. This is desirable if the adjustments are right, but it also means that the homogenization uncertainty must be considered when interpreting trend estimates.

8. Homogenization cannot fix fundamental flaws or non-climatic effects. If a station is dominated by local artifacts (e.g. intense urban canopy, major exposure problems) with no good references, WMO guidance suggests such series may be unsuitable for climate trend analysis altogether.

9. Homogenization of extremes has fundamental limitations. For instance daily maximum temperatures may be affected by turbulent eddies during warm sunny days. And daily minimum temperatures may be extremely sensitive to small breezes during stable nights. Even the existence of high-quality parallel measurements cannot fully account for these effects.

3.1.2 Premises of this study

Homogenization is generally applied with caution in order to minimize the risk of introducing artificial changes. In practice, this means that it is performed only when there is clear evidence of significant inhomogeneities that warrant adjustment.

The homogenized series should be presented and stored with version numbers and alongside the measured series. This enables (a) the construction of new versions of the homogenized series when new techniques and/or knowledge become available, and (b) facilitates comparison between measured and homogenized series.

We also prefer to apply robust methods that are relatively insensitive to the specific choice of calibration data of the homogenization algorithm.

We do not attempt to adjust for gradual environmental changes (e.g., urbanization and slow tree growth). Currently this is difficult, mainly because the right metadata is often lacking (Pielke Sr et al., 2007). Therefore, the effects of such changes may still be present in the homogenized time series. Nonetheless, it is important to estimate the potential impact of these environmental changes on long-term temperature trends. As an example, Brandsma et al. (2003) estimated for De Bilt a long-term effect of urbanization

of about 0.1°C per century. For the European average temperatures Chrysanthou et al. (2014) estimated this effect as 0.07°C per century for summer and 0.026°C for the annual average. All these trends are small compared to the total observed trend (about 2°C in the Netherlands), which in this case justifies omitting an adjustment for urbanization.

Furthermore, we only adjust the effects of known changes in instrument or location, and only when these are expected to have a material effect on long-term climate trends. Therefore, small step changes or anomalies limited to short periods are not adjusted.

In the Netherlands, a record of major instrument changes and displacements has been kept, which reduces the risk of missing undocumented inhomogeneities. Therefore, we do not try to find breakpoints in the time series where undocumented step changes may have occurred. This is far from simple, and using these breakpoints for homogenization may not only remove errors but also introduce new ones. However, for the purpose of quality control, automatic breakpoint detection (Venema et al., 2012) on the homogenized records may still be useful for for quality control, in particular to identify possible undocumented changes for further examination.

3.2 The previous version 1.0

Version 1.0 (Brandsma, 2016a) is based on a statistical method that aligns the probability distributions of two datasets. This method is named quantile matching (QM), also known as quantile delta mapping (QDM) in the literature. See Cannon et al. (2015) e.g. for its use to adjust climate model output for bias.

We explain the QDM method for the case that no parallel measurements are available, which is the most complicated one.

For a given probability p, the quantile $Q(p)$ is the value exceeded with probability $1-p$, where we assume that the function Q is both continuous and increasing. Therefore, if F is the cumulative distribution function, then $F(Q(p)) = p$ and $Q(F(x)) = x$. The QDM method adjusts quantiles; from these adjustments, adjustments of the observations are derived.

Besides the time series of observations at the candidate station, it also considers a reference time series, for example of the temperature at another site, or of the average of the temperatures at several other sites.

This reference time series is assumed to be homogeneous, i.e. free of errors due to changes in instrument, location, etc.

Furthermore, we assume that if the observations of the candidate station would also be homogeneous, then differences between the quantiles of the observations and of the reference data would be the same before and after the breakpoint. Therefore, for each probability p,

$$\hat{Q}_1(p) - Q_1^r(p) = Q_2(p) - Q_2^r(p), \tag{3.1}$$

with (see Figure 3.1)

$\hat{Q}_1(p)$ the estimated quantile of the homogenized observations before the breakpoint,

$Q_1^r(p)$ the quantile of the reference data before the breakpoint,

$Q_2(p)$ the quantile of the observations after the breakpoint,

$Q_2^r(p)$ the quantile of the reference data after the breakpoint.

Subtracting the quantile $Q_1(p)$ of the observations before the breakpoint in (3.1), we obtain the adjustment of this quantile:

$$\hat{Q}_1(p) - Q_1(p) = \Psi(p) := (Q_1^r(p) - Q_1(p)) - (Q_2^r(p) - Q_2(p)), \tag{3.2}$$

Let $x(t)$ be the observation at a day t before the breakpoint, which needs to be adjusted. Then by taking $p = Q_1^{-1}(x(t)) = F_1(x(t))$ in (3.2), we obtain for the adjusted observation $\hat{x}(t)$:

$$\hat{x}(t) - x(t) = \hat{Q}_1(F_1(x(t)) - Q_1(F_1(x(t)) = \Psi(F_1(x(t)). \tag{3.3}$$

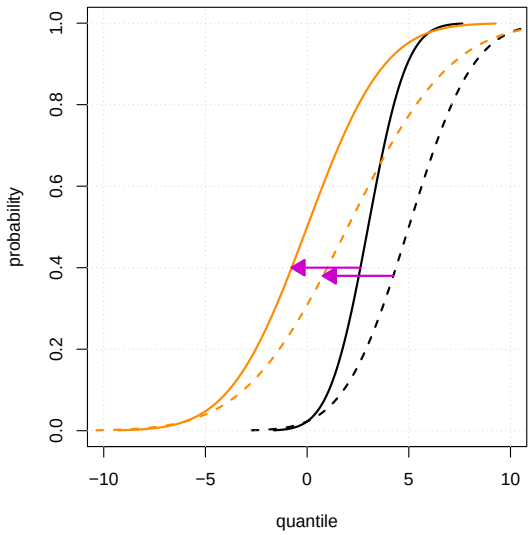

Figure 3.1: Illustration of QDM (eq. (3.1)): quantiles (horizontal) and probabilities of non-exceedance (vertical). Black: reference series; orange: homogenized series. Dashed: after breakpoint; full: before breakpoint.

This model is estimated separately for each month of the year to reflect the variation of prevailing weather over the year, which may affect the relationship between temperatures at different locations. In the computations, $\Psi(p)$ is estimated for probabilities $p = 0.05, 0.1, ..., 0.95$ from the data of 3-month intervals centered on the month of interest, and then successively smoothed over the probabilities and the months using a LOESS

filter of order 2 with a tricube weight function and a span of 0.6. For $p < 0.05$ and $p > 0.95$, the estimates for $p = 0.05$ resp. $p = 0.95$ are used.

For version 1.0, the algorithm described above was applied for the homogenization of daily TN, TX and TG at De Bilt, using the corresponding measured daily temperature at Eelde as reference time series. Eelde was chosen because of all the principal temperature stations, its situation is most similar to De Bilt. For estimating Ψ in (3.2), data were used in 4-year intervals before and after the period containing the known breakpoints at the Bilt (see Table 2.1). The reason for not using longer intervals was to limit the impact of any violations of assumption (3.1). The distribution function F_1 in (3.2) was estimated from all data before the breakpoints. Between the breakpoints at 1950-09-16 and 1951-08-27, the adjustment was linearly interpolated.

For the other four principal stations (see Table 2.1), parallel measurements over several years are available, so there is no need for a reference time series. For these, the same algorithm was used as above, but with $Q_1^r = Q_2^r = 0$ and using the parallel measurements to estimate Ψ.

3.3 Version 2.0 using parallel measurements

Version 2.0 for the four principal stations De Kooy, Eelde, Beek, and Vlissingen, where parallel measurements are available (see Table 2.1), is based on an entirely different method from version 1.0.

The motivation for developing this new method is to incorporate more physical information in the homogenization than in version 1.0. In particular, the variation of temperature quantiles with season in version 1.0 originates from the prevailing weather in different months, meaning that in version 1.0, the month was used as a crude proxy for varying weather conditions. It would be preferable to predict the dependence of temperature differences between sites or instruments directly from weather indicators like wind speed or solar radiation.

In principle, this could be done by modelling the weather dependence of quantiles as in version 1.0. However, because incorporation of weather dependence makes the model more complex, other aspects need to be simplified to prevent over-fitting of the model to the available data. In this case, we replace explicit modelling of conditional temperature distributions (conditioned on the non-temperature covariates) by a model predicting the daily temperatures, followed by a re-scaling of the spread in predicted temperatures to correct their distribution.

In de Valk and Brandsma (2023), a flexible nonparametric regression model is used, predicting daily TN or TX at the new site based on measurements at the old site and additional weather-related variables. These variables are (subsets of)

U	wind vector [m/s],
N	cloud cover [okta],
HUM	specific humidity [g/kg],
SST	sea surface temperature [deg C] (only for stations near the coast),

TN	model	RMS	r^2
den Helder → de Kooy	$\{T_{src}\} + \{U, N\} + \{SST\}$	0.96	0.97
Souburg → Vlissingen	$\{T_{src}, SST\} + \{U\} + \{N\}$	0.60	0.99
Maastricht → Beek	$\{T_{src}\} + \{U\} + \{HUM\}$	0.62	0.99
Groningen → Eelde	$\{T_{src}\} + \{U, N\}$	0.93	0.98

Table 3.1: Estimated/selected models for homogenization of daily minimum temperature TN with cross-validation root mean squared error RMS (deg C) and r^2, the fraction of the variance predicted in cross-validation.

TX	model	RMS	r^2
den Helder → de Kooy	$\{T_{src}, HUM\} + \{U, SEAS\}$	0.57	0.99
Souburg → Vlissingen	$\{T_{src}, SST, U\}$	0.45	1.00
Maastricht → Beek	$\{T_{src}\} + \{HUM\} + \{U, N\}$	0.51	1.00
Groningen → Eelde	$\{T_{src}, HUM\} + \{U\} + \{N\}$	0.50	1.00

Table 3.2: As Table 3.1 for TX.

SEAS season (represented by a vector on the unit circle).

These variables have been measured during the years covered by the temperature measurements, and they are potentially useful to explain and predict differences in temperature between the old site/instrument and the new site/instrument. Further details of the data, pre-processing and homogenization are found in Section 2.3

The spread in predicted temperature is inflated to obtain accurate temperature distributions, and cross-validation is employed to obtain a robust selection of covariates and their interactions. We refer to this model by the acronym GAM (generalized additive model), which is the technical term for the class of flexible regression models employed (Wood, 2020). GAM is a versatile and widely used non-linear regression technique.

Tables 3.1 and 3.2 list the covariates selected for each station; braces separate the groups of covariates for which interactions have been estimated (the contributions of different groups are simply added). In almost all cases, the physical covariates appear to be better predictors than season.

de Valk and Brandsma (2023) find that the fitted models exhibit realistic features, such as stronger nighttime cooling at inland locations under calm, clear conditions.

TG	model	RMS	r^2
den Helder → de Kooy	$\{T_{src}, \Delta TX, \Delta TN\} + \{SEAS\}$	0.39	1.00
Souburg → Vlissingen	$\{T_{src}\} + \{SEAS, \Delta TX, \Delta TN\}$	0.30	1.00
Maastricht → Beek	$\{T_{src}, SEAS\} + \{\Delta TX, \Delta TN\}$	0.29	1.00
Groningen → Eelde	$\{T_{src}\} + \{SEAS, \Delta TX, \Delta TN\}$	0.31	1.00

Table 3.3: As Table 3.1 for TG.

An additional advantage of the method is that the precision of the homogenization of individual observations can be estimated from the fitted GAM model, using a variance adjustment to account for the serial dependence of the daily observations; see Appendix C of de Valk and Brandsma (2023).

For the homogenization of daily mean temperature TG, a similar GAM model was developed, which uses season and the adjustments ΔTN of TN and ΔTX of TX (both in deg C) as optional covariates; see Table 3.3 for the selected models and cross-validation statistics. The motivation for this choice is that the information of the physical covariates is already used in the homogenization of TN and TX, so the resulting adjustments of TN and TX already contain this information. Furthermore, using these adjustments improves the consistency of TG with TN and TX. The covariate selection for TG did not drop any covariates; only the choices of interactions to be modelled differ among the stations.

A comparison between the root mean squared (RMS) residuals of version 1.0 (v1) and version 2.0 (GAM) for the four stations with parallel measurements is shown in Table 3.4.

	TN		TX		TG	
station	v1	GAM	v1	GAM	v1	GAM
De Kooy	1.08	0.91	0.63	0.54	0.47	0.37
Eelde	1.11	0.90	0.55	0.47	0.36	0.28
Beek	0.65	0.62	0.53	0.49	0.31	0.28
Vlissingen	0.82	0.56	0.52	0.38	0.39	0.25

Table 3.4: Residual root mean squared error (°C) of homogenization of daily minimum temperature TN, daily maximum temperature TX and daily mean temperature TG using version 1.0 (v1) and version 2.0 (GAM), determined from the parallel measurements.

Table 3.4 shows that GAM consistently reduces residual RMS errors across all stations and variables relative to version 1.0. The reductions in mean square error are generally modest: between 6% and 76%.

The residual RMS values for version 2.0 (GAM) are only slightly lower than the cross-validation RMS error estimates in Tables 3.1 and 3.3. As even the latter are uniformly lower than the residual RMS errors of version 1.0, version 2.0 is more accurate than version 1.0 in RMS.

For a more detailed impression of the differences between version 2.0 and version 1.0 for the stations with parallel measurements, Figure B.1–B.3 in Appendix B shows the annual minimum of TN (TNn), the annual maximum of TX (TXx) and the annual mean temperature TGg from version 1.0 (v1), version 2.0 (GAM), and the parallel measurements at the new site. For TNn, the differences between the versions are consistent with more extreme nighttime cooling when moving away from the coast (to De Kooy or to Souburg) or away from the city (to Eelde). For TXx, the changes may reflect a dampening effect of surrounding vegetation on extreme daytime temperature maxima (De Kooy, Eelde). For the extremes TNn and TXx at De Kooy and Eelde and for TNn

at Vlissingen, it is difficult to ascertain for every year that version 2.0 is an improvement over version 1.0, because the temperature range of the parallel measurements at the old site does not fully cover the range of the preceding measurements to be homogenized. For TGg, the three versions are almost indistinguishable.

Further comparisons of unadjusted temperatures and version 1.0 (`hom_v1`) and version 2.0 (GAM) homogenized temperatures are found in Appendix D, showing aggregated values of 15 climate indices listed in Appendix A. Most of these 15 indices relate to extreme weather: annual maxima and minima, heatwaves, cold waves, etc.

Furthermore, Appendix E shows nonlinear trend estimates of these indices from the different versions, and Appendix F shows additional details on the sampling uncertainty of the trends from version 2.0 data.

The nonlinear trends shown in Appendix E and Appendix F are determined using local linear regression (LOESS) using a standard tricube kernel with support of 42 years (which gives almost the same variance as a 30 year average); see de Valk (2020); Scherrer et al. (2024).

Confidence intervals in these figures represent the sampling uncertainty. They are determined by the normal approximation using an estimate of the variance, truncated to zero where needed. The variance of the sampling uncertainty due to homogenization is estimated using a block bootstrap method (Kunsch, 1989), and the variance of nonlinear trend estimates due to year-to-year variability is obtained from the local linear regression estimator, which assumes that fluctuations of annual values around the true trend line are white noise.

Aggregated indices (Figure D.1) differ little between the two versions of the homogenized data. The same applies to the trends (Figures E.1–E.4). The 95% confidence intervals of the trend lines show that the version 2.0 homogenization does not noticeably increase the sampling uncertainty in the trend line estimates, which is already substantial without homogenization; see Appendix F showing the total sampling uncertainty and the sampling uncertainty due to homogenization.

3.4 Version 2.0 without parallel measurements

3.4.1 Issues with the use of a reference time series instead of parallel measurements

For De Bilt no complete parallel measurements exist for the combination of the screen change and the almost simultaneous relocation. Therefore, the adjustment for these changes is derived from two different models predicting the temperature (TN, TX or TG) at De Bilt from the temperature at one or more chosen reference station(s): one model calibrated to the data gathered before the known breakpoint, and another calibrated to the data gathered after the breakpoint. Then the differences between the predictions of these two models are used to adjust the data before the break. This approach assumes that with ideal measurements (no sensor change or relocation), temperature differences between De Bilt and the reference data would have remained the same over the calibration intervals (see Section 3.2 for a discussion of this assumption as it applies

to the method used for version 1.0).

Adjustment using data from reference stations in this manner has two fundamental limitations:

1. For the reference station(s), we can choose from Vlissingen, De Kooy, Eelde and Beek: data from these stations are sufficiently quality-controlled and homogenized to be used for long-term monitoring. However, they are all far away from De Bilt, so their temperatures are often quite different from the temperature at De Bilt, or in other words, there correlations with station De Bilt are low. As a consequence, the two predictions of the temperature at De Bilt (calibrated on measurements before and after the breakpoints) are quite uncertain, and subtracting these two uncertain predictions to derive temperature adjustments (which are themselves relatively small) further amplifies the uncertainty. In particular, this affects the reliability of adjustments of extreme temperatures and derived extreme temperature indices such as the number of heatwaves or the number of cold waves.

2. Because adequate parallel measurements are missing at the Bilt, the temperature adjustments cannot be validated, so we cannot check how good they really are. All we can do is test the method(s) on a different dataset for which parallel measurements do exist, and choose a method which works well on that dataset. In addition, we can check the robustness of the method with respect to its chosen parameter settings: a small change in a setting should not have a large impact on the adjustments.

Because of these two limitations, the homogenization of the temperature data of De Bilt is more challenging and less certain than the homogenization for the four other stations with parallel measurements.

3.4.2 Method

For De Bilt, modelling of covariate dependence of the relation between temperature at this station and temperature at a reference station chosen from the H4 does not yield significant improvements when tested on a proxy dataset; see Appendix G. This can be explained by the relatively large distances between the H4 stations and De Bilt, which result in relatively low temperature correlations, and, consequently, high noise levels (where noise denotes the unpredictable component of De Bilt's temperature signal). This noise overwhelms modest model improvements, but most importantly, it impedes accurate calibration of more complex models like the GAM models used for the H4 stations; see Appendix G for further details of the test.

Therefore, for version 2.0, we test and improve the QDM method used earlier for version 1.0 (Section 3.2). The QDM method requires the choice of:

1. a reference time series: the data from one of the four stations Vlissingen, De Kooy, Eelde and Beek with quality-controlled homogenized temperature data, or a time series constructed from a combination of two or more of these stations;

2. the calibration interval: the length of the sub-records before and after the breakpoint which are used for calibration.

We consider the following eight options for the reference time series: (1) Beek, (2) Eelde, (3) average Beek and Eelde, (4) De Kooy, (5) Vlissingen, (6) average De Kooy and Vlissingen, (7) average Eelde and Vlissingen, and (8) average all four H4 stations (the version 2.0 homogenized series). Table 3.5 shows the Pearson correlation coefficients of the differenced reference time series with the differenced time series of De Bilt. Their values do not support skipping any of these reference series.

The reason for considering averages of temperatures at several stations is that these averages may be more representative for temperatures at De Bilt than the data of individual stations, since all four reference stations are far from De Bilt. Also, averaging may reduce the variation in the daily temperature differences between the reference data and the data from De Bilt (or equivalently: increase the correlation between the reference data and the data from De Bilt). This is confirmed by the correlation coefficients in Table 3.5.

For the calibration interval, we consider periods of 5, 10, 15, 20, and 25 years (somewhat extending the range considered in Dijkstra et al. (2022). Short intervals have the advantage that violations of the assumption (3.1) underlying the QDM method tend to have less impact on the outcomes, but long intervals have the advantage of higher precision.

reference series	TN	TX	TG
Beek	0.70	0.75	0.84
Eelde	0.72	0.77	0.85
Beek/Eelde	0.81	0.85	0.91
De Kooy	0.67	0.75	0.83
Vlissingen	0.66	0.77	0.84
De Kooy/Vlissingen	0.76	0.84	0.90
Eelde/Vlissingen	0.81	0.87	0.93
All	0.84	0.89	0.94

Table 3.5: Pearson correlation coefficients of differenced reference series with the differenced time series of De Bilt.

In the present context, a robust model is a model which is not sensitive to the exact choice of the calibration interval or to the reference time series.

To some extent, this is related to precision: imprecise models, which are sensitive to the values of the data sample used to estimate them (in theory: to the choice of the ensemble member), are generally also sensitive to other choices, like in this case the calibration interval and reference time series. However, also bias (usually due to violation of assumptions underlying the model) can cause differences between models estimated using different choices, even if these choices should not matter, in theory. Here, we focus on robustness, as we want to avoid bias even more than we want to reduce sampling uncertainty.

To evaluate robustness, we consider the 15 annual climate indices listed in Appendix A, which encompass mean indices (TGg, TXg, TNg) as well as indices of extremes (TGn,

TGx, TNn, TX3n, TXx, TX3x, no. of summer days, tropical days, heatwaves, ice days, severe frost days, cold waves). We aggregate these annual indices to temporal averages (for temperatures) or sums (for day counts) over 1907-1950. Figure H.1 in Appendix H shows the estimates of these aggregated indices and their confidence intervals for different choices of calibration interval and reference time series. These plots already provide a visual indication of robustness, but it is not easy to recognize patterns in all these lines, so a more structured assessment would be desirable.

To find out which choice(s) of reference time series lead to robust models, we compute for each reference time series and each climate index the sample variance of the 5 values of the index for the 5 calibration interval lengths. From this, we derive for each reference time series two types of metrics:

- MINSUM: the number of indices for which the variance is the lowest over all reference time series,
- RANKSUM:
    a. for each climate index, we rank the variances corresponding to the 8 reference time series, with rank defined as increasing with decreasing variance (these ranks are normalized, in contrast to the original variances);
    b. then for each reference series, we sum these ranks over the climate indices.

A high score for a given reference time series indicates that the derived indices are relatively insensitive to the choice of calibration interval, reflecting a robust homogenization.

	robustness		precision	
reference series	MINSUM	RANKSUM	MINSUM	RANKSUM
Beek	0	53	0	42
Eelde	0	61	0	50
Beek/Eelde	8	96	5	104.5
De Kooy	0	44	0	22.5
Vlissingen	2	65	0	49.5
De Kooy/Vlissingen	2	64.5	0	68
Eelde/Vlissingen	1	71	4	96
All	2	85.5	6	107.5

Table 3.6: Robustness and precision metrics MINSUM and RANKSUM (see text) for eight reference time series.

The results are found in Table 3.6 under "robustness". Using data from individual stations, Vlissingen, Eelde and Beek come out best. However, we see that averaging of time series from multiple stations generally gives more robust estimates, as could be expected, given that this makes them more precise (see Table 3.5). Both MINSUM and RANKSUM indicate that the most robust combination of stations is Beek/Eelde.

In order to find out to what extent the values of the robustness metrics can be simply explained by sampling uncertainty, we can compute MINSUM and RANKSUM metrics expressing precision similar to those expressing robustness. To express precision, instead of computing for each reference time series and each aggregated index the sample variance of the index over the different calibration intervals, we compute for each reference time series and each index the mean of the estimated variances corresponding to the different calibration intervals (estimated previously using the block bootstrap). These mean variances are then converted to metrics as above, and listed under "precision" in Table 3.6.

The precision metric agrees closely with the robustness metrics: for RANKSUM, Kendall's tau is 0.79 and the Pearson correlation coefficient is 0.93, which are both very high. The main difference is that for the precision metric, the average of all H4 stations as reference gives the highest value, as expected.

That the combination Beek/Eelde scores high in robustness is partly explained by its also high precision metric. That it scores higher than the average of all H4 stations seem to indicate that (the) coastal stations contribute relatively high sampling uncertainty or bias. Indeed, all combinations involving De Kooy score relatively low in robustness and in precision, so it appears that this station is better avoided.

	robustness		precision	
calibration interval	MINSUM	RANKSUM	MINSUM	RANKSUM
5 years	0	16	0	19
10 years	0	36	0	29
15 years	8	61	2	47
20 years	1	49	0	59
25 years	6	63	13	71

Table 3.7: Robustness and precision metrics MINSUM and RANKSUM (see text) for five calibration intervals.

To check which calibration intervals give the most robust estimates, we can simply switch the roles of calibration interval and reference series in the procedure above. The outcomes are listed in Table 3.7 under "robustness". They indicate that a calibration interval of at least 15 years is required to obtain robust estimates. Beyond this, there is no systematic improvement, so there seems to be no reason to choose a calibration interval longer than 15 years.

This is at odds with the trend in the precision metric in Table 3.7 which increases monotonically with the length of the calibration interval (which is expected, as precision generally improves when a larger sample of data is used for estimation). This points to a substantial departure from the time-invariance assumption 3.1 for some reference time series, as that would increase spread among the outcomes of different series and should affect estimates using long calibration intervals more than estimates using short calibration intervals. This possibility is a reason to exclude calibration intervals longer than 15 years.

Apart from this issue, the strong agreement between the robustness and precision metrics shows that most of the variation in aggregated indices among the different reference series and calibration intervals is due to the variation in sampling uncertainty among the alternatives, or in other words, that precision is the main determinant of the observed variation in the robustness metric.

That does not leave much room for bias as a determinant, which is therefore either relatively small, or varies little among reference series and calibration intervals of up to 30 years. The latter option seems unlikely (it would mean that De Bilt is the outlier among the five stations and homogenization using reference series would be impossible), so the results are most compatible with a relatively small bias in comparison to the sampling uncertainty.

This has an advantage: since we cannot use parallel measurements for De Bilt, bias and root mean squared error cannot be estimated directly, but the standard deviation (representing only the sampling uncertainty) still provides a fair indication of the root mean squared error.

This is a general statement applying to the fifteen climate indices in Appendix A lumped together, and does not necessarily hold for every climate index individually.

Additional detailed comparisons of unadjusted temperatures and version 1.0 (`hom_v1`) and version 2.0 (Beek/Eelde) homogenized temperatures are found in Appendix I, showing nonlinear trends of 15 climate indices listed in Appendix A. In particular, Figure I.2 shows that the version 2.0 homogenization based on the Beek/Eelde reference series and 15-year calibration intervals does not cause a large increase in the sampling uncertainty in the trend line estimates, which is already substantial without homogenization.

Chapter 4

Results

This chapter presents the main outcomes of the homogenization process. We first evaluate how version 2.0 affects monthly mean temperature differences between old and new measurement sites. Next, we examine the impact on annual temperature indices and long-term trends, both for individual stations and at the country scale. Particular attention is given to differences between version 1.0 and version 2.0 of the homogenization, and to the remaining uncertainties, especially for De Bilt for which no parallel measurements are available.

4.1 Impact on monthly mean differences

First, we examine how version 2.0 of the homogenization affects the differences in monthly mean TN, TX, and TG between the old and new sites, as shown in Figure 2.8.

For the H4 stations, Figure 4.1 (top and middle rows) shows that monthly mean differences in TN, TX and TG have become very small. This is noteworthy: 7 of the 8 eight models used for adjusting TN and TX do not contain season as covariate (see Tables 3.1-3.2), so the use of physical covariates like wind and humidity also effectively removes the seasonality of the effect of station relocation shown in Fig. 2.8.

For De Bilt, Figure 4.1 (bottom) shows larger residual differences in monthly means. Lacking parallel measurements at De Bilt, it shows the monthly difference between homogenized values of TN, TX and TG before and after the breakpoint minus the same difference from the mean of the H4 stations. However, for the homogenization of De Bilt, deviations from the mean homogenized values of TN, TX and TG of Eelde and Beek are used (see Section 3.4). Therefore, the residuals in Figure 4.1 (bottom) reflect the effect of choosing a reference time series different from the one chosen in this report. These residuals are still small: their root mean squared values are 0.17° for TN, 0.22° for TX and 0.12° for TG.

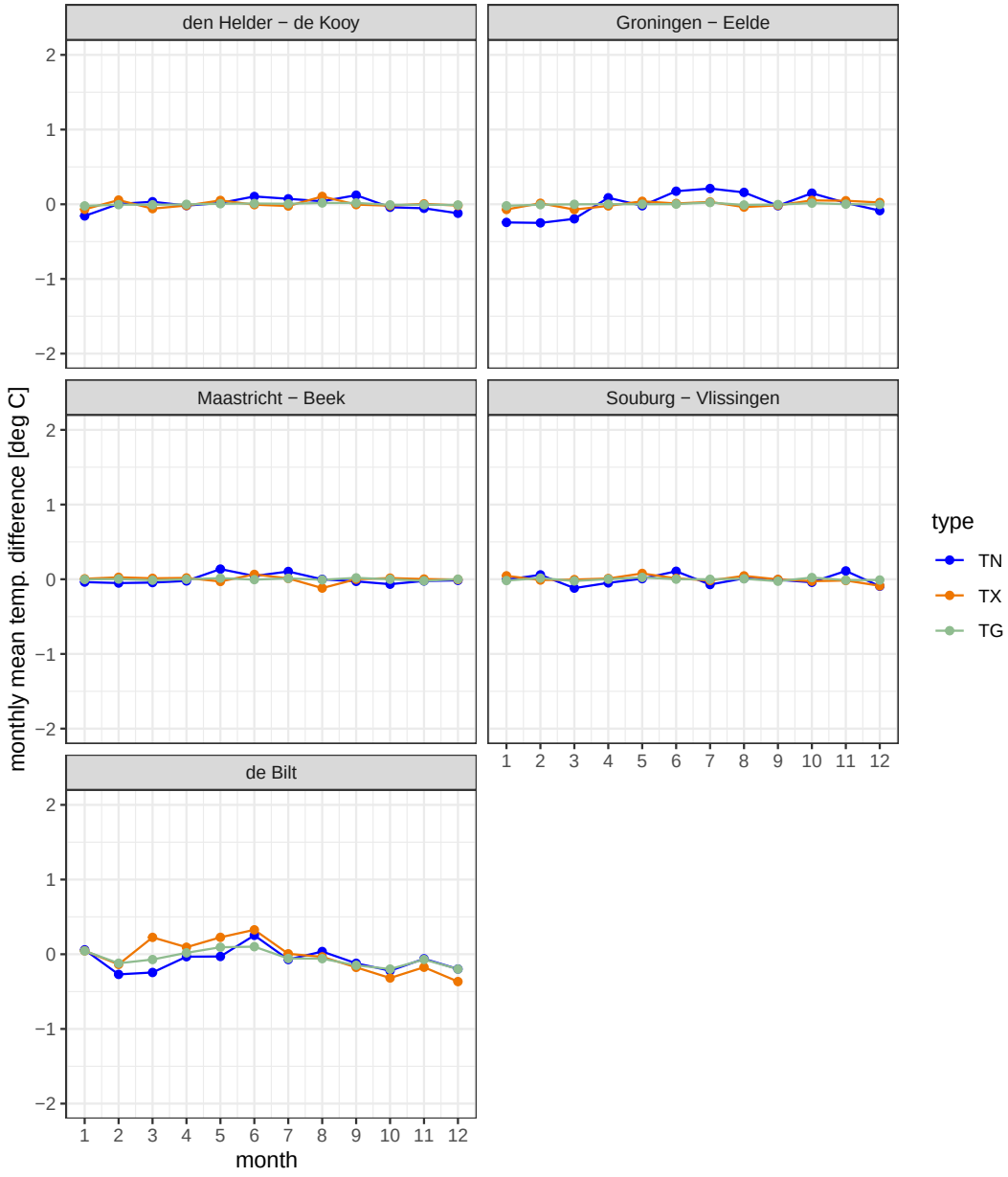

Figure 4.1: Monthly mean differences in homogenized TN, TX and TG (version 2.0) at the old site and the new site from the parallel measurements of the H4 stations (top and middle). For De Bilt (bottom): monthly mean differences of homogenized (version 2.0) TN, TX and TG of De Bilt over 15-year intervals before and after the breakpoints minus the corresponding differences of the mean of homogenized (version 2.0) TN, TX and TG over the four H4 stations.

4.2 Impact on temperature quantiles

Figs. C.1–C.3 show the difference between a quantile of the homogenized temperature at the new site and the corresponding quantile of the measured temperature at the old site (the *quantile adjustment*), as function of the latter quantile.

The lines concern the homogenized data, so the years with parallel measurements at both sites are excluded (as for these, the data from the new site have been used).

In addition, the orange dots show the difference between the empirical quantiles at the new and the old site, derived directly from the parallel measurements from which the GAM model has been estimated (see Section 3.3).

In several cases, the versions 1 and 2 of the homogenization for the H4 stations differ considerably in the extreme ranges of the temperature scales. Toward the lowest and highest temperatures, the quantile adjustments by version 1.0 (green) tend to level off to a constant value. For version 2.0 (red), we see increasing divergence from version 1.0, with temperature adjustments changing roughly linearly in the original measured temperature and often displaying larger fluctuations than seen in the curves for version 1.0.

The deviation of version 2.0 from version 1.0 may seem suspicious when we look at Figs. C.1–C.3 through the lens of quantile matching as applied in version 1.0 (see Section 3.2). However, it can be understood better if we look in more detail into what these two methods do.

Version 1.0 basically attempts to match the empirical quantile differences from the parallel measurements (the blue dots in these plots), with the only refinement being that this is done per season. To improve robustness, the estimated monthly quantile adjustments are required to be constant for the lowest and highest 5% of temperatures. The need for this also stems from the use of the standard LOESS smoother (with degree 2, performing local quadratic regression), which can result in wild (quadratic) divergence of quantile adjustments toward and beyond the measured temperature extremes.

However, there is no physical reason why a difference in temperature quantiles at two adjacent stations should tend to a constant value at the extremes of the temperature range.

In contrast, version 2.0 predicts the daily temperature values at the new site using flexible models which have linear tendencies near and beyond the extremes, extrapolating the tendencies present in the observations (the thin-plate spline model; see de Valk and Brandsma (2023)). This is inherently more robust than local quadratic regression (for the same reason, linear regression is still ubiquitous in statistics). Furthermore, the model is made more robust by using cross-validation for the selection of covariates and their interactions, and by using a similar technique (marginal likelihood) to estimate the smoothness (see Fong and Holmes (2020)).

The empirical quantile differences from the parallel measurements (the orange dots in Figs. C.1–C.3) are generally matched well by the version 2.0 curves (blue), also in the extreme ranges. This cannot be said of the version 1.0 curves (green). This indicates that the linear tendencies of this type of model near and beyond the extremes gives a good approximation of the data.

The fluctuations in the quantile adjustments by version 2.0 at the ends of the

temperature range in Figs. C.1–C.3 are not due to lack of robustness or smoothness (the version 2.0 models are very smooth in the multidimensional space of all covariates), but reflect the impact of non-temperature covariates on the daily temperature values (there are only a few data points in these ranges, so the effects of the covariates are not averaged out in the computed quantiles).

4.3 Effects on annual temperature indices and trends

Next, we examine how the homogenization affects annual indices and long-term trends. In appendix J, values of annual minimum temperature TNn derived from homogenized TN, annual mean temperature TGg derived from homogenized TG, and annual maximum temperature TXx derived from homogenized TX are shown for all five stations, and their averages over the stations are shown as H5 mean.

The nonlinear trends shown are determined using local linear regression; see Section 3.3 for a description.

Due to substantial year-to-year variability in TNn and TXx, the uncertainties in the estimates of their trends are relatively large. In contrast, TGg fluctuates much less and the trends are more precise, and vary less among the stations.

Figure 4.2 shows for TXx, TGg and TNn the trends from the original data and from version 1.0 (`hom_v1`) and version 2.0 (new) of the homogenization:

- For all stations, homogenization tends to make trends for different stations more similar.
- For TGg, version 2.0 gives practically the same trends as version 1.0.
- For TXx and TNn, the long-term temperature increases at De Kooy and Eelde in the new version are slightly larger than in version 1.0.
- For TXx, the long-term temperature increase at Bilt is considerably smaller in version 2.0 than in version 1.0; for TGg, the difference is much smaller and for TNn, the effect of homogenization is negligible.

Differences in the nonlinear trends of other climate indices between the versions are found in Appendix E and Appendix I (see Appendix A for explanations of the indices). For the H4 stations, these differences are small (see Appendix E). For De Bilt, the version 2.0 trends in the "warm" indices TGx, TXx, TX3x, no. of summer and tropical days and no. of heatwaves are about halfway the version 1.0 trends and the trends from the original data (see Figure I.1).

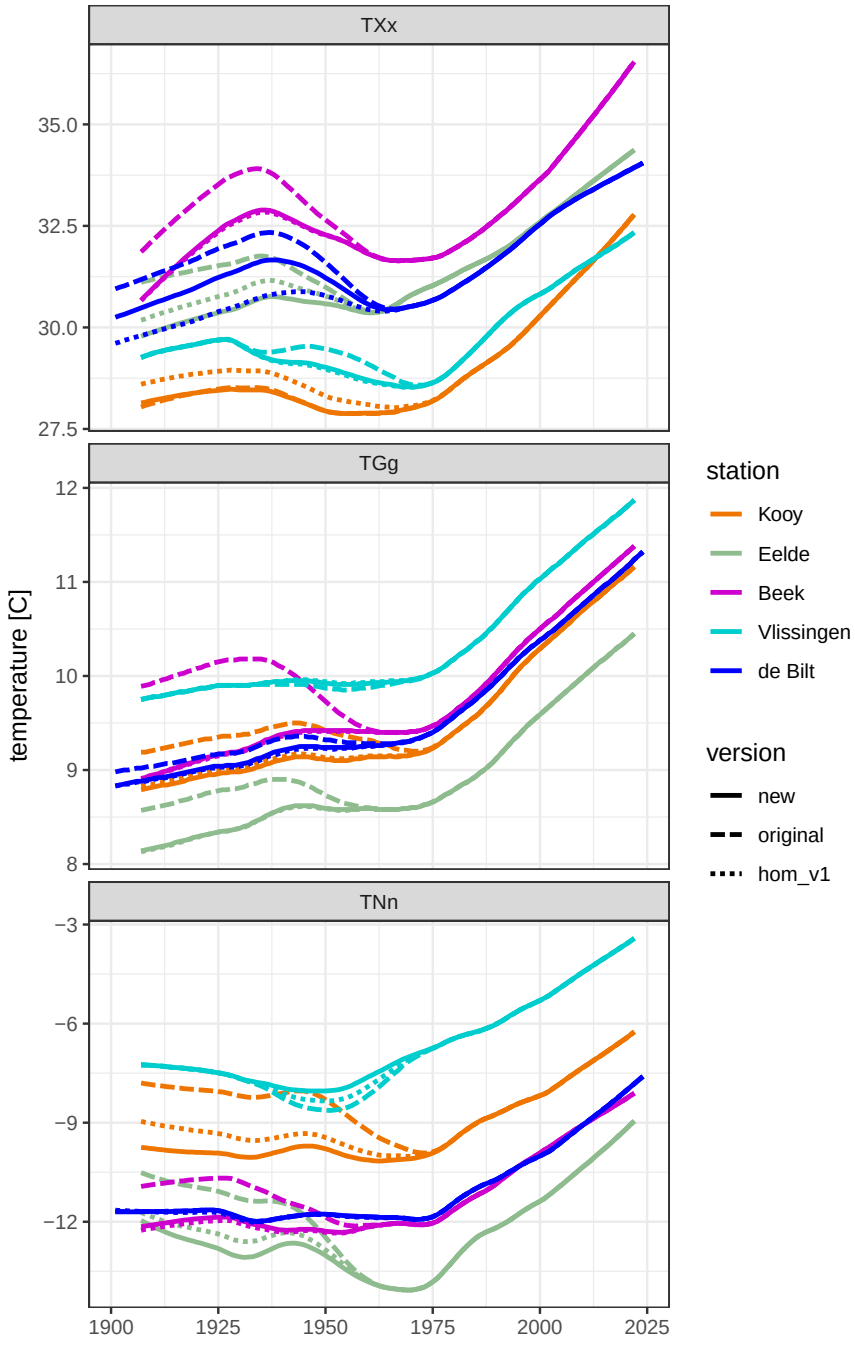

Figure 4.2: Nonlinear trends of annual maximum temperature TXx, annual mean temperature TGg and annual minimum temperature TNn from the original measurements and the data homogenized by version 1.0 (`hom_v1`) and version 2.0 (new), for all five stations.

period	length [year]	original	version 1.0	version 2.0
1901–1950	50	23	7	14
1951–2000	50	9	9	9
2001–2025	25	16	16	16
1901–2025	124	48	32	39

Table 4.1: Numbers of heatwaves in different periods derived from the original data and from version 1.0 and version 2.0 homogenized data of De Bilt.

The resulting heatwave counts for De Bilt[1] for three periods are shown in Table 4.1 (last column). The original data indicated a steep decline in heatwave frequency from the first to the second half of the 20th century and an even steeper increase after that. The decrease over the 20^{th} century is less steep in version 2.0 of the homogenized data.

Compared to version 1.0, the number of heatwaves over 1901-1950 has doubled in version 2.0 from 7 to 14.

These values should be interpreted with caution, as heatwave counts exhibit high year-to-year variability and their counts and long-term trends are highly uncertain; see Fig. H.1 (N.heatwave, Beek/Eelde, 15 years) and Fig. I.1 for the long-term trend. This is because heatwaves are fairly rare, but also because the probability of a heatwave in a given year (and therefore also the number of heatwaves in a given period) is very sensitive to a small adjustment of the temperature.

For cold waves, there is little difference between the versions of the data for De Bilt: in version 2.0, one additional cold wave is found before 1951.

4.4 Differences between the long-term trends at different stations

Comparing the long-term trends at different stations from version 2.0 in Figure 4.2, we observe:

- For TGg, trends for the different stations are very similar.
- Before 1975, the trend in TXx shows a distinct oscillation at Beek and De Bilt – with a peak around 1935 – but is flatter at the other three stations.
- The pronounced dip in TNn at Eelde between 1950 and 1980 appears inconsistent with trends at other stations. However, it could indicate a genuine spatial difference: the trend line for the nearest station De Kooy also shows a slight dip there. Relative to version 1.0, the depth of the dip at Eelde has become a little smaller.

[1]For the Netherlands, KNMI determines heatwaves based on the data of only De Bilt; see Appendix A. There is no internationally agreed standard of what constitutes a heatwave.

4.5 Differences between annual extremes and annual means

The Figures in Appendix J indicate a steeper increasing trend in the annual extremes TNn and TXx than in the annual means TGg. To see this in more detail, the top rows of Figure 4.3 show (a) the trend in TXx and, superimposed on this, (b) the trend in the annual mean daily maximum temperature TXg, shifted upwards to have the same mean as the trend in TXx; separate plots are shown for each station and for their average (H5 mean). The bottom rows show the trend in TNn and the downward shifted trend in TNg.

Compared to the trends in the annual means TNg and TXg, the trends in the annual extremes TNn and TXx show more pronounced long-term oscillations. Note that these oscillations are fairly coherent across stations.

The 95% confidence intervals for the trends in TNn are wide, so they seem compatible with the trends in TNg. For TXx, the confidence intervals are narrower, but they nevertheless contain the trend lines for TXg for almost all years except the most recent. It should be noted that nonlinear trend estimates are generally less reliable near the start and end of the time series. In particular, the local linear regression method used for trend estimation tends to straighten the curves there, exaggerating differences. In this respect, it is worth noting that the changes between 1925 and 2000 in annual extremes and in annual averages are very similar.

4.6 Discussion of recent trends and uncertainties

In Vautard et al. (2023), the excess of the increase in TXx since 1950 above the increase in TXg (particularly high in a region extending from Western/Northern France to the Netherlands) is attributed to a change in circulation, leading to more frequent southerly flow patterns[2]. Other factors such as aerosols (Schumacher et al., 2024) and hydrology (groundwater table, drainage) may also have contributed to deviations between the trends in annual extremes and annual averages.

To characterize current (2025) trends in annual extremes of temperature reliably, we will need in the order of twenty additional years of measurements (one half of the width of the kernel used in the local regression to determine the trends.).

Some further checks of the version 2.0 homogenized time series are reported in Appendix K. These involve the application of an independent homogenization method to the homogenized time series to see if any additional inhomogeneities can be detected, and checks on the daily temperature range and asymmetry of the daily temperature distribution.

[2]The spatial patterns of the total and dynamical contributions to trends in TXx and TXg in Figure 1 of Vautard et al. (2023) are strikingly similar, in spite of the large differences in magnitude of the estimated total and dynamical contributions. This may indicate that the dynamical contributions have been underestimated in this article, e.g. due to noise and/or to the reduction of circulation to a single variable, whereas circulation may be more complex.

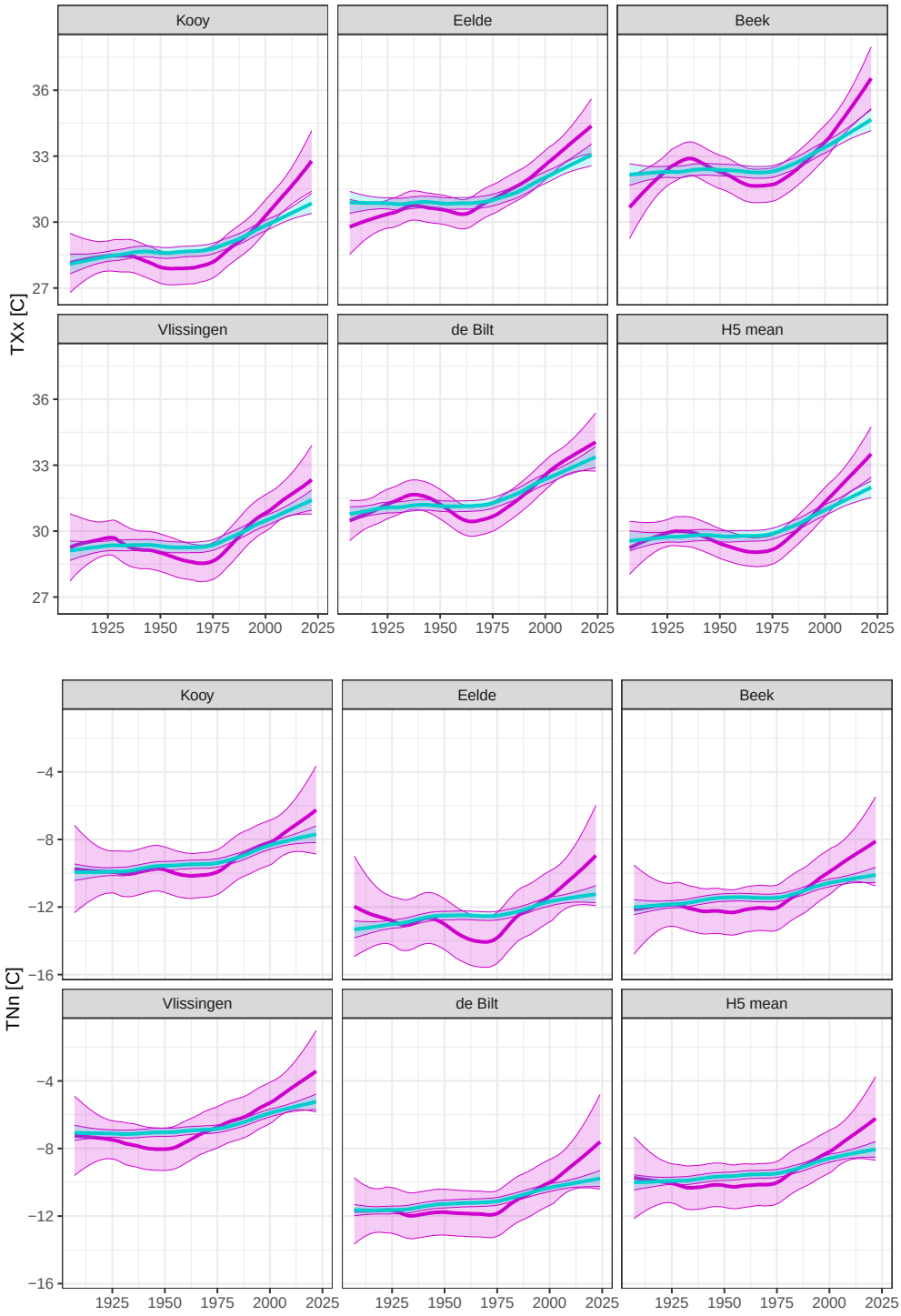

Figure 4.3: Top rows: Nonlinear trends (version 2.0) in annual maximum temperature TXx (magenta) and trend in annual mean daily maximum temperature TXg shifted to have the same mean (cyan), both with their 95% confidence intervals. Bottom rows: same for TNn (magenta) and shifted TNg (cyan).

Overall, these checks support the robustness of the homogenized temperature series, while also highlighting a few localized or recent changes that warrant continued monitoring or documentation.

Chapter 5

Key results and conclusions

This chapter summarizes the main findings from the homogenization of daily minimum, maximum, and mean temperature data (TN, TX, TG) for the five principal stations in the Netherlands. Conclusions are grouped by station type (H4 stations, De Bilt), overarching effects across all stations, and practical guidance for data users.

Homogenization of TN, TX and TG from the H4 stations (De Kooy, Eelde, Vlissingen and Beek/Maastricht Airport)

1. In version 2.0, which incorporates variables such as wind, humidity, cloud cover, sea surface temperature, and/or season, day-to-day variations during periods with parallel measurements are better captured compared to version 1.0. However, the improvement is modest overall (reductions in mean square error between 6% and 76%)
2. Also climate indices and their long-term trends in version 2.0 differ little from those in version 1.0 (Figs. B.1–B.3, D.1, E.1–E.4), suggesting that version 1.0 was already adequate (Table 3.4).
3. The version 2.0 adjustments of temperature extremes are more consistent with tendencies present in the calibration data than those of version 1.0 (Figs. C.1–C.2). However, for the extremes of TNn and TXx at De Kooy and Eelde and of TNn at Vlissingen, the parallel measurements do not fully capture the full range of temperatures to be adjusted, making it difficult to confirm improvements unambiguously (Figs. B.1–B.2).
4. The sampling uncertainty introduced by homogenization is relatively small, even for indices of extremes.

Homogenization of TN, TX and TG from De Bilt

5. Due to the lack of sufficient parallel measurements at De Bilt, homogenization relies on reference station data. Consequently, the homogenization is less precise than the homogenization for the H4 stations.

6. Attempts to model the relationship between temperatures at De Bilt and at reference stations using non-temperature covariates such as wind, humidity, and cloud cover did not improve the homogenization. Therefore, version 2.0 builds on a refined version of the QDM method used in version 1.0.

7. Using the average of the homogenized temperatures from Eelde and Beek as reference and selecting 15-year calibration intervals before and after breakpoints produces robust and relatively precise adjustments.

8. Although the lack of parallel measurements prevents direct validation, the similarity between robustness and precision metrics across tested variants suggests that potential bias is small compared to sampling uncertainty.

9. Concerning indices for warmth at De Bilt (TXx, TGx, and numbers of summer days, tropical says, and heatwaves), the nonlinear trends from version 2.0 differ substantially from the trends of version 1.0 and from the trends of the unadjusted data, so introducing a new version appears to be justified. In particular, the number of heatwaves in 1901-1950 is now estimated to be 14, twice as many as in the previous version. However, the long-term trends of these indices from the two versions cannot be conclusively distinguished, due to the large natural year-to-year variability of these indices. For other indices, differences from unadjusted data and version 1.0 are very small; see Appendix I.

Homogenization of TN, TX and TG for all stations

10. The homogenization appears to align trends across stations, improving spatial consistency (Fig. 4.2).

11. Homogenization can introduce errors if the models used are not calibrated with sufficient precision, for example when calibrated on a dataset that is too small. We find that the version 2.0 homogenization has been calibrated with sufficient precision; the calibration has little impact on the overall precision of long-term climate trends, which is determined mainly by the natural year-to-year variability y (Figs. F.1–F.4, I.2).

12. At all stations, TNn and TXx have increased more than the corresponding annual means (TNg and TXg). These differences may be related to low-frequency oscillations visible in the annual extremes but less in the annual averages.

Guidance for use

13. Bias introduced by homogenization appears to be small compared to the sampling uncertainty. Therefore, confidence intervals based on sampling uncertainty offer a reasonable indication of homogenization accuracy.

14. The sampling uncertainty due to homogenization does not include the random component (not predictable from available measurements) of the difference in temperature between the old site and the new site, which is larger.

15. For estimating long-term temperature trends, the impact of homogenization on trend uncertainty can be safely ignored.

16. Users are advised to consult the stated limitations of homogenization in Section 3.1.

Bibliography

Aguilar, E., Auer, I., Brunet, M., Peterson, T. C., Wieringa, J., et al. (2003). Guidelines on climate metadata and homogenization. Wcdmp no. 53, wmo-td no. 1186, World Meteorological Organization, Geneva, Switzerland.

Brandsma, T. (2011). Parallel air temperature measurements at the knmi observatory in de bilt (the netherlands) may 2003 - june 2005. Technical Report WR-2011-01, Royal Netherlands Meteorological Institute.

Brandsma, T. (2016a). Homogenisatie van dagelijkse temperaturen van de knmi hoofdstations. *Meteorologica*, 25(2):4–8.

Brandsma, T. (2016b). Homogenization of daily temperature data of the five principal stations in the netherlands (version 1.0). Technical Report TR-356, Royal Netherlands Meteorological Institute.

Brandsma, T. (2019). Pagode metingen. *Meteorologica*, 28(1):5–8.

Brandsma, T. (2022). Effects of lowering thermometers screens in the netherlands around 1960. Technical Report TR-400, Royal Netherlands Meteorological Institute.

Brandsma, T. (2025). Impact of a solar farm on temperature and radiation measurements at groningen airport eelde in the netherlands. Technical Report TR-25-01, Royal Netherlands Meteorological Institute.

Brandsma, T., Können, G., and Wessels, H. (2003). Empirical estimation of the effect of urban heat advection on the temperature series of de bilt (the netherlands). *International Journal of Climatology*, 23(7):829–845.

Cannon, A. J., Sobie, S. R., and Murdock, T. Q. (2015). Bias correction of gcm precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? *Journal of Climate*, 28(17):6938–6959.

Chrysanthou, A., van der Schrier, G., van den Besselaar, E. J. M., Klein Tank, A. M. G., and Brandsma, T. (2014). The effects of urbanization on the rise of the european temperature since 1960. *Geophysical Research Letters*, 41(21):7716–7722.

de Valk, C. (2020). Standaardmethode voor berekening van een trend. Technical Report TR-389, Royal Netherlands Meteorological Institute.

de Valk, C. and Brandsma, T. (2023). Homogenization of daily temperatures using covariates and statistical learning—the case of parallel measurements. *International Journal of Climatology*, 43(15):7170–7182.

Dijkstra, F., de Vos, R., Ruis, J., and Crok, M. (2022). Reassessment of the homogenization of daily maximum temperatures in the netherlands since 1901. *Theoretical and Applied Climatology*, 147(3):1185–1194.

Fong, E. and Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, 107(2):489–496.

Guijarro, J. A. (2014). Quality control and homogenization of climatological series. In *Handbook of Engineering Hydrology*, pages 517–530. CRC Press.

Kunsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The annals of Statistics*, 17(3):1217–1241.

Pielke Sr, R. A., Davey, C. A., Niyogi, D., Fall, S., Steinweg-Woods, J., Hubbard, K., Lin, X., Cai, M., Lim, Y.-K., Li, H., et al. (2007). Unresolved issues with the assessment of multidecadal global land surface temperature trends. *Journal of Geophysical Research: Atmospheres*, 112(D24).

Scherrer, S. C., de Valk, C., Begert, M., Gubler, S., Kotlarski, S., and Croci-Maspoli, M. (2024). Estimating trends and the current climate mean in a changing climate. *Climate Services*, 33:100428.

Schumacher, D. L., Singh, J., Hauser, M., Fischer, E. M., Wild, M., and Seneviratne, S. I. (2024). Exacerbated summer european warming not captured by climate models neglecting long-term aerosol changes. *Communications Earth & Environment*, 5(1):182.

Van der Hoeven, P. (1982). Observations of water temperature and salinity, state office of fishery research (rivo): 1880–1981. Technical Report WR-82-8, Royal Netherlands Meteorological Institute.

Vautard, R., Cattiaux, J., Happé, T., Singh, J., Bonnet, R., Cassou, C., Coumou, D., D'andrea, F., Faranda, D., Fischer, E., et al. (2023). Heat extremes in western europe increasing faster than simulated due to atmospheric circulation trends. *Nature Communications*, 14(1):6803.

Venema, V. K. C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., Rasol, D., Rustemeier, E., Kolokythas, K., Marinova, T., Andresen, L., Acquaotta, F., Fratianni, S., Cheval, S., Klancar, M., Brunetti, M., Gruber, C., Prohom Duran, M., Likso, T., Esteban, P., and Brandsma, T. (2012). Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8(1):89–115.

Williams, C. N., Menne, M. J., and Thorne, P. W. (2012). Benchmarking the performance of pairwise homogenization of surface temperatures in the united states. *Journal of Geophysical Research: Atmospheres*, 117(D05116).

WMO (2014). Guide to meteorological instruments and methods of observation. Technical Report WMO-No. 8, World Meteorological Organization.

WMO (2020). Guidelines on homogenization. Technical Report WMO-No. 1245, World Meteorological Organization.

Wood, S. N. (2020). Inference and computation with generalized additive models and their extensions. *Test*, 29(2):307–339.

World Meteorological Organization (2004). Guidelines on climate metadata and data homogenization. Wcdmp no. 53, wmo-td no. 1186, World Meteorological Organization, Geneva, Switzerland.

World Meteorological Organization (2017). Manual on climate data management. Wmo-no. 1131, World Meteorological Organization, Geneva, Switzerland.

Part II

Appendix

Appendix A

List of climate indices

Index/Indices	Meaning
TGg/TGn/TGx	Annual mean/minimum/maximum of daily mean temperature TG
TXg/TXx	Annual mean/maximum of daily maximum temperature TX
TX3x	Annual maximum of running 3-day mean of daily maximum temperature TX
TNg/TNn	Annual mean/minimum of daily minimum temperature TN
TN3n	Annual minimum of running 3-day mean of daily minimum temperature TN
Nd_summer	Annual number of summer days: days with TX greater than or equal to 25.0 C
Nd_tropical	Annual number of tropical days: days with TX greater than or equal to 30.0 C
N_heatwave	Annual number of heatwaves (see caption)
Nd_ice	Annual number of ice days: days with TX below 0.0 C
Nd_severefrost	Annual number of days with severe frost: days with TN below -10.0 C
N_coldwave	Annual number of cold waves (see caption)

Table A.1: A heatwave is defined as a period of at least 5 consecutive summer days in which at least three tropical days occur. A cold wave is defined as a period of at least 5 consecutive ice days in which at least three days with severe frost occur.
Strictly, these definitions apply only to the temperature at station De Bilt. However, in the present study, we have also applied these definitions to the data of the other (H4) stations in order to compare and evaluate the results of different homogenization methods.

Appendix B

H4 station temperatures homogenized with two different methods: annual means and extremes

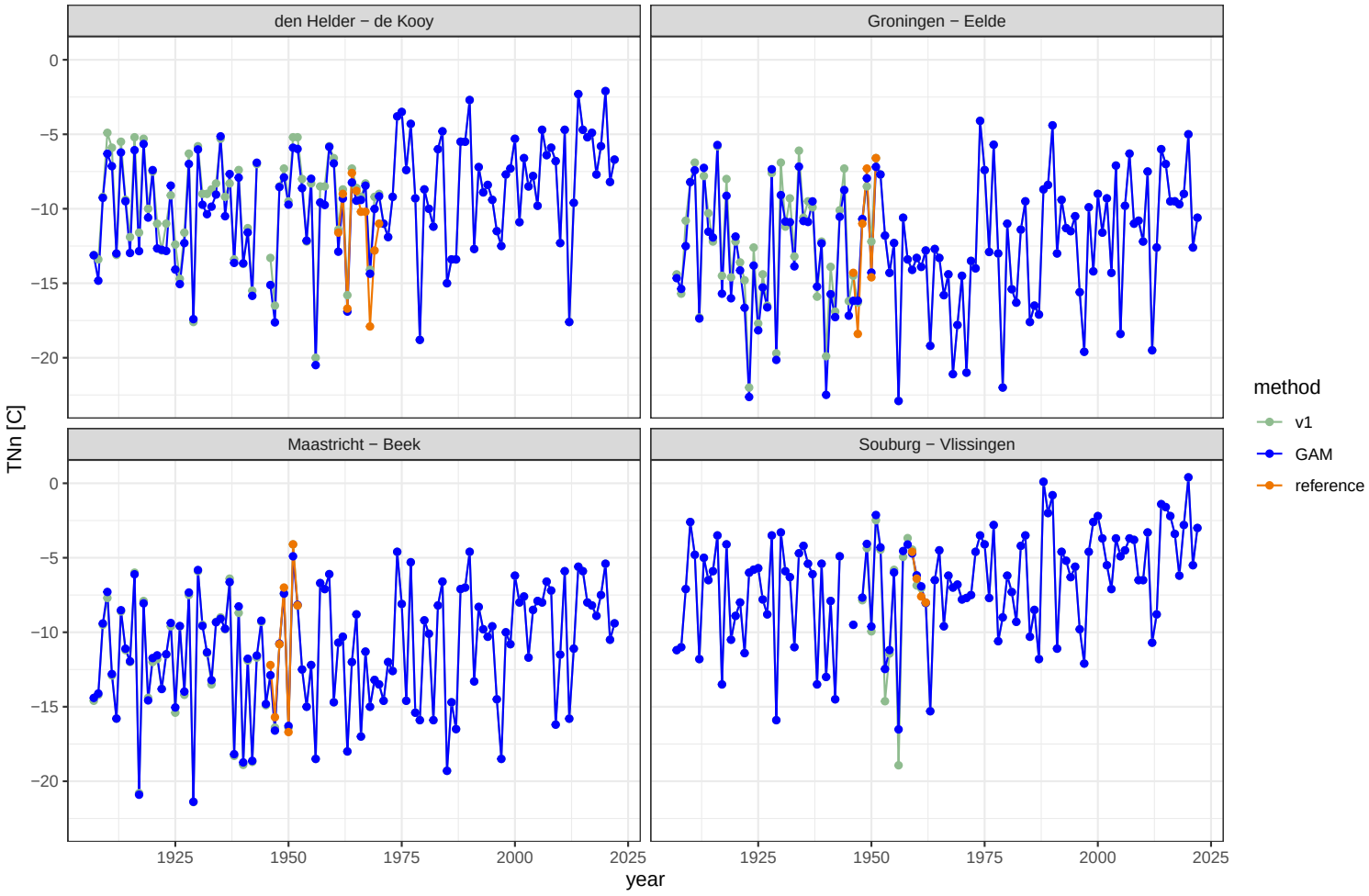

Figure B.1: Annual minima of daily minimum temperature TNn homogenized by version 1.0 (v1) and by version 2.0 (GAM), and from measurements at the new site used for model fitting. Also showing the homogenized values over the calibration period (not used in the final product).

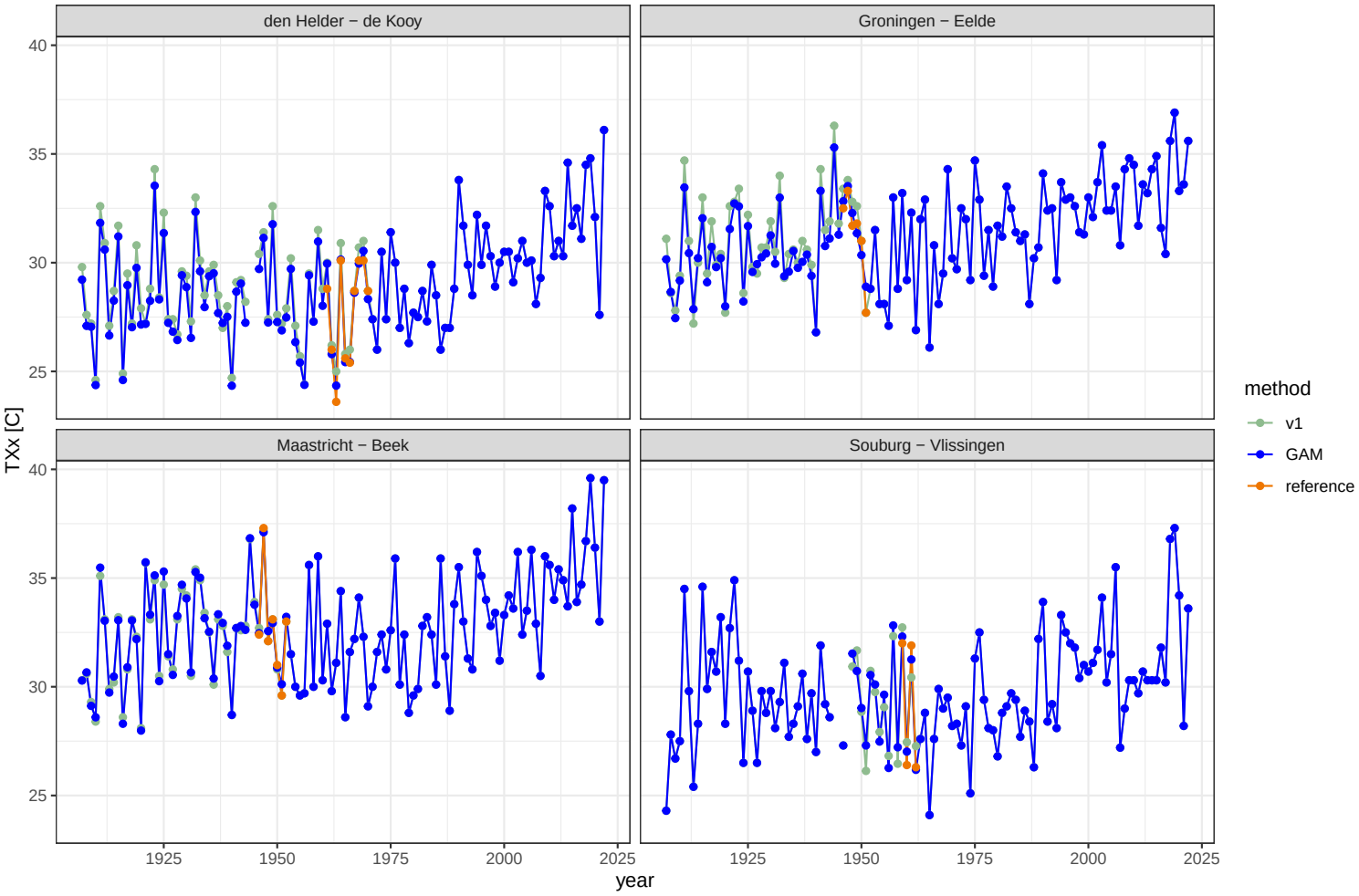

Figure B.2: As Figure B.1 but for annual maxima of daily maximum temperature TXx.

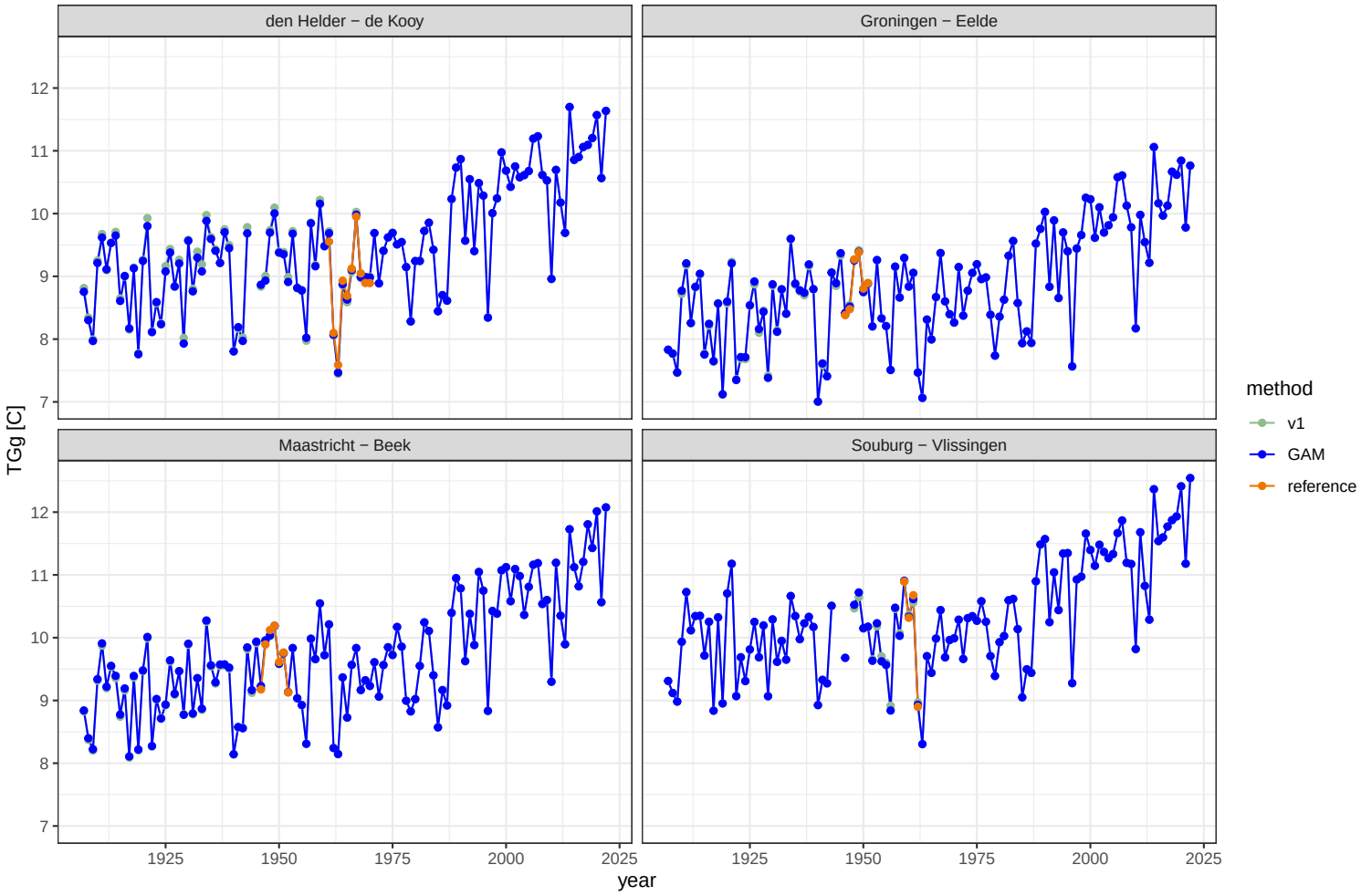

Figure B.3: As Figure B.1 but for annual means of daily mean temperature TGg.

Appendix C

H4 station temperatures homogenized with two different methods: average quantiles

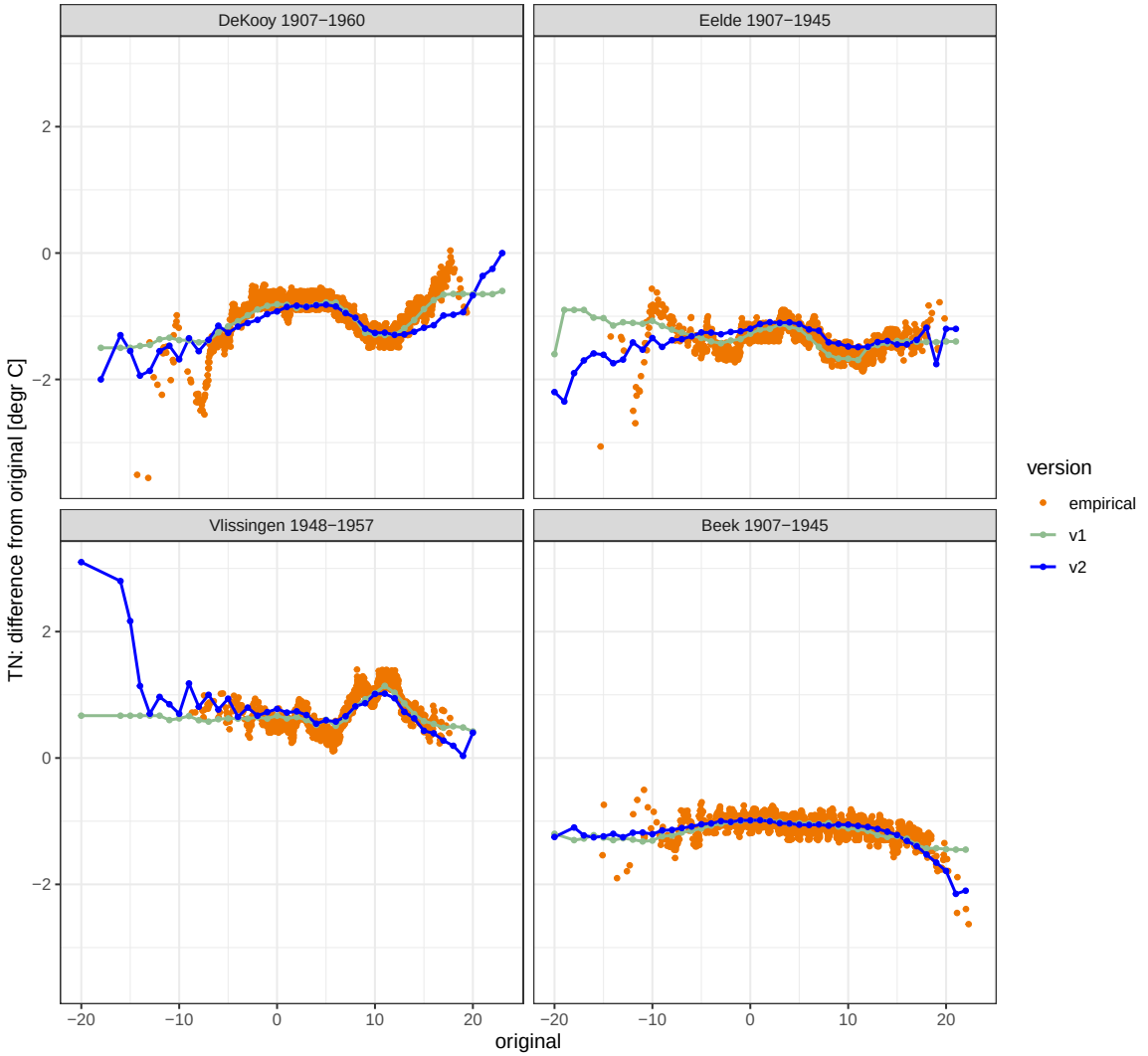

Figure C.1: Green/blue: difference between a quantile of the homogenized TN at the new site and the corresponding quantile of the measured TN at the old site, as function of the latter quantile (for the years indicated in the headers). Colour indicates version. Orange dots: difference between the empirical quantiles of TN at the new and the old site, derived directly from the parallel measurements.

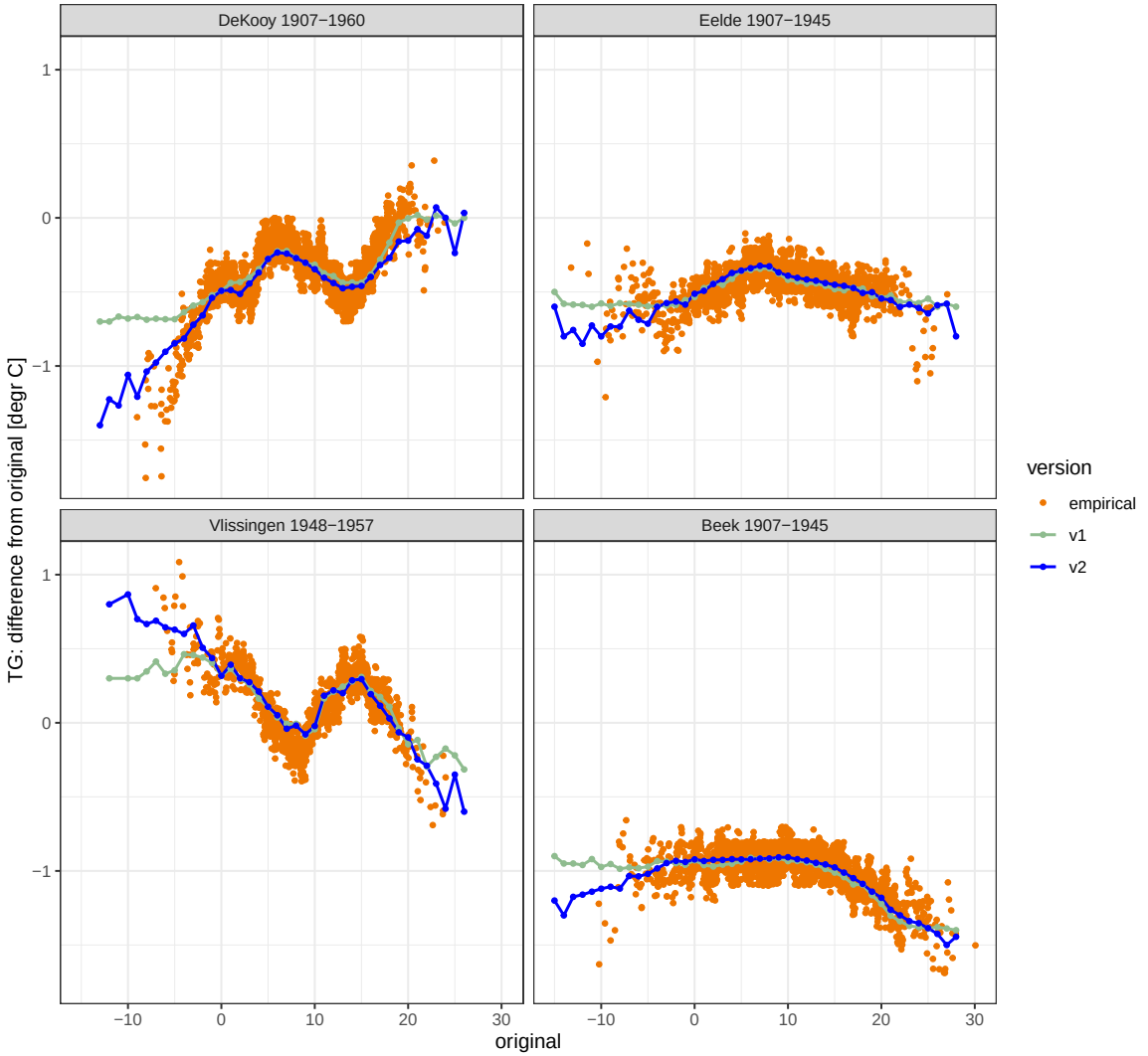

Figure C.2: As Fig. C.1 but for TG.

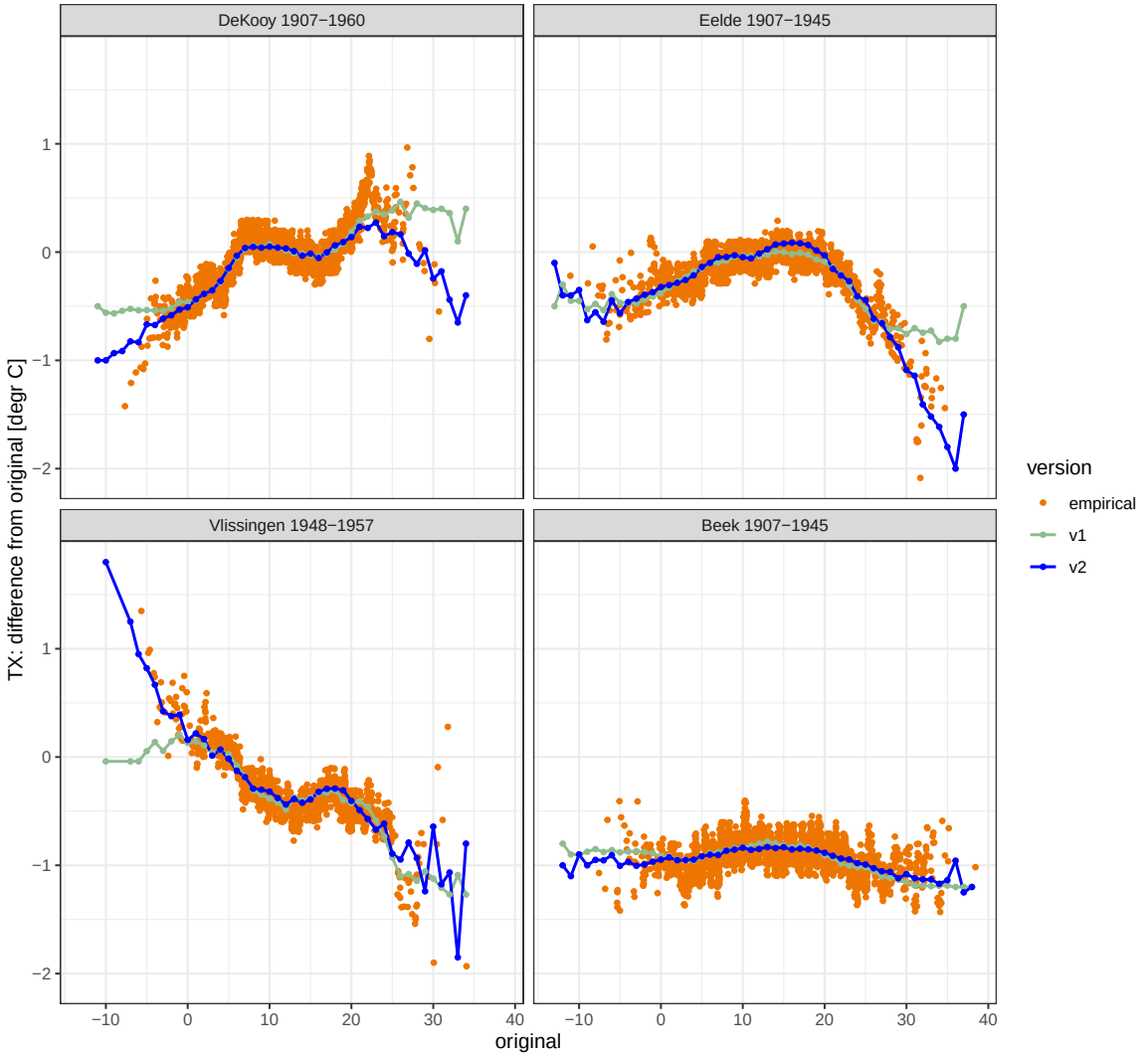

Figure C.3: As Fig. C.1 but for TX.

Appendix D

Aggregated indices and their sampling uncertainty for the H4 stations

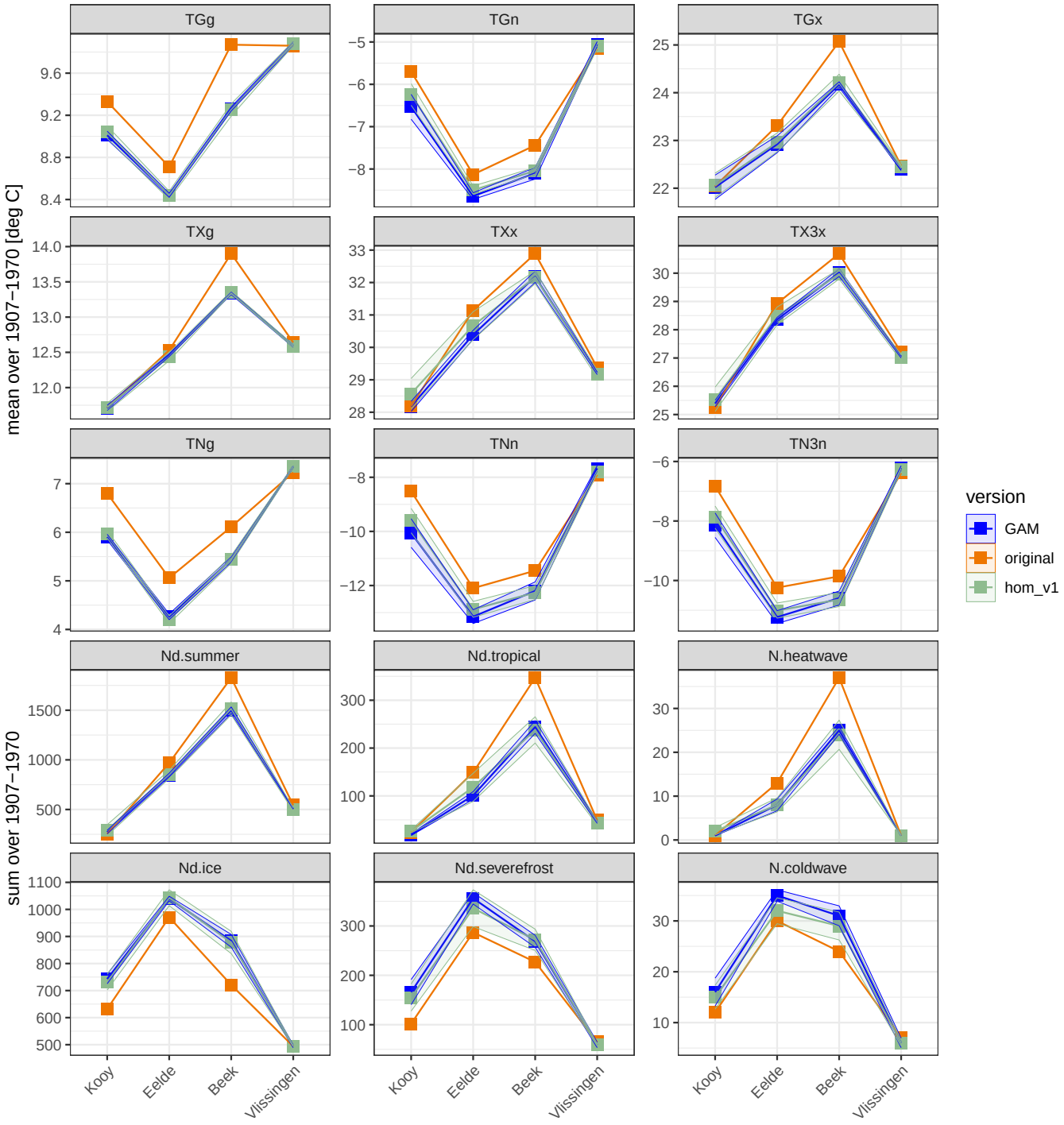

Figure D.1: Climate indices (see Appendix A) for the H4 stations aggregated over 1907-1970 by averaging (upper 9 panels) and summation (lower 6 panels) from original data and data homogenized by version 1.0 (`hom_v1`) and version 2.0 (GAM), with for the latter also indicative 95% confidence intervals of the sampling uncertainty due to homogenization.

Appendix E

Trends of indices for the H4 stations

The figures in this Appendix show long-term trends of the climate indices defined in Appendix A for the H4 stations.

They can be used to compare long-term trends from different versions of the dataset. The confidence intervals shown here only reflect the sampling uncertainty in the trend lines due to year-to-year variability, but do NOT include sampling uncertainty introduced by the calibration of the homogenization.

The latter is shown for version 2.0 in the figures in Appendix F, together with the total sampling uncertainty (both due year-to-year variability and to calibration).

The plots show that for version 2.0, the calibration does not increase the uncertainty much, so the easier to compute confidence intervals in this Appendix are sufficient in practice.

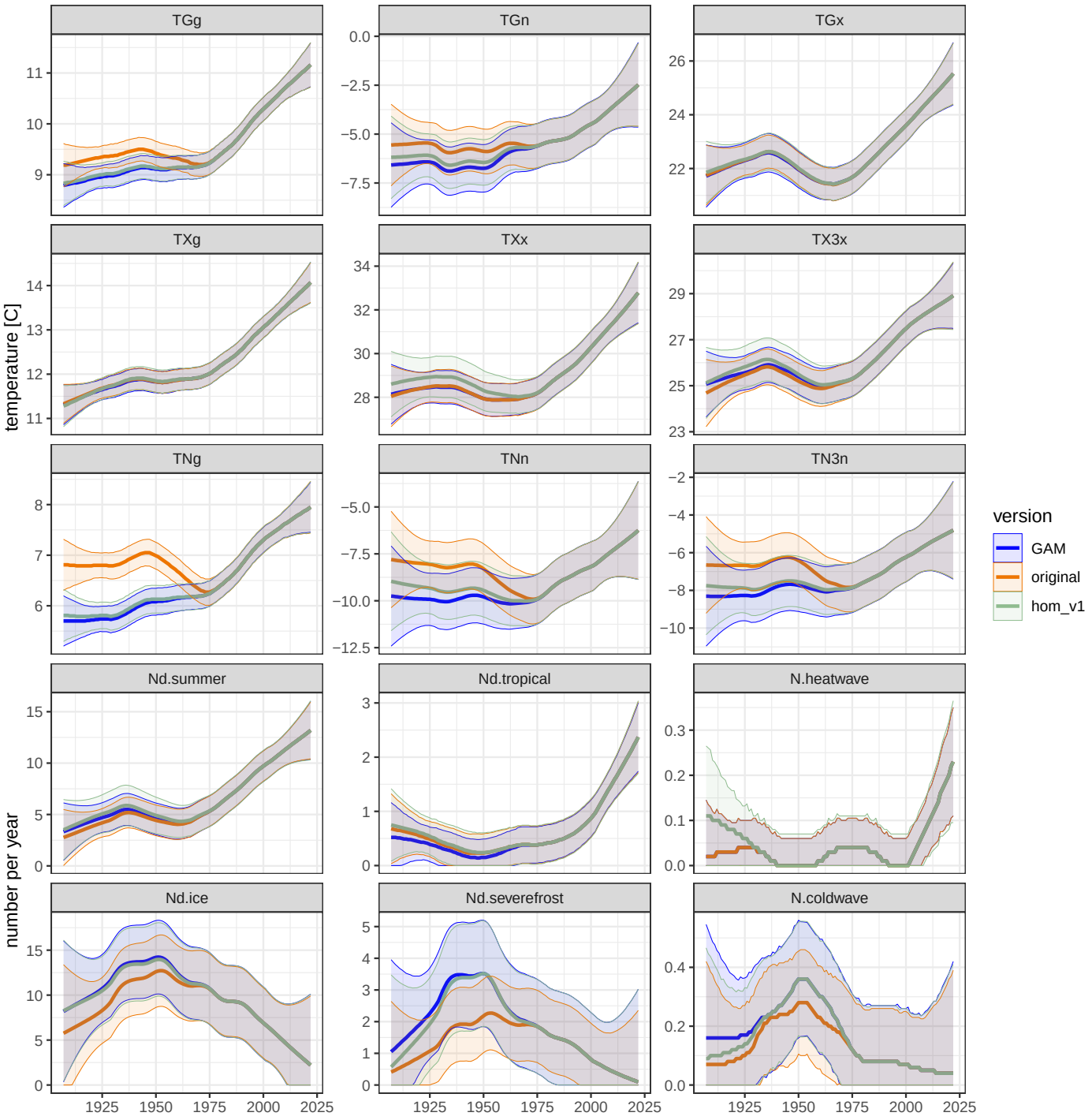

Figure E.1: Trend lines of indices (see Appendix A) for **De Kooy** from original data and data homogenized by version 1.0 (`hom_v1`) and GAM, with indicative 95% confidence intervals of the sampling uncertainty due to year-to-year variability (NOT including the sampling uncertainty due to calibration of the homogenization).

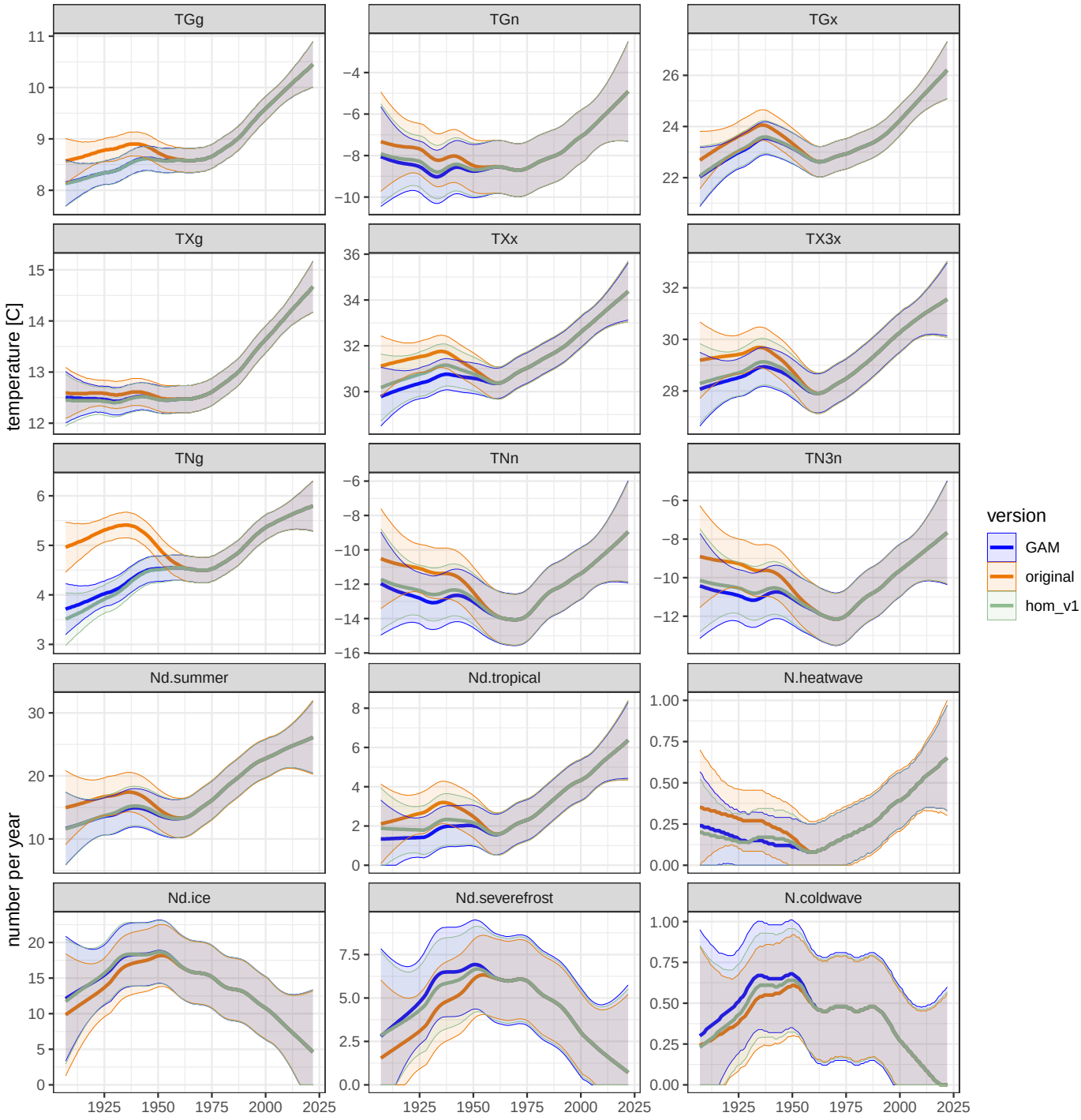

Figure E.2: As Figure E.1 but for **Eelde**.

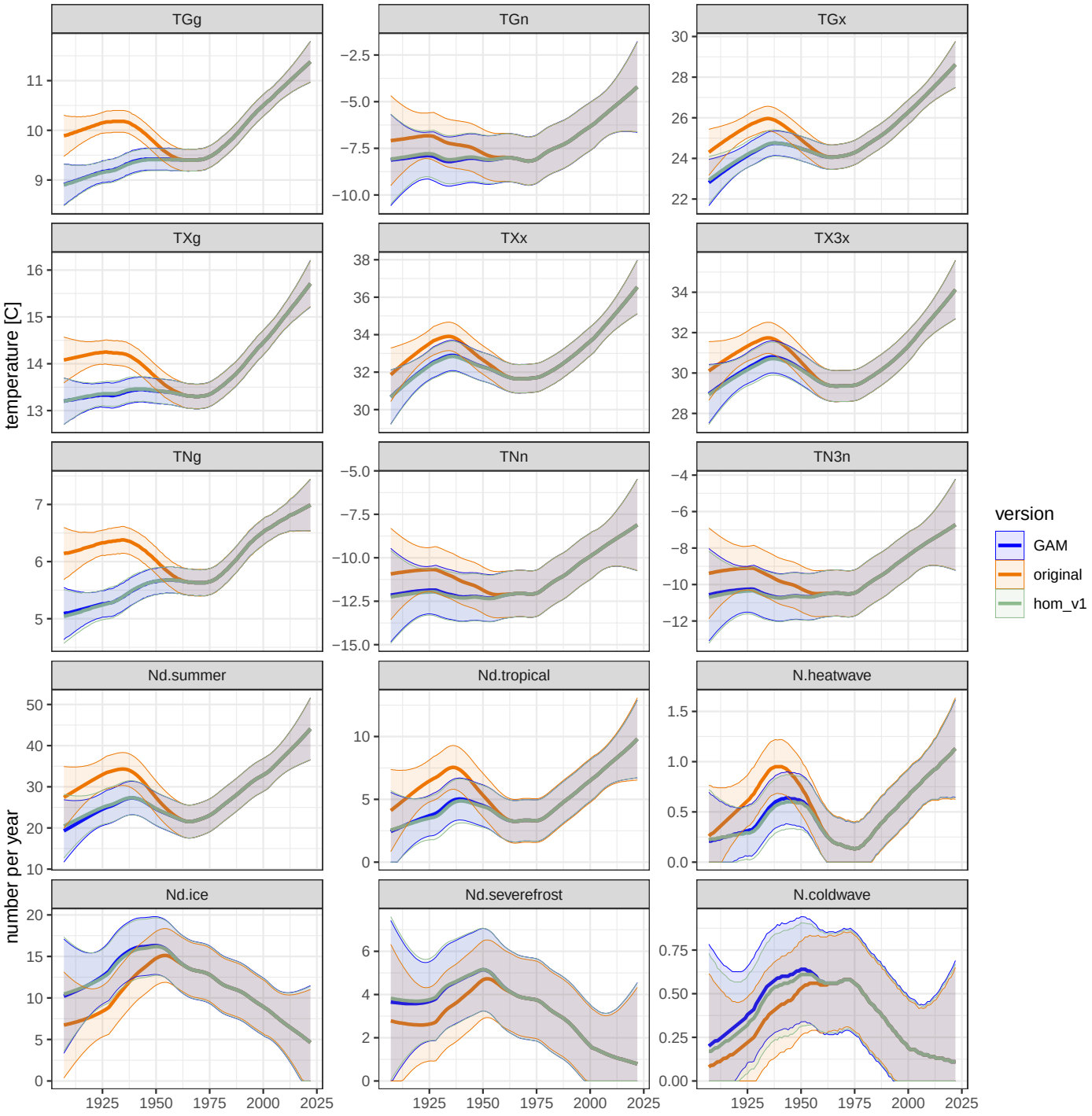

Figure E.3: As Figure E.1 but for **Beek**.

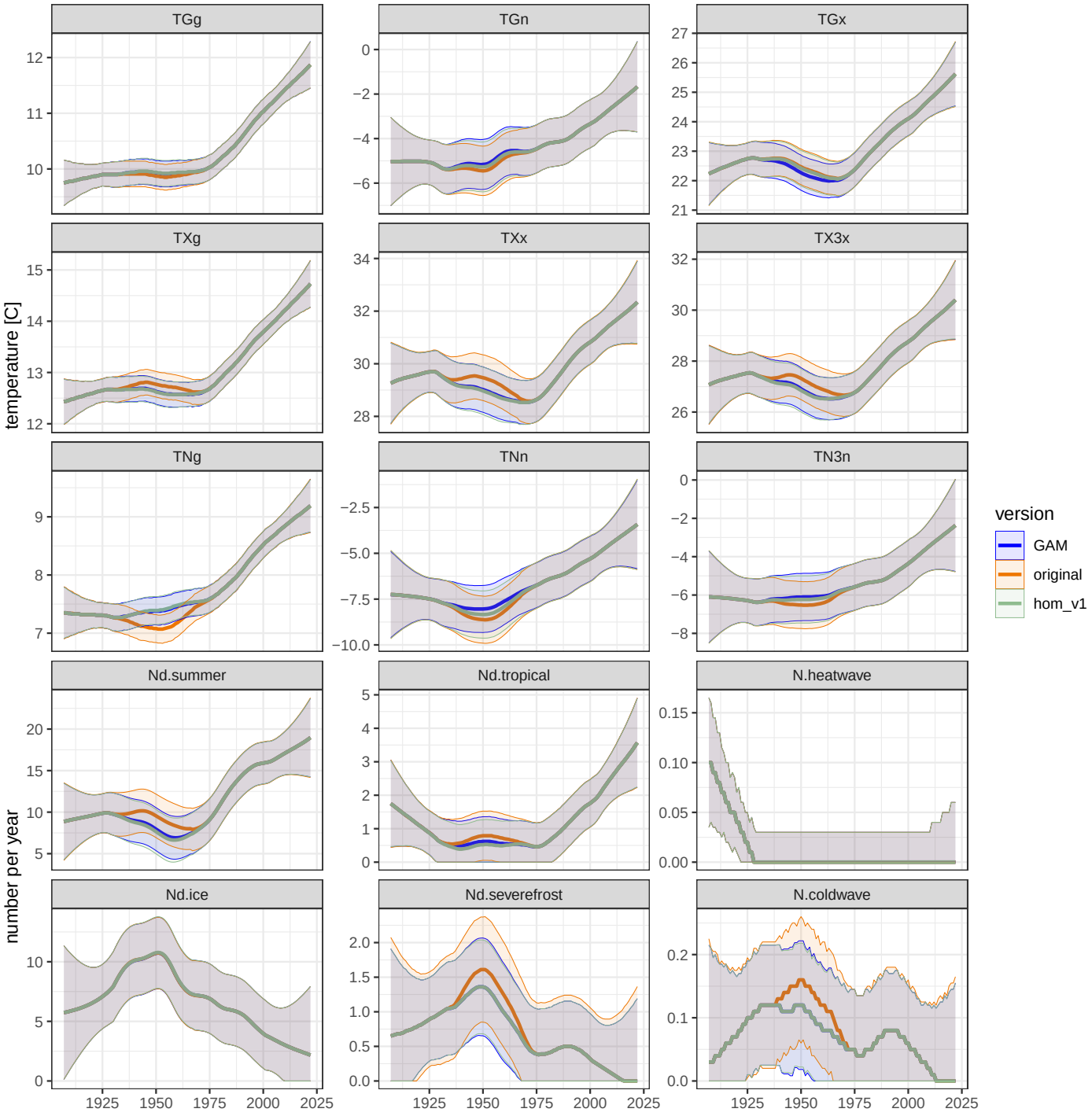

Figure E.4: As Figure E.1 but for **Vlissingen**.

Appendix F

Sampling uncertainty of trends for the H4 stations due to homogenization by GAM

The figures in this Appendix show long-term trends of the climate indices defined in Appendix A for the H4 stations, derived from the version 2.0 data and from the unadjusted data.

The confidence intervals for version 2.0 represent the sampling uncertainty introduced by the calibration of the homogenization (dark blue), and the total sampling uncertainty (light blue) which also includes the effect of year-to-year variability.

These plots and comparison with Appendix E show that for version 2.0, the calibration does not increase the uncertainty much, so the easier to compute confidence intervals in Appendix E are sufficient in practice.

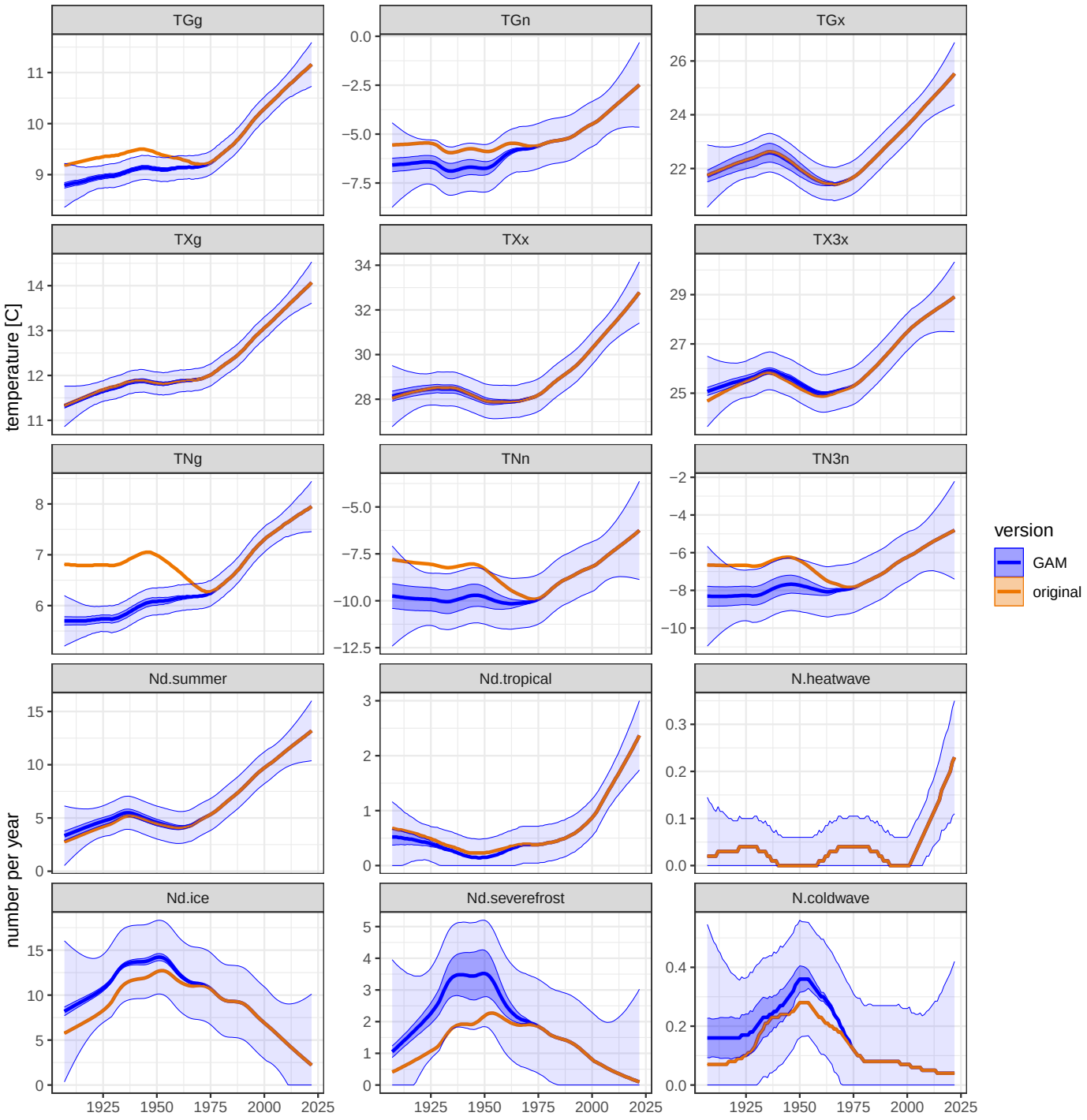

Figure F.1: Trend lines of indices (see Appendix A) for **De Kooy** from original data and data homogenized by version 2.0 (GAM), with for the latter indicative 95% confidence intervals of the total sampling uncertainty of the trend estimate (light blue) and of the component of this uncertainty due to calibration of the homogenization (dark blue).

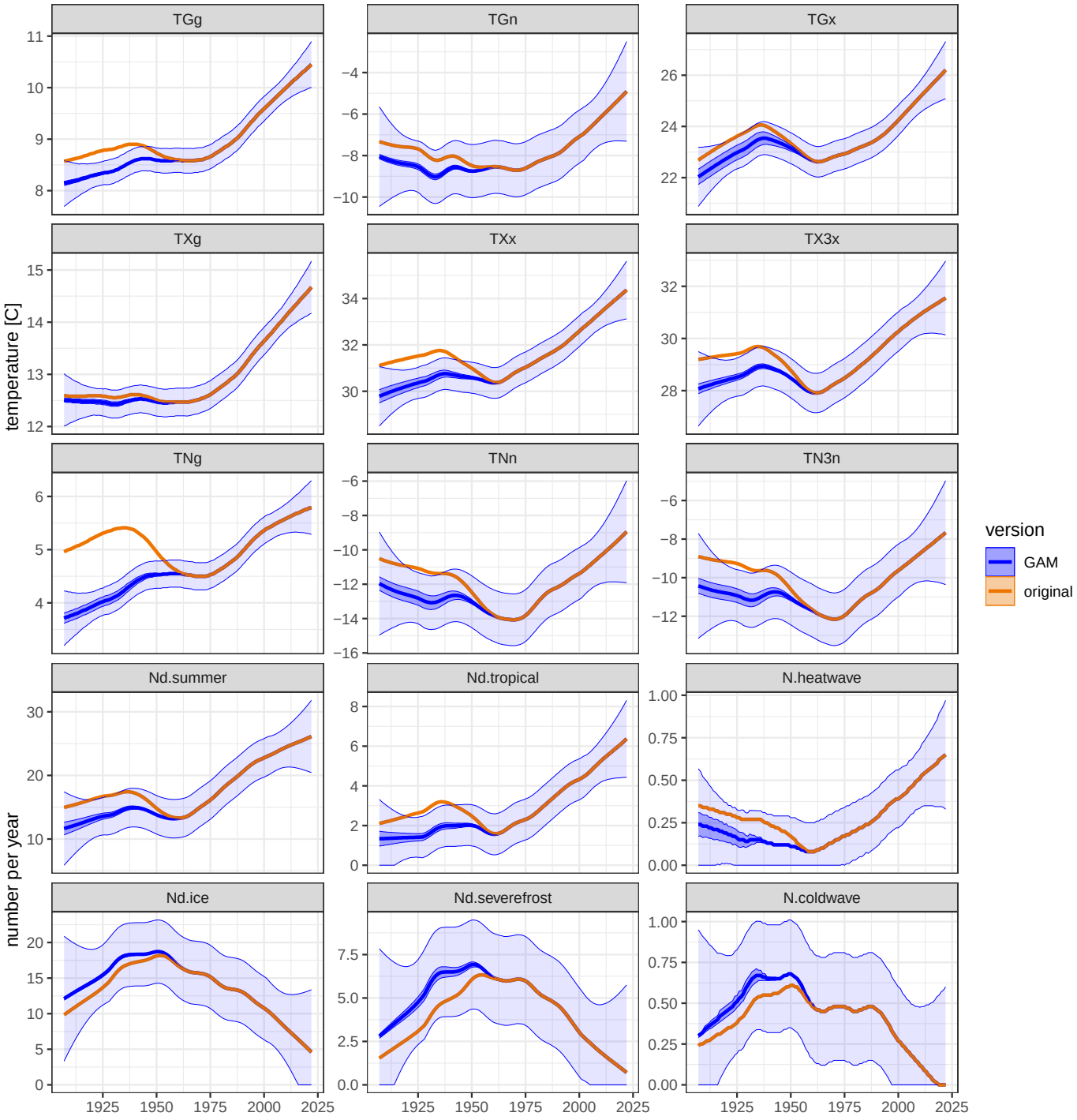

Figure F.2: As Figure F.1 but for **Eelde**.

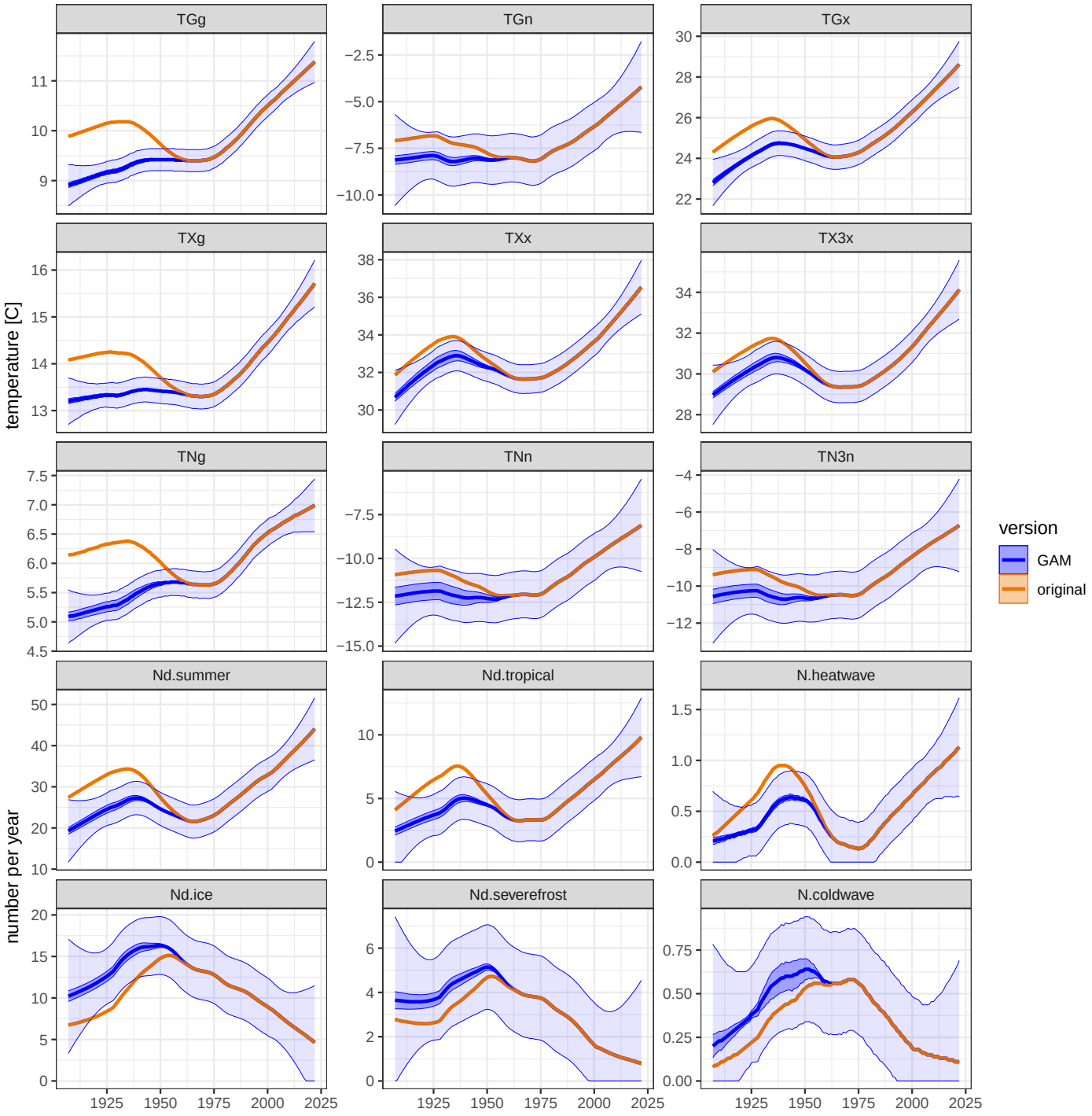

Figure F.3: As Figure F.1 but for **Beek**.

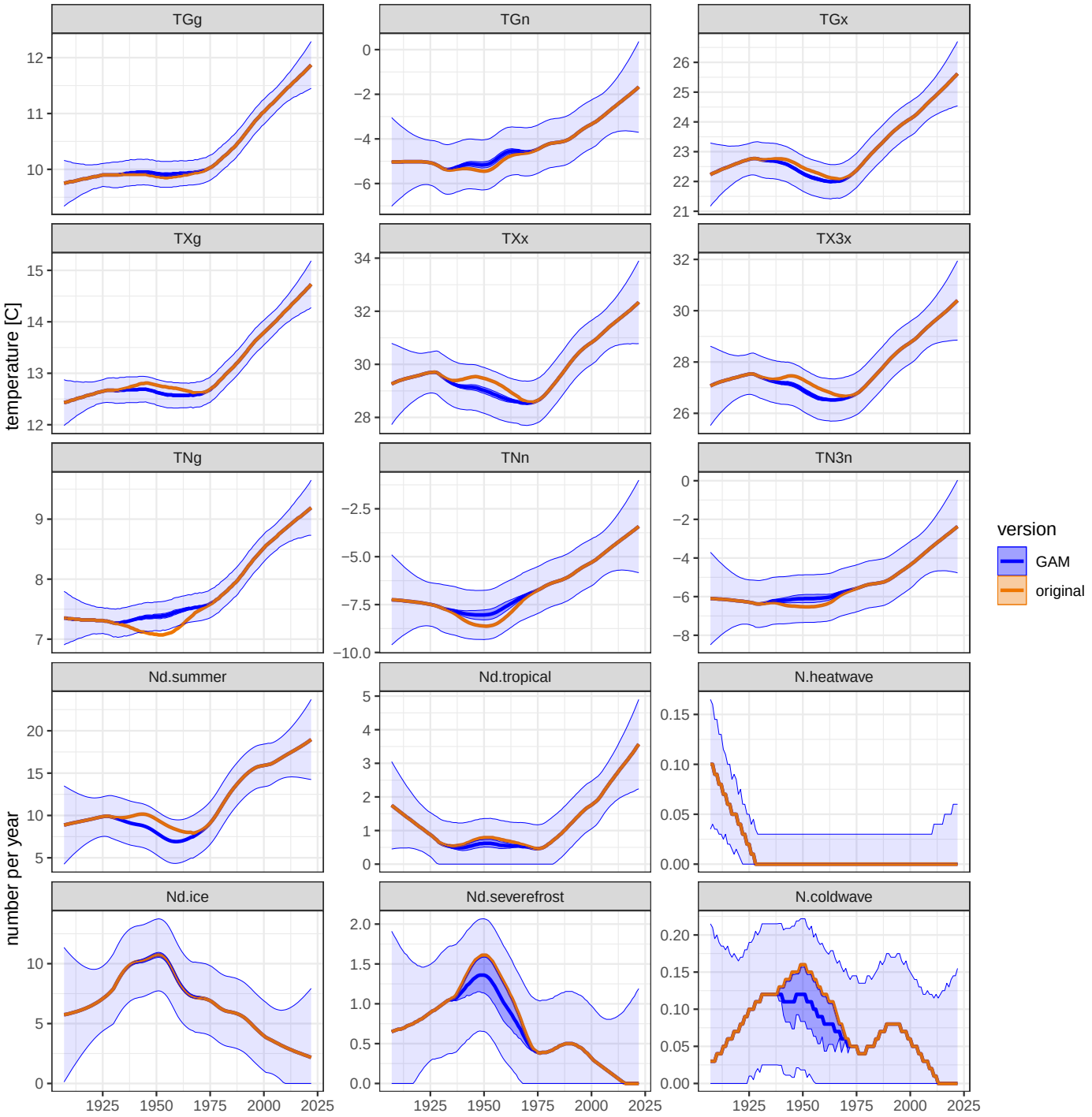

Figure F.4: As Figure F.1 but for **Vlissingen**.

Appendix G

Comparison of methods for homogenization of the temperature record of De Bilt

Due to the absence of parallel measurements at De Bilt around the known changes in sensor type and location, temperature adjustments must be estimated from data collected at reference stations, for which we use a subset of the H4 stations.

As described in Section 3.3, the temperature adjustments are derived by subtracting predictions of the temperature at the Bilt by models calibrated on data collected before and after the known breakpoints.

We evaluated two modeling approaches to predict temperatures at De Bilt using reference station data:

- the QDM (Quantile Delta Mapping) model similar to the model used in version 1.0 (see Section 3.2),
- the GAM model from de Valk and Brandsma (2023) which accounts for effects of covariates like wind, humidity, cloud cover and season (see Section 3.3), applied in version 2.0 for the H4 stations with parallel measurements.

  Several versions of this model were tested. Here, we report on a version in which the relationship between the temperatures at De Bilt and the reference station is constrained to be linear in order to enhance robustness against noise (see the first limitation in Section 3.4).

Because suitable parallel measurements around the breakpoints are missing for De Bilt, it is not possible to test the homogenization of De Bilt using data from De Bilt only. Therefore, we performed tests in which the data of De Bilt before the known breakpoints were replaced by data from the nearby station Soesterberg. The idea behind this is that a method for homogenization of the data from two different instruments/locations at De Bilt using data from (a) reference station(s) far away should also show skill in adjusting

the data from Soesterberg to match those of De Bilt. One might object that this setup is not representative because the correlation between temperatures at Soesterberg and de Bilt is much lower than between temperatures at the two sites in De Bilt, However, since there are no suitable parallel measurements at De Bilt, this high correlation cannot benefit the homogenization in any way; hence, the test involving Soesterberg is in fact representative and valuable for predicting the performance of homogenization using data from H4 station(s) as reference.

Only data from 1961-2000 were used for testing; there are no known sensor relocations or replacements within this period. Breakpoints were assigned within this period to cover the entire dataset. Test results for the various breakpoints were aggregated to assess overall performance.

Tests were carried out with different reference stations and calibration time windows (used before and after the breakpoint). In Tab. G.1, the RMS errors for TN and TX are shown for two window lengths (5 and 10 years) with Beek as reference station.

	TN		TX	
window	QDM	GAM	QDM	GAM
5 yr	0.82	0.85	0.67	0.82
10 yr	0.81	0.87	0.63	0.74

Table G.1: RMS errors (°C) of predictions of daily minimum temperature TN at De Bilt from TN at Soesterberg, using TN from Beek as reference, and similarly for TX instead of TN, for two methods (QDM and GAM/covariates) and two lengths of the calibration time window before and after the assumed breakpoint.

Tab. G.1 shows that the GAM model does not perform better than QDM. Similar tests were performed with Eelde as reference station, and considering other metrics besides RMS error. All these results are in line with the ones reported here.

This outcome contrasts with the favorable outcomes of the GAM model for homogenizing the H4 stations based on on parallel measurements (see de Valk and Brandsma (2023)). To understand this, we need to realize that homogenization with and without parallel measurements are two quite different problems.

Including the effects of non-temperature covariates in the GAM models for the H4 stations with parallel measurements works because the correlation between the temperatures and the old and the new site is already fairly high: roughly the same weather passes the two sites, so differences in temperature are for a substantial part attributable to differences between the local environments. The effects of the latter depend on the weather situation (wind, cloud cover, etc.) which is modelled by the GAMs.

This does not hold for temperature differences between De Bilt and any of the H4 stations: because of the much larger spatial distances, the weather passing these sites may differ considerably, resulting in lower correlations (higher noise). The potential reduction in the error of the prediction of the temperature at De Bilt from the data of the reference station by using non-temperature covariates is therefore much smaller,

not only in a relative sense, but for the RMS error also in an absolute sense. Also, the increased noise level will result in a less well calibrated model with larger scale bias (a well-known phenomenon in least squares regression), requiring a larger inflation of the predicted spread in temperatures to match the observed spread.

This problem is exacerbated by the subtraction of the predictions by the model fitted to measurements before the breakpoints from the predictions by the model fitted to measurements after the breakpoints. This further amplifies the noise.

Furthermore, this procedure does not directly optimize the homogenization itself. Instead, it combines two models of limited quality (see above) optimized to predict the temperature at De Bilt from a temperature at (a) station(s) far away, but not optimized for the intended purpose.

This explains that without parallel measurements or data from a reference station nearby, very simple models tend to perform better than more sophisticated models. In fact, a simple linear model for predicting the temperature at the Bilt from only the temperature at the reference site (the simplest "GAM") performs similar to QDM in the test using data of Soesterberg and De Bilt.

Our conclusion is that the use of non-temperature covariates does not improve the prediction of temperatures at De Bilt using data from H4 stations as reference. Therefore, version 2.0 of the De Bilt homogenization is based on a refined QDM method.

Appendix H

Aggregated indices and their sampling uncertainty for De Bilt

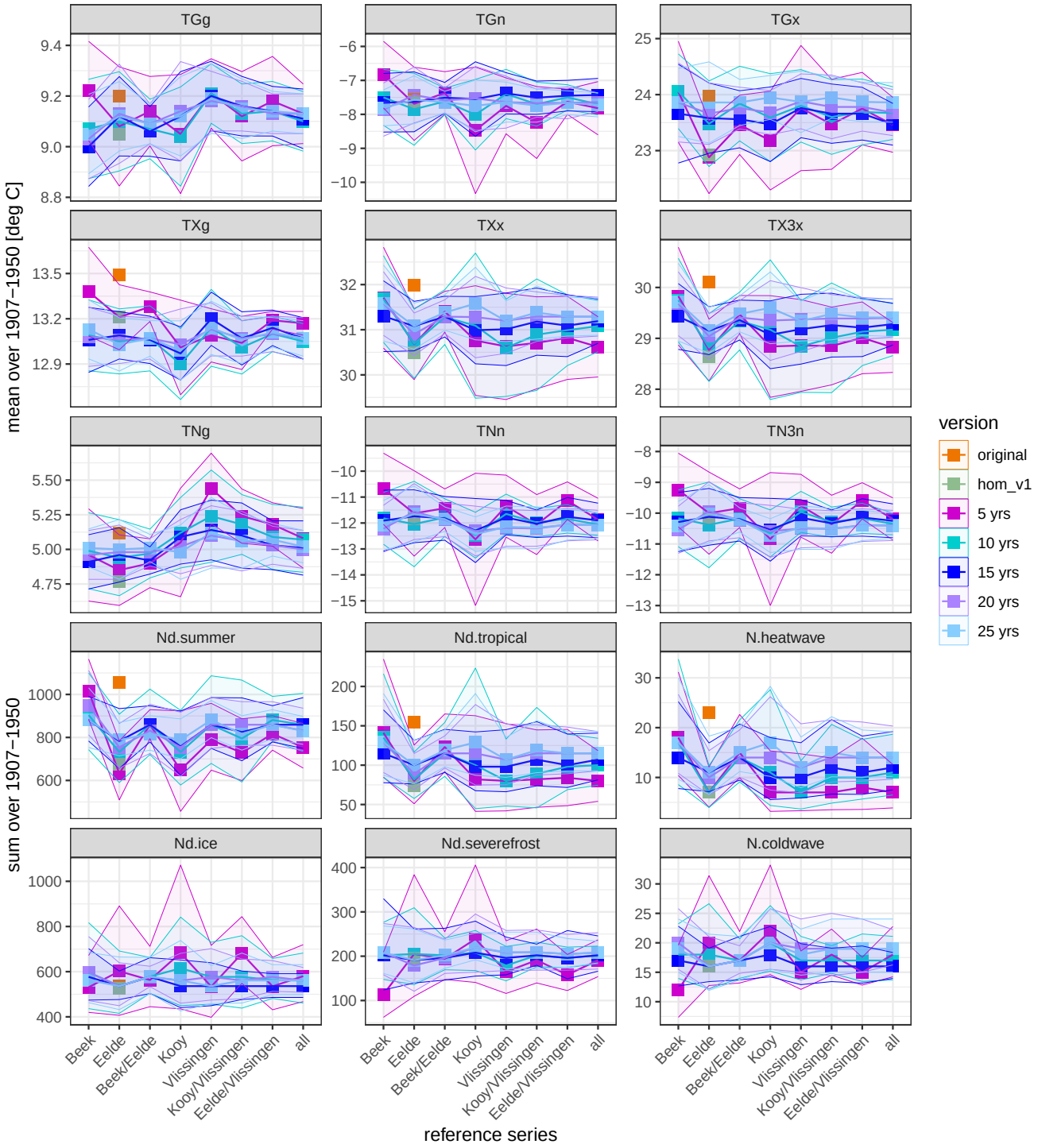

Figure H.1: Climate indices (see Appendix A) for **De Bilt** aggregated over 1901-1950 by averaging (upper 9) and summation (lower 6) from original data, data homogenized by version 1.0 (`hom_v1`) and data homogenized by QDM with different window lengths, with for the latter also indicative 95% confidence intervals of the sampling uncertainty due to homogenization.

Appendix I

Trends of indices for station De Bilt

The figures in this Appendix show long-term trends of the climate indices defined in Appendix A for De Bilt.

Fig. I.1 can be used to compare long-term trends from different versions of the dataset. The confidence intervals shown here only reflect the sampling uncertainty in the trend lines due to year-to-year variability, but do NOT include uncertainty introduced by the calibration of the homogenization.

For version 2.0, the latter is shown in the next Fig. I.2, together with the total uncertainty (including year-to-year variability and calibration). Comparison of the two figures shows that the calibration does not increase the uncertainty much, so the easier to compute confidence intervals in Fig. I.1 are sufficient in practice.

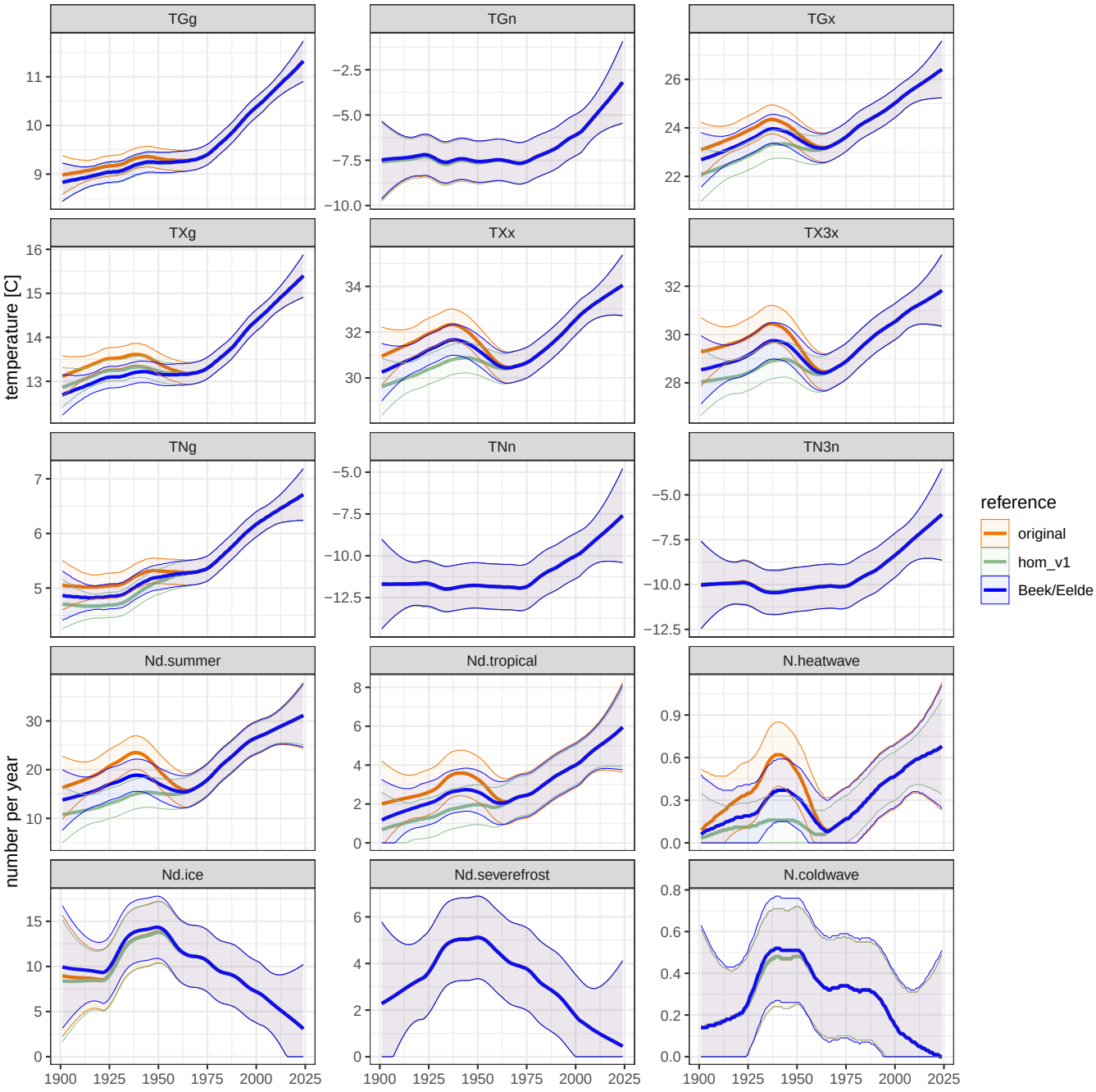

Figure I.1: Trend lines of indices (see Appendix A) for **De Bilt** from original data and data homogenized by version 1.0 (`hom_v1`) and version 2.0 (Beek/Eelde), with indicative 95% confidence intervals of the sampling uncertainty of the trend estimate due to year-to-year variability only.

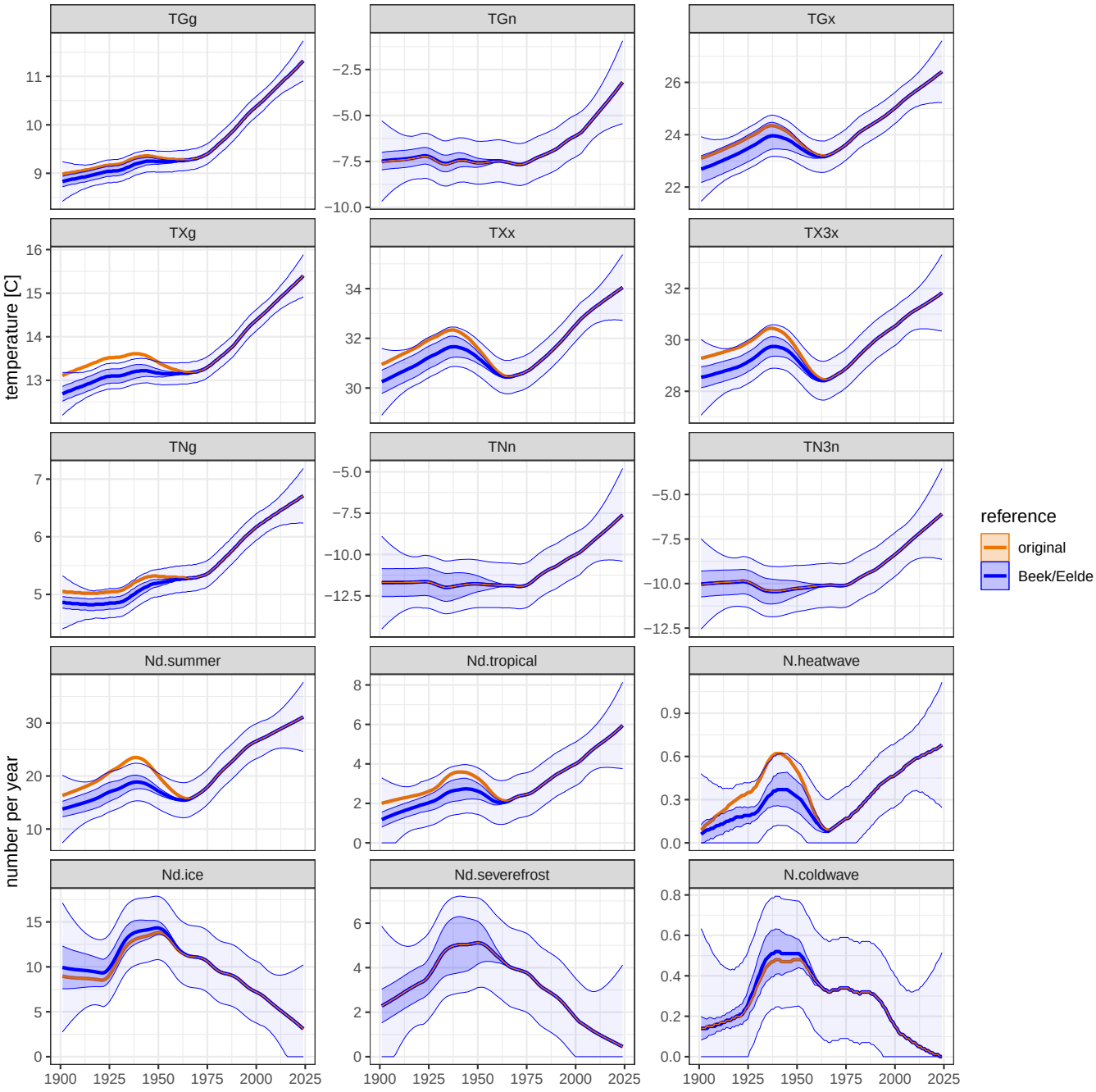

Figure I.2: Trend lines of indices (see Appendix A) for **De Bilt** from original data and version 2.0 (Beek/Eelde), with indicative 95% confidence intervals of the total sampling uncertainty of the trend estimate (light blue) and of the component of this error due to homogenization (dark blue).

Appendix J

Homogenized annual values and their trends

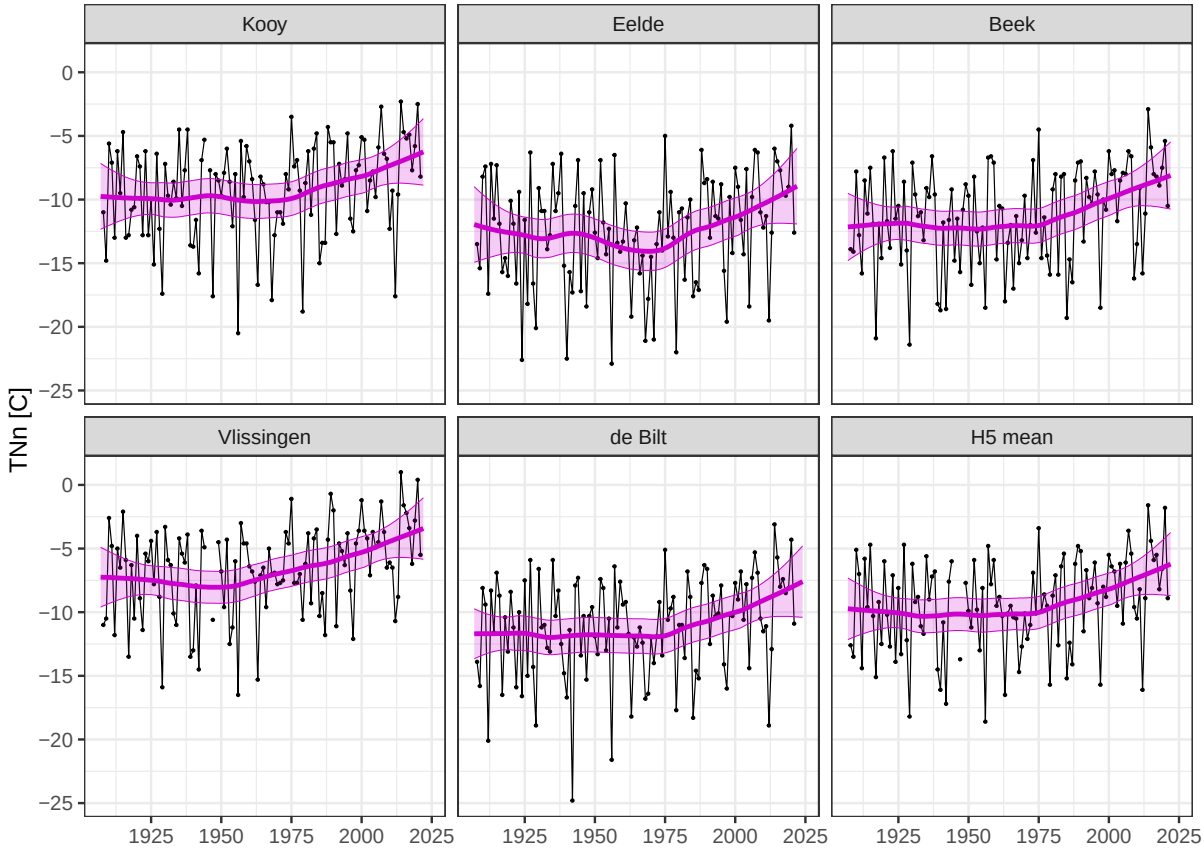

Figure J.1: Annual minimum temperature TNn and its nonlinear trend, with 95% confidence intervals for the trend (only accounting for year-to-year fluctuations) from five stations and from the mean of daily minimum temperature TN from these stations.

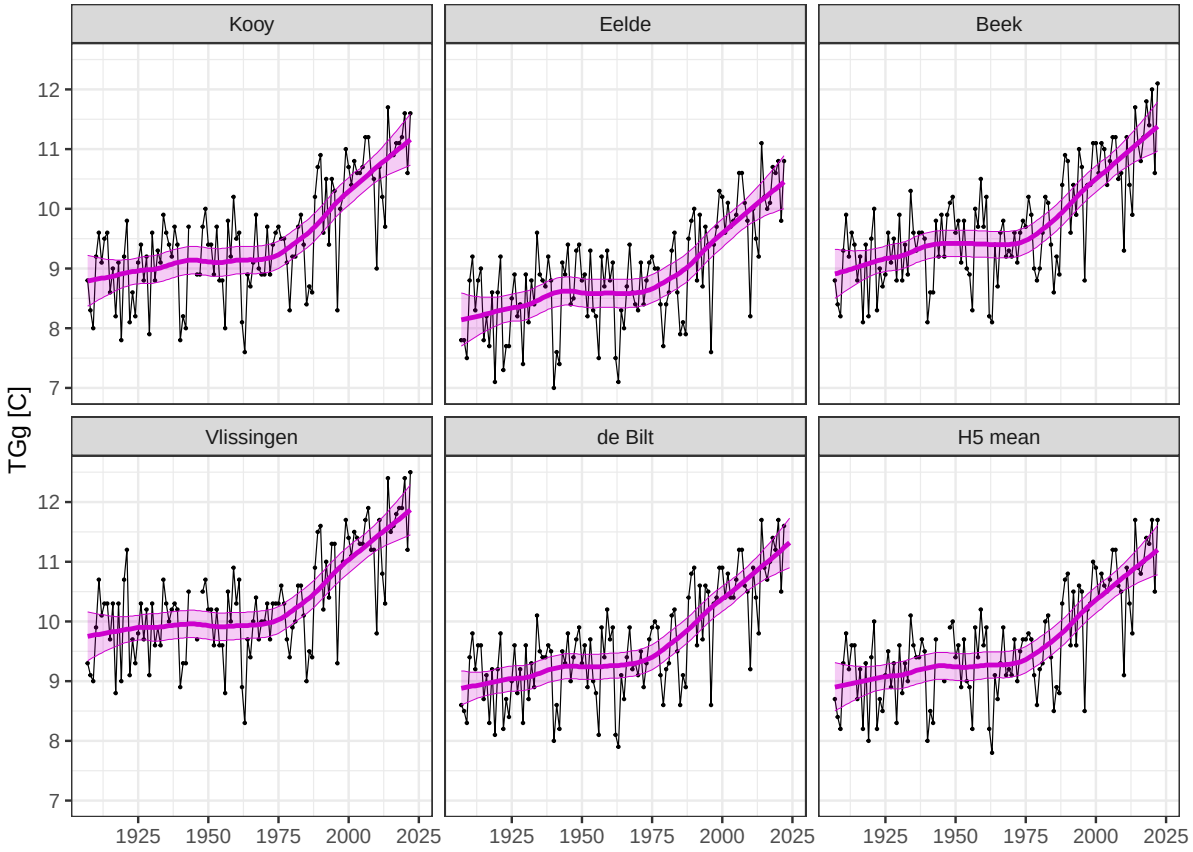

Figure J.2: Annual mean temperature TGg and its nonlinear trend, with 95% confidence intervals for the trend (only accounting for year-to-year fluctuations) from five stations and from the mean of daily mean temperature TG from these stations.

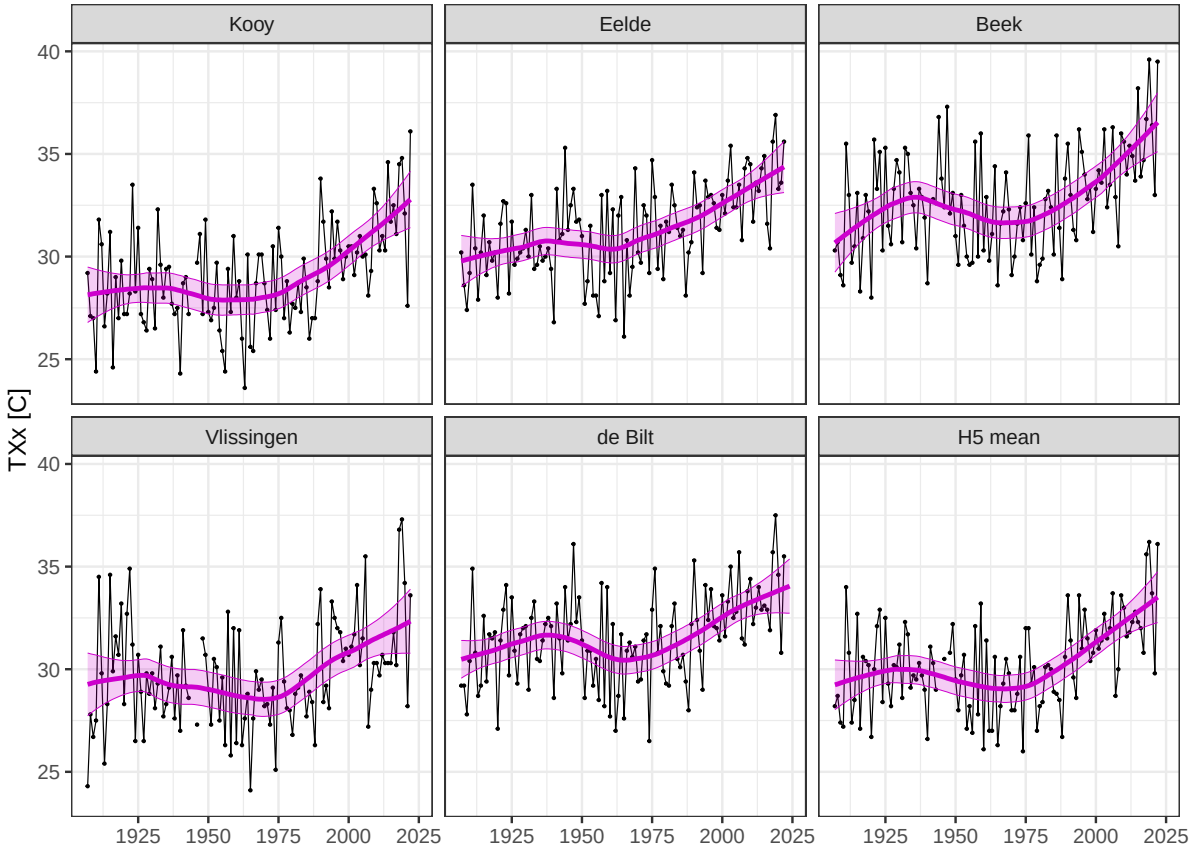

Figure J.3: Annual maximum temperature TXx and its nonlinear trend, with 95% confidence intervals for the trend (only accounting for year-to-year fluctuations) from five stations and from the mean of daily maximum temperature TX from these stations.

Appendix K

Additional checks of the homogenization version 2.0

K.1 Check using an automated homogenization method

A straightforward way to verify the quality of a homogenized dataset is to apply an independent homogenization method to the already homogenized series and assess whether it identifies any remaining significant inhomogeneities.

For this purpose, we chose the homogenization package Climatol version 4.1.0 (Guijarro, 2014). This package is designed for automatic detection and adjustment of step changes and outliers in monthly averaged temperature time series based on comparisons of series from multiple sites. Therefore, applying Climatol to the version 2.0 time series primarily serves to detect any overlooked but significant changes in instrumentation or environmental conditions. We applied Climatol separately to the version 2.0 homogenized TN, TX and TG time series over the period 1907-2022.

The monthly and annual mean adjustments suggested by Climatol are shown in Figure K.1. The only substantial adjustments identified are short-term spikes. There are only a few (3-4 for each of TN, TX and TG for all stations). We do not adjust these spikes, as they have little influence on long-term trends. The suggested temperature adjustments for detected step changes are generally small: only in two periods of a few years, $0.5°$ is reached. Averaged over the five stations, the suggested annual mean adjustments are negligible: their root mean squared values are $0.04°$ for TN and TX and $0.01°$ for TG.

We reviewed earlier studies and conducted a targeted search in the station metadata to identify possible causes for the most significant inhomogeneities.

Eelde

Groningen/Eelde shows a step changes in TN around 1924 and 2018. Regarding 1924, the metadata notes a relocation of the thermometer screen in February 1928. As this occurred a few years after 1924, it is unclear whether this is the cause of the step change.

The step change around 2018 is probably related to the installation of a solar farm in 2019 about 30 m north of the temperature screen. Brandsma (2025) found an annual average nighttime cooling of about 0.3°C due to the installation of the solar farm. This is of the same order of magnitude as the adjustment proposed in Figure K.1. On 11 November 2025 the measurement site moved to a new location 35 m further from the solar farm. It is expected that this will strongly reduce the impact of the solar farm.

De Bilt

For De Bilt the only noteworthy residual inhomogeneity is a temporary change in TN around 2003 of about 0.5°C annually. Brandsma (2011) showed that this is related to the growth of trees and bushes close the the measurement site of that time, causing a reduction in minimum temperatures of the same order of magnitude. After pruning the shrubs and trees in September/October 2004 the minimum temperatures went back to normal.

Vlissingen

Vlissingen shows some discrepancies in TN and TX before 1935. The only notable event found in the metadata was in January 1930, mentioning the use of a larger type of screen at the new measurement field.

De Kooy

Den Helder/De Kooy shows a step change in TX around 1925 of about 0.35°C annually. For TN there is a step change of about 0.25°C annually around 1989. We did not find metadata evidence for these steps.

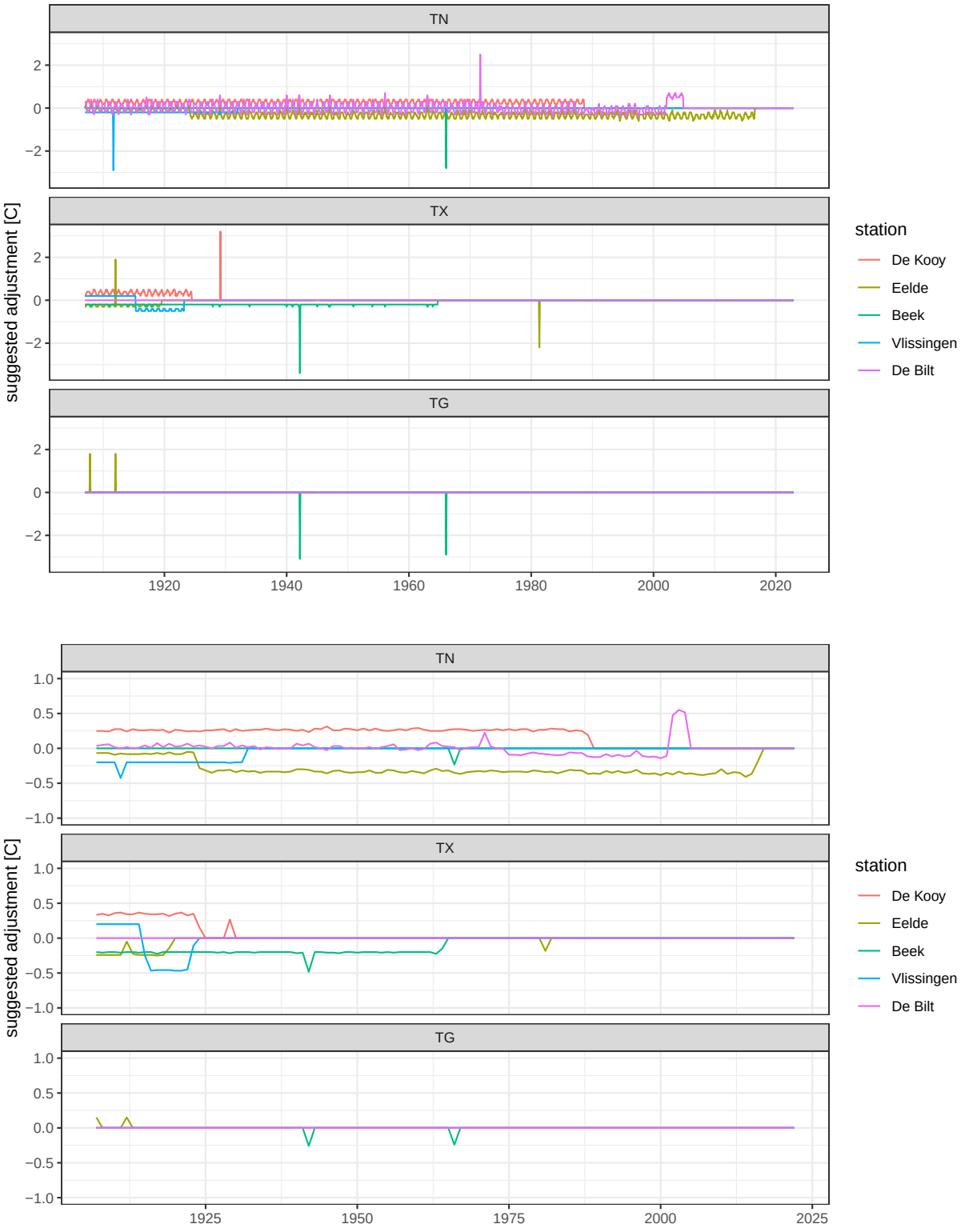

Figure K.1: Monthly (top) and annual (bottom) mean adjustments of the version 2.0 homogenized temperature series suggested by Climatol.

K.2 Checks on the daily temperature distribution

Figure K.2 (top) shows the annual mean of the daily temperature range (TXg - TNg) and the nonlinear trends. The mean values differ greatly (low for De Kooy and Vlissingen near the coast, and high and almost identical for the three inland stations Eelde, De Bilt and Beek). However, the temporal variations of the trends for the five stations are very similar. This consistency is reassuring and suggests that variation in the daily temperature range is primarily driven by regional-scale factors. Even the fluctuations in annual mean range at the three inland stations are very similar, and the same holds for the two stations near the coast.

The annual maxima of the daily temperature range shown in Figure K.2 (bottom) show a similar pattern, although the year-to-year variability is much larger. For De Kooy, the trend shows an anomalous bump before 1950, which may indicate inhomogeneity in the extremes.

A final check considers the asymmetry of the daily temperature distribution, using two derived metrics (see Appendix A):

- the difference between the annual mean daily midpoint temperature (TXg+TNg)/2 and the annual mean temperature TGg, and
- the ratio of the annual mean upward excursion |TXg-TGg| to the annual mean downward excursion |TNg-TGg| relative to the daily mean.

Figure K.3 shows considerable differences between the temporal profiles of these metrics among the stations. The patterns for both metrics are very similar. For some stations (notably Eelde and Beek), the metrics exhibit a smooth trend; for others (in particular Vlissingen), less so. Differences in long-term trends between stations indicate inhomogeneity, possibly due to gradual changes in the surroundings, in sensors or sensor siting etc. For no station does the long-term variation in (TXg+TNg)/2-TGg exceed 0.2° C, so the impact on long-term trends is small.

Overall, the independent checks in this Appendix support the robustness of the homogenized temperature series, while also highlighting a few localized or recent changes that warrant continued monitoring or documentation.

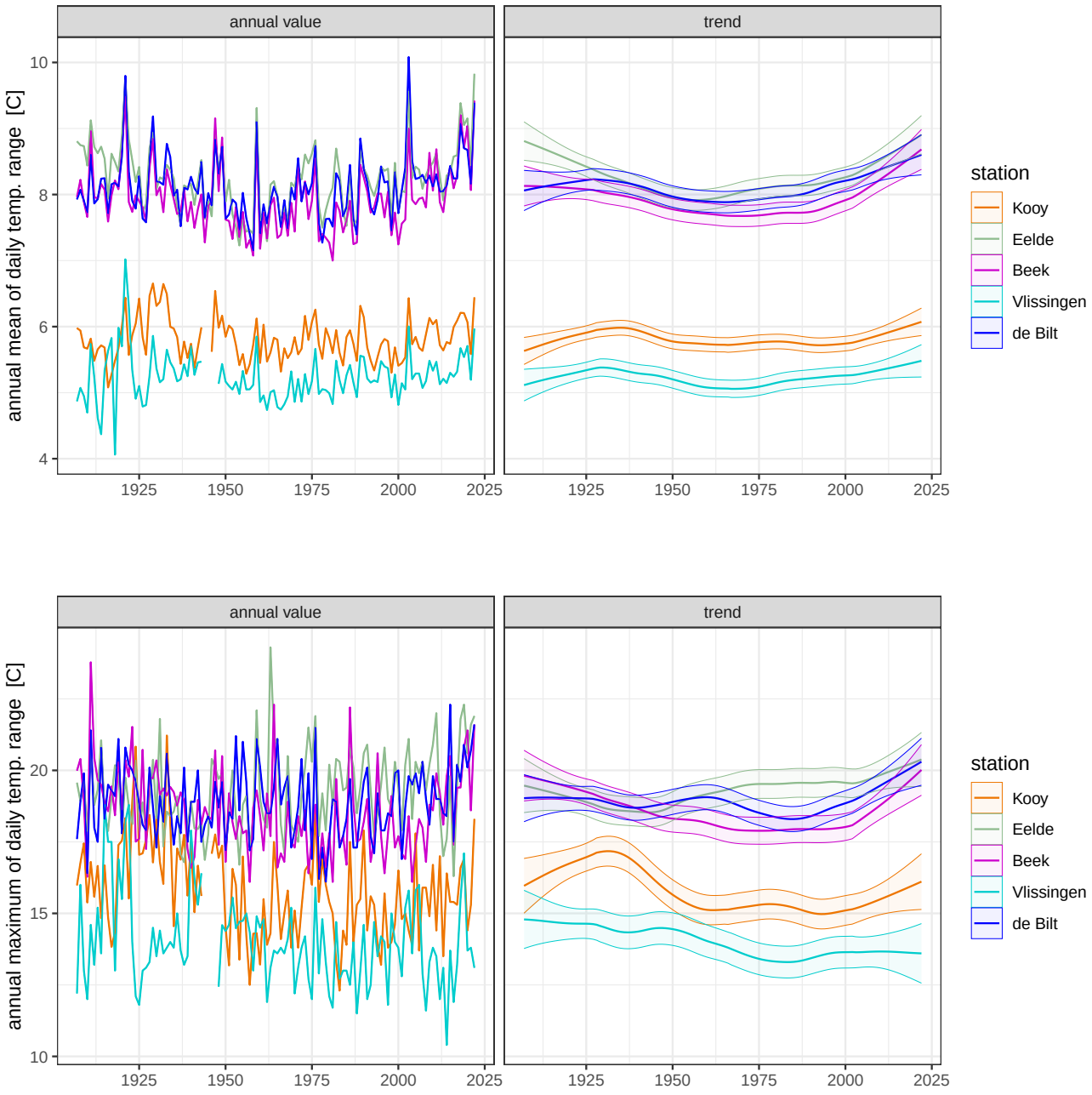

Figure K.2: Annual mean (top) and maximum (bottom) of the daily temperature range for all five stations, and their nonlinear trends

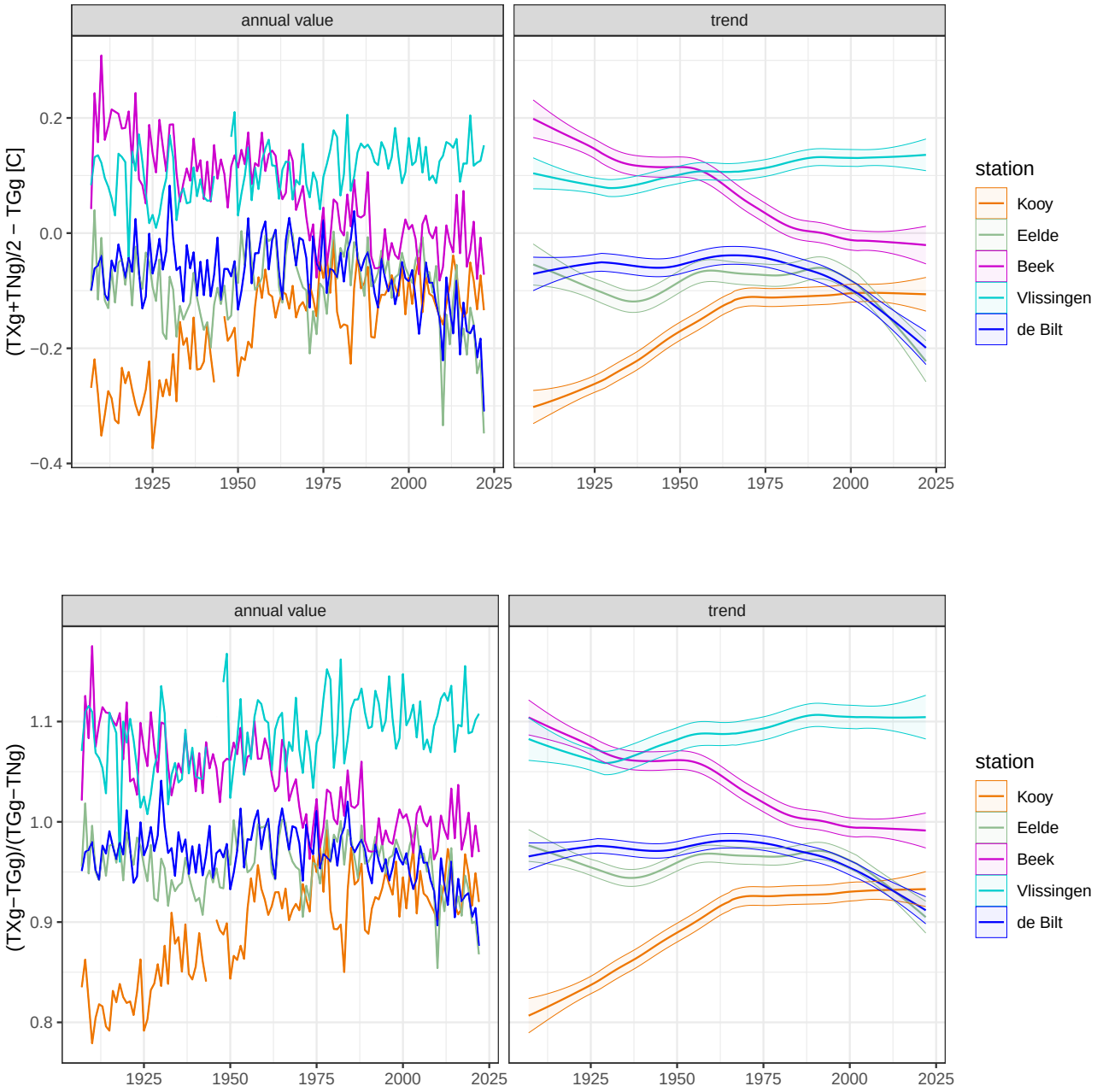

Figure K.3: Asymmetry of the daily temperature distribution according to two metrics: difference metric (top) and ratio metric (bottom); see formulas on the vertical axis labels (see also Appendix A).

Appendix L

Metadata

This appendix provides detailed metadata for the five principal meteorological stations used in this study: De Kooy (Den Helder), Eelde (Groningen), Vlissingen, Beek (Maastricht Airport), and De Bilt. The information includes the types of thermometer screens, temperature sensors, and measurement heights over time, as well as known relocations. This appendix extends the overview provided in Chapter 2, offering additional documentation that supports interpretation and validation of the homogenization process.

L.1 Den Helder/De Kooy

Table L.1 summarizes the thermometer screens, sensors, and measurement heights used at Den Helder and De Kooy.

Known relocations are specified below. Figs. L.1 and L.2 show the location at Den Helder and at airport De Kooy, respectively.

period	screen	sensor	sensor height (m above msl)
19060101-19611123	Stevenson	thermograph	2.20
19611124-19720731	Stevenson	thermograph	1.50
19720801-19921231	Stevenson	resistance	1.50
19930101-present	round multi-plated	pt-500	1.50

Table L.1: Thermometer screen, sensor and sensor height at Den Helder/De Kooy. Thermograph readings were routinely combined with thermometer (regular and min/max) readings at 8, 14, and 19 UTC.

Relocations

- August 1, 1972 Relocation from Den Helder to airport De Kooy.
- August 1972 Relocation of the meteorological instruments on the airport over a distance of 1000 m
- September 1980 The measurement field was relocated to a new area on airport.
- April 2007 The measurement field relocated to the other side of the runway.

Figure L.1: Photograph of the measurements in Den Helder from 1955 taken in northerly direction.

L.2 Groningen/Eelde

Table L.2 summarizes the thermometer screens, sensors, and measurement heights used at Groningen/Eelde.

Known relocations are specified below. Figs. L.3 and L.4 provide visual documentation of the measurement sites in the city of Groningen and at Groningen Airport Eelde.

Relocations

- February 1928 Within the area of the station of the city of Groningen, the Stevenson was relocated to a somewhat more exposed location.
- January 1, 1951 Relocation from the city of Groningen to Groningen airport Eelde.
- May 1973 The measurement location at the airport was relocated over a distance of 750 m.

Figure L.2: Photograph of the automatic weather station De Kooy from 21 November 2024 taken in northerly direction.

period	screen	sensor	sensor height (m above msl)
19060101-19590730	Stevenson	thermograph	2.20
19590731-19730228	Stevenson	thermograph	1.50
19730301-19910315	Stevenson	resistance	1.50
19910316-present	round multi-plated	pt-500	1.50

Table L.2: Thermometer screen, sensor and sensor height at Groningen/Eelde. Thermograph readings were routinely combined with thermometer (regular and min/max) readings at 8, 14, and 19 UTC.

- May 2009 The measurement location at the airport was relocated from south to the runway to north of the runway.

Figure L.3: Photograph of the measurements in the city of Groningen from 1952 taken in northerly direction.

L.3 Vlissingen

Table L.3 provides the history of thermometer screens, sensors, and measurement heights used at Vlissingen.

In addition, Vlissingen experienced temporary relocations as specified below. Figs. L.5 to L.7 give an impression of these locations.

Figure L.4: Photograph of the automatic weather station Groningen Airport Eelde from 19 November 2024 taken in northerly direction.

period	screen	sensor	sensor height (m above msl)
19060101-19300114	unspecified screen enclosure*	thermograph	2.20
19300115-19431102	Stevenson	thermograph	2.20
19431103-19450731	Stevenson	thermograph	8.80
19450801-19611012	Stevenson	thermograph	2.20
19611013-19740930	Stevenson	thermograph	1.50
19741001-19930430	Stevenson	resistance	1.50
19930501-present	round multi-plated	pt-500	1.50

Table L.3: Thermometer screen, sensor and sensor height at Vlissingen. Thermograph readings were routinely combined with thermometer (regular and min/max) readings at 8, 14, and 19 UTC. Notes: * in some documents specified as 'a kind of cage'; perhaps an early Stevenson variant.

Relocations

- 19431103-19450731 Hotel Brittannia, located at 51°27' N. 03°33' E, with ground level at 8.0 m above msl.
- 19450801-19470815 Hotel Noordzee Boulevard, located at 51°26' N 03°35' E, with ground level at 8.0 m above msl.
- 19470815-19580429 West Souburg, located at 51°28' N 03°35' E, with ground level 0.5 m below msl.

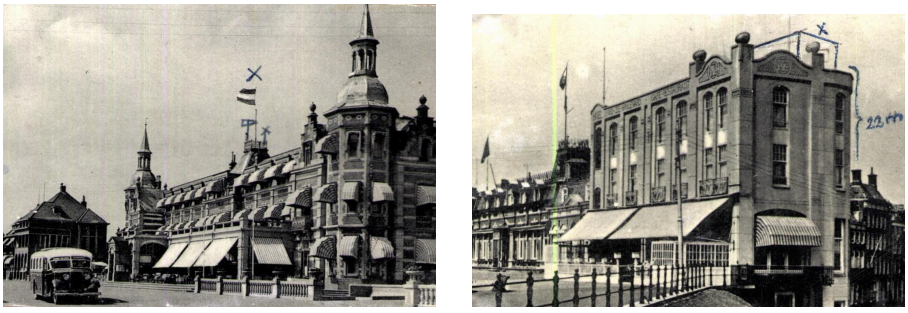

Figure L.5: Photographs of the temporary locations Hotel Brittannia (left) and Hotel Noordzee Boulevard (right).

L.4 Maastricht/Beek

Table L.4 summarizes the thermometer screens, sensors, and measurement heights used at Maastricht/Beek.

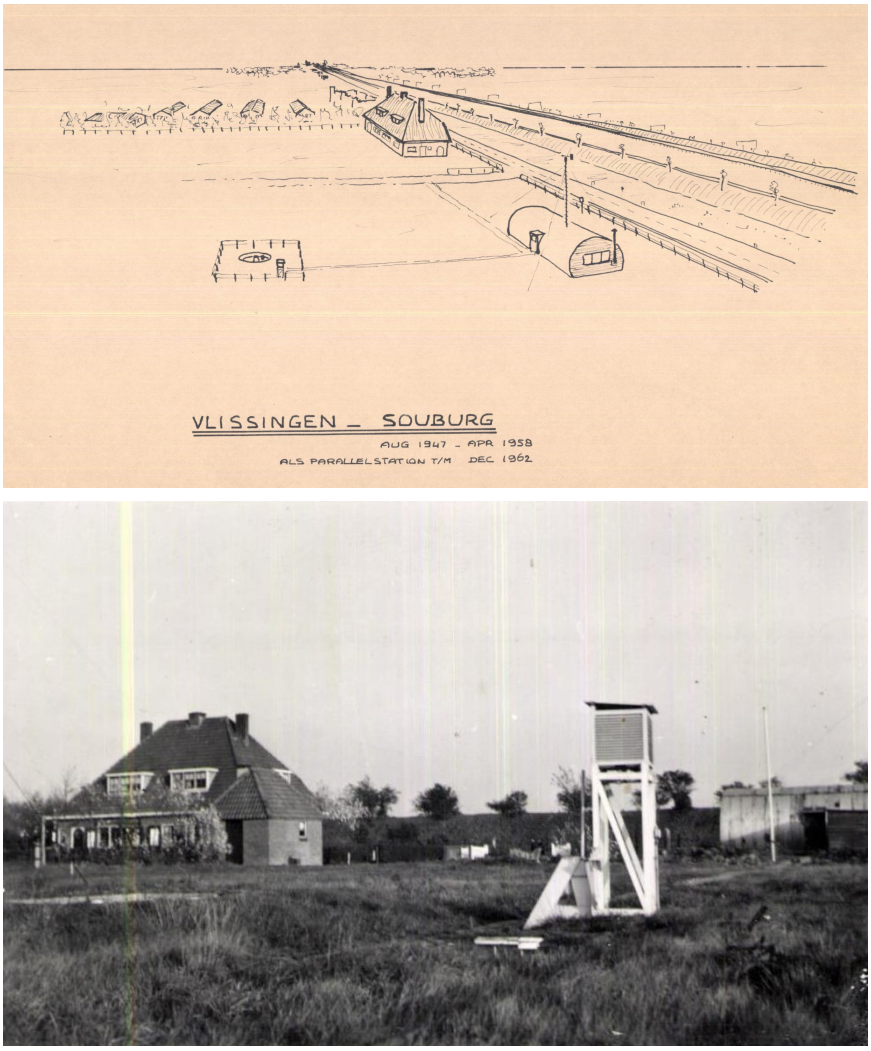

Figure L.6: Drawing of the temporary location Souburg (top) and a photograph of the thermometer screen from 25 October 1955 taken in east-northeasterly direction (bottom).

Figure L.7: Photograph of Vlissingen from 5 September 2024 taken in northerly direction.

In addition, Maastricht/Beek underwent relocations as listed below. Figs. L.8 to L.9 give an impression of these locations.

period	screen	sensor	sensor height (m above msl)
19060101-19451130	Stevenson	thermograph	20.10
19451201-19610413	Stevenson	thermograph	2.20
19610414-19760131	Stevenson	thermograph	1.50
19760201-19910228	Stevenson	resistance	1.50
19910301-present	round multi-plated	pt-500	1.50

Table L.4: Thermometer screen, sensor and sensor height at Maastricht/Beek. Thermograph readings were routinely combined with thermometer (regular and min/max) readings at 8, 14, and 19 UTC.

Relocations

- January 1, 1951 Relocation from the city of Maastricht to Beek airport.
- November 2005 Measurement field relocated on the airport over a distance of 1770 m.

L.5 De Bilt

Table L.5 summarizes the thermometer screens, sensors, and measurement heights used atDe Bilt.

Figure L.8: Photograph of the measurements on top of the HBS school in the city of Maastricht taken in 1939.

Figure L.9: Photograph of the automatic weather station Maastricht Airport from 11 December 2024 taken in southerly direction.

In addition, De Bilt experienced relocations as specified below. Figs. L.10 to L.13 give an impression of these locations.

period	screen	sensor	sensor height (m above msl)
19010101-19500516	pagoda	thermograph	2.20
19500517-19610628	Stevenson	thermograph	2.20
19610629-19930625	Stevenson	thermograph	1.50
19760201-19910228	Stevenson	resistance	1.50
19930326-present	round multi-plated	pt-500	1.50

Table L.5: Thermometer screen, sensor, and sensor height at De Bilt. Thermograph readings were routinely combined with thermometer (regular and min/max) readings at 8, 14, and 19 UTC.

Relocations

- 16 September 1950 The measurement site in De Bilt relocated 80 m westward.
- 27 August 1951 The measurement site was moved 300 m southward.
- 26 September 2008 The measurement site relocated 220 m to the east.

The relocation on 26 September 2008 is discussed in Brandsma (2011). It followed a period with a temporary inhomogeneity in temperature at the older site. Note that this inhomogeneity is also discussed in Appendix K.

Figure L.10: Photograph of the measurements at the KNMI site in De Bilt from 1932 taken in northerly direction.

Figure L.11: Photograph of the temperature measurements at the KNMI site in De Bilt from April 1950 taken in northwesterly direction.

Figure L.12: Photograph of the temperature measurements at the KNMI site in De Bilt from 1953 taken in southerly direction.

Figure L.13: Photograph of the measurement field at the KNMI site in De Bilt from 27 August 2024 taken in easterly direction.

Royal Netherlands Meteorological Institute

PO Box 201 | NL-3730 AE De Bilt
Netherlands | www.knmi.nl