Received: 3 July 2023 | Accepted: 8 September 2023

DOI: 10.1002/joc.8258

RESEARCH ARTICLE

International Journal of Climatology RMetS

Homogenization of daily temperatures using covariates and statistical learning—The case of parallel measurements

Cees de Valk | **Theo Brandsma**

KNMI, De Bilt, The Netherlands

Correspondence
Theo Brandsma, KNMI, PO Box 201, 3730 AE De Bilt, The Netherlands.
Email: theo.brandsma@knmi.nl

Abstract

A data driven method based on generalized additive modelling (GAM) has been developed for homogenizing daily minimum and maximum temperature (TN, TX) series using parallel measurements and covariates. The method is applied to two coastal and two inland stations in the Netherlands. Between 1950 and 1972, these stations were relocated from cities to airports, accompanied by parallel measurement of at least 5 years at the old and new sites. Separating these parallel measurements in training and test data, the method compares numerous models involving covariates like the wind vector, cloudiness, specific humidity and sea surface temperature, and selects a model for each station. The resulting models offer an improvement compared to models based on temperature and season only: seasonal dependence is largely replaced by dependence on physical quantities. However, quantitatively, the impact is not large in the cases studied. One of the reasons might be that some covariates have only been measured at specific times not coinciding with the occurrences of the temperature minima or maxima. Additional benefits of the method are robustness and estimation of the sampling error variance of the daily homogenized daily temperature values.

KEYWORDS

climatology, cross-validation, generalized additive model, homogenization, machine learning, parallel measurements, temperature

1 | INTRODUCTION

While methods for homogenizing temperature time series on a monthly scale are well described (e.g., Venema et al., 2012), the homogenization of series on the daily scale is still a topic of ongoing research. The latter is of special interest for studying weather extremes and variability, and is the topic of the current article.

Methods for homogenizing daily temperatures may involve two steps: break detection and break adjustment. Break detection is usually performed on aggregated seasonal or monthly values. Once the breakpoints are known, daily adjustments are calculated using data from nearby stations not having breaks close to the breakpoint of interest. Squintu et al. (2020) compared several of these methods on an observation-based benchmark dataset. The methods used quantile matching or regression or a combination of these to adjust the daily values. Not one single method seemed to outperform the other; Squintu et al. (2020) concluded that the choice of a method should be guided by the user needs.

The purely statistical daily adjustment techniques usually do not take into account other weather variables than temperature. It is known, however, that temperature differences between sites and sensors depend not only on temperature itself, but also on variables like wind speed and cloudiness (e.g., Brandsma & Van der Meulen, 2008). Physics-based correction methods can account for this. For instance, Auchmann and Brönnimann (2012) used such a method for homogenizing sub-daily temperature series. Their method requires daily wind speed and cloud cover and is applied to a transition of the Wild temperature screen to the Stevenson screen in Basel, Switzerland. It is evaluated using parallel measurements.

The WMO Guidelines on Homogenization (WMO, 2020) stress the importance of performing parallel measurements, stating: 'The best possible set of candidate and reference series is a parallel observation of the old and new situation at the same site, for example, when a new instrument system is introduced, while the old system continues to be in operation for a period of time'. When all variables of interest are measured while performing parallel measurements, the weather dependence of the candidate variable can be studied and accounted for when homogenizing the variable. The latter is, however, only possible when the weather variables of interest are also present for the period to be homogenized.

In the present study, we introduce a data driven method based on generalized additive modelling (GAM) for homogenizing daily temperatures. The method is in between the physical approach and the purely statistical approach. It has been developed for homogenizing daily temperature series using parallel measurements and covariates. The main advantage of GAM models is that they are flexible: they can incorporate any variable of interest and do not require a prescribed form of the relationships to be modelled. Furthermore, these models are interpretable, and theory and software for GAM models are mature (Hastie & Tibshirani, 1990; Wood, 2020). We combine GAM model fitting with a data driven approach for selection of covariates and their interactions. We demonstrate the method for homogenizing daily maximum and minimum temperatures at two coastal and two inland weather stations in the Netherlands. These stations have been subject to a major relocation, accompanied with multi-year parallel measurements between the old and new location.

The framework can in principle be applied to other variables than temperature and to inhomogeneities due to other causes (like changes in instruments and setup). Furthermore, we anticipate to extend it to the situation without parallel measurements, employing time series from nearby sites—homogeneous over common time-intervals before and after the break—as covariates.

2 | DATA AND METHODS

2.1 | Data

The data to be homogenized consists of the daily (0–0 UTC) minimum (TN) and maximum (TX) temperatures of four of the five principal stations of KNMI in the Netherlands (Figure 1). The following major relocations took place:

1. Measurement at De Kooy started on 1 August 1972 as a continuation of Den Helder. Den Helder is located behind the North Sea dike on the Western edge of the city, whereas De Kooy is an exposed location on the airport on the SE edge of Den Helder, about 1 km from the Wadden Sea.
2. Measurement at Eelde started in 1951 as a continuation of Groningen. Groningen was located in the city whereas Eelde is an exposed location on the airport at about 10 km south of Groningen.
3. Vlissingen was temporally relocated to Souburg from 1947 to 1958. Vlissingen is an exposed location in the harbour on the Westerschelde estuary; Souburg was an exposed airport location further inland at 1.8 km NNW of Vlissingen.
4. Maastricht was until 1950 situated in the city of Maastricht, and then relocated to an exposed location on

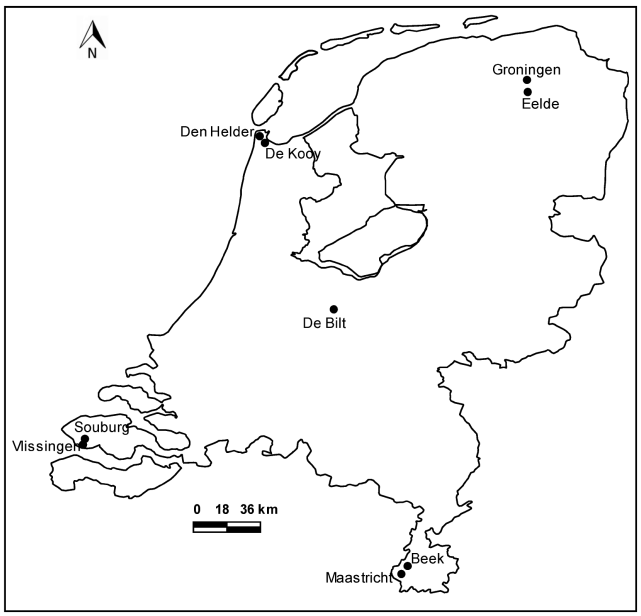

FIGURE 1 Locations of the five principal stations of the Netherlands. Four of them moved from a city location to an airport location: Den Helder/De Kooy (06235), Groningen/Eelde (06280), Vlissingen/Souburg (06310) and Maastricht/Beek (06380) and are the subject of the present study. Values in parentheses are WMO station numbers.

Beek airport in 1951, at about 9 km NE of Maastricht. The altitude of Beek airport is about 65 m higher than of Maastricht. Temperature measurements in Maastricht were taken at 20 m above ground level, strongly deviating from the standard measurement height of 2.2 m above ground level at that time.

Table 1 presents further details about the stations. Station Den Helder has a gap from September 1944 to May 1945.

Figure 2 shows the monthly mean temperature differences in the overlapping periods. Moving to locations further inland (Den Helder to de Kooy and Vlissingen to Souburg) results—as expected—in a more continental type seasonal temperature variation with lower minimum temperatures and higher maximum temperatures. Also an effect on the timing of the peaks of the differences in TN and TX is visible. The relocation from the city of Groningen to airport Eelde mainly affected TN. The relocation from the city of Maastricht to airport Beek affected TN and TX equally. A possible explanation is that the measurements in Maastricht took place on the top of a roof, so both sites are exposed.

Besides the daily TN and TX data, hourly data of wind speed (F), wind direction (D), cloudiness (N) and relative humidity (RH) are available as potential covariates. 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 calculated specific humidity (HUM) from T and RH (WMO, 2014) as a

TABLE 1 Details of the stations shown in Figure 1.

Station	LAT (N)	LON (E)	ALT (m)	Operational period	Overlap
Den Helder/	52.967	4.750	4.4	1906–1972/07	1960–1970
De Kooy	52.924	4.785	0.5	1972/08–present	(10 years)
Groningen/	53.217	6.550	2.1	1906–1950	1946–1951
Eelde	53.125	6.586	3.5	1951–present	(6 years)
Vlissingen/	51.442	3.596	8.0	1906–present	1958/05–1962
Souburg	51.467	3.583	−0.5	1947–1958	(∼5 years)
Maastricht/	50.850	5.693	49.4	1906–1950	1946–1952
Beek	50.910	5.768	114.0	1951–present	(7 years)

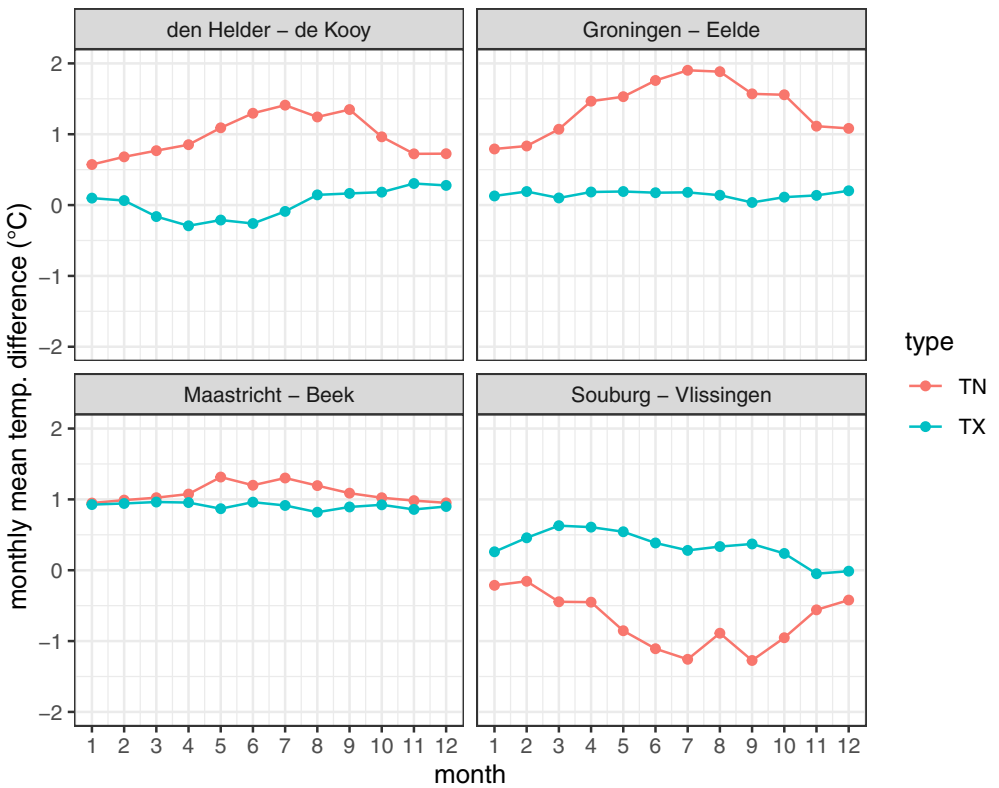

FIGURE 2 Monthly mean temperature differences between the old and new station locations (see Figure 1 for locations). [Colour figure can be viewed at wileyonlinelibrary.com]

measure of the dryness of the air. We then used HUM as a covariate instead of RH. For TN we used the morning values of N and HUM and for TX the afternoon values. F and D are combined into a wind vector U with an easterly and northerly component. For TN we used U in the hour that TN occurs (tTN) and for TX 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 included SST as a covariate. 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, we estimated daily values from the observed monthly averages using a GAM model (see Section 2.2) variant which can use linear functionals as observations, with the degree of smoothing tuned on daily values from Den Helder over the sub-period.

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

- A tiny part (about 0.01%) of the needed hourly values of T, FF, D and RH were missing. These were filled in with the data of principal station De Bilt in the centre of the Netherlands (Figure 1).
- From 1951/03/01 onward D and F were no longer measured at Maastricht. Instead D and F of Beek have been used for the period (1951/03/01–1952/12/31). As Maastricht and Beek have different windspeeds, we used the overlap part 1948/01/01–1950/12/31 to correct Beek to Maastricht using quantile matching.
- The F data of 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 the overlap period 1931/01/01–1950/12/31.
- For 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 in the period 1906–1950. We used a loess smoother (span = 0.4) to find monthly correction values to correct the F values before 1928 to the values thereafter.
- From 1961/01/01 onwards N is no longer measured at Souburg. Instead cloud cover of Vlissingen for the period 1961/01/01–1962/12/31 has been used.

2.2 | Methods

2.2.1 | Regression model and estimation

Say we want to homogenize the measurements y_{src} ('source') of a variable y (in the present case TN and TX), regularly sampled in time t (in our case, t is indicated by the date). The homogenized value $y_{hom}(t)$ is a nonlinear function of $y_{src}(t)$ and of other covariates (in our case wind vector components, cloud cover, specific humidity, sea surface temperature and seasonality).

Over a certain time-interval, we have parallel measurements of y: y_{src} from the original site/sensor, and y_{ref} from the reference site/sensor, which we try to approximate based on $x_1 = y_{src}$ and possibly other covariates $x_2, \ldots, x_m$. The approximation is based on an assumed relationship between y_{ref} and $x_1 = y_{src}, x_2, \ldots, x_m$ of the general form

$$y_{ref}(t) = s(x_1(t), \ldots, x_m(t)) + \varepsilon(t), \tag{1}$$

with s a smooth real function on $\mathbb{R}^m$ and ε an error term, indicating that the approximation may not be exact.

The function s adds the effects of the individual covariates and of all their interactions. Because interactions between certain (groups of) covariates may be assumed to be absent, we distribute the covariates over d groups

$$\mathbf{x}_j = \{x_{m_{j-1}+1}, \ldots, x_{m_j}\}, \quad j = 1, \ldots, d \tag{2}$$

(with $m_0 = 0$ and $m_d = m$), and represent s as the sum of d smooth functions $s_1, \ldots, s_d$ modelling the effects of these groups, plus an intercept s_0:

$$s(x_1, \ldots, x_m) = s_0 + \sum_{j=1}^{d} s_j(\mathbf{x}_j). \tag{3}$$

We require that the empirical mean $\frac{1}{n}\sum_{i=1}^{n} s_j(\mathbf{x}_j(t_i))$ of $s_j(\mathbf{x}_j)$ over the period of parallel measurements vanishes for $j = 1, \ldots, d$, so the intercept s_0 is determined unambiguously.

Each s_j for $j > 0$ is assumed to be of the form

$$s_j(\mathbf{x}_j) = \sum_{k=1}^{N_j} \beta_{k,j} b_{k,j}(\mathbf{x}_j), \tag{4}$$

with the $b_{1,j}, b_{2,j}, \ldots$ a large number of known smooth functions which are precomputed, and $\beta_{1,j}, \beta_{2,j}, \ldots$ the parameters to be estimated, so s_j is linear in the parameters (Wahba, 1990). Such models are examples of GAMs, for which powerful likelihood-based estimation methods have been developed, which estimate the smoothness of each function s_j from the data (Wood, 2020). We use the function gam.R from R-package mgcv with the thin-plate smoothing spline and standard settings (e.g., Gaussian and independent daily $\varepsilon(t)$), except for the 'method',

which is set to REML. Some further details are provided in Appendix A.

This produces a smooth regression model of the form (1). Inevitably, the variance of the approximation $s(x_1,...,x_m)$ of y_{ref} is smaller than the variance of y_{ref} itself (as the former approximates the mean of y_{ref} conditional on the covariates). For climatological analysis, this is undesirable. Ideally, we would estimate s under the additional constraint that the distribution of the prediction $s(x_1,...,x_m)$ matches the distribution of y_{ref}, or to simplify, that the empirical variance of $s(x_1,...,x_m)$ matches the empirical variance $VAR\left(y_{ref}\right)$ of y_{ref}:

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left|s(x_1(t_i),...,x_m(t_i))-s_0\right|^2\\&=VAR\left(y_{ref}\right)\\&=\frac{1}{n}\sum_{i=1}^{n}\left|y_{ref}(t_i)-\bar{y}_{ref}\right|^2\end{aligned}\tag{5}$$

with $\bar{y}_{ref}$ the mean of y_{ref}. However, we do not have software capable of estimating the functions $s_1,...,s_d$ including their smoothness under this constraint. Therefore, we directly adjust the variance of $s(x_1,...,x_m)$ to ensure that the empirical variances of the homogenized values y_{hom} and of the reference values y_{ref} match:

$$y_{hom}(t)=s_0+\alpha(s(x_1(t),...,x_m(t))-s_0),\tag{6}$$

with

$$\alpha=\sqrt{\frac{nVAR\left(y_{ref}\right)}{\sum_{i=1}^{n}\left|s(x_1(t_i),...,x_m(t_i))-s_0\right|^2}}.\tag{7}$$

This variance matching is equivalent to multiplying all parameters $\beta_{k,j}$ in Equation (4) by the same factor α. In Appendix B, we show that variance matching is in some sense equivalent to estimation under the constraint (5).

Variance matching (or 'inflation') is commonly applied in for example, paleo-climate reconstruction (Bürger & Cubasch, 2005) and in bias-correction of climate model projections (Maraun, 2013). For these applications, it may have serious limitations due to mismatch in the temporal and/or spatial scales resolved in the two datasets involved (model output/measurements, or proxy data/measurements) (Maraun, 2013; Von Storch, 1999). However, in the present application, scale mismatch is not an issue.

In the discussion of these models in Section 3, we will avoid their mathematical details and only indicate the grouping of covariates, as in, for example,

$$\{y_{src},x_2,x_3\}+\{x_4,x_5\}\tag{8}$$

for a model (1), (3), (6) and (7) with two additive smooth functions of 3 and 2 covariates, respectively ($d=2$, $\mathbf{x}_1=(y_{src},x_2,x_3)$, and $\mathbf{x}_2=(x_4,x_5)$).

2.2.2 | Selection of covariates and interactions

A further problem is the selection of covariates to include in the model and of the interactions between covariates to be estimated (the groups in Equation 3). Simple models, with few covariates and small groups (and therefore, little interaction between covariates), are generally easier to estimate than complex models, even though the latter may have smaller residual mean squared error (*MSE*).

We use cross-validation to select models: for every year Y with simultaneous temperature measurements from the original and reference sites or sensors, the model is estimated from the data of the remaining years, and the *MSE* of the model predictions for the year Y is computed. Then these *MSE*'s are averaged.

Several strategies are known to select a model from a set of alternative covariate combinations and groupings (James et al., 2013); for example, (1) an exhaustive search through all possible models, or (2) forward selection: starting from the simplest models, add a covariate and/or grow groups, test these, and only retain the best model found so far before further increasing the number of covariates in the model.

After some experimentation, we adopted a cautious variant of forward selection:

- Certain covariates such as wind vector components are only added or skipped together.
- Limits can be imposed on the number of covariates and on group size.
- To avoid getting stuck too early with a sub-optimal specification of the interactions, we allow exhaustive search for models having up to three covariates.
- When a new covariate is to be added, we do not necessarily start from the best model (with lowest *MSE*) found so far, but also consider models with the same covariates but somewhat simpler interactions, and start from the best of these if its *MSE* is less than 5% higher than the lowest *MSE*. The idea is that if we take such a simpler model to the next level, then it will generate all models which would be generated from the best, more complex, model, but also a number of other similar models.

This selection method runs automatically and produces a list of all models tested, sorted with increasing

MSE. This allows the analyst to either take the best model (which is likely to generalize well, due to cross-validation) or apply an additional trade-off between performance and other desirable properties such as simplicity, consistency of model choice across sites or physical plausibility.

In the present study, we selected the simplest model (with the least number of variables and then with the least number of interactions between variables) from all models with *MSE* less than 1% above the lowest *MSE*.

2.2.3 | Uncertainty of the homogenized time-series

Due to the flexibility of the model class considered, bias in the regression is likely a minor issue. Distributional bias in the homogenized time series is handled by the variance matching, which is adequate for temperature data (but probably not for variables with heavily skewed distributions such as precipitation). Aside from bias, we address two types of error:

1. deviations between homogenized values and the values which have or would have been measured at the new (reference) site or sensor (prediction errors)
2. errors in predictions due to sampling error in model parameters estimated from a limited dataset of simultaneous measurements from the two sites or sensors.

Error (1) includes error (2), but also the effects of spatial fluctuations in weather and the loss of coherence between temperatures and covariates by temporal averaging or interpolation in time; see Section 2.1. We can only estimate an overall mean squared error (*MSE*) from the parallel measurements, or a derived skill score such as r^2:

$$r^2 = 1 - MSE/VAR\left(y_{ref}\right). \tag{9}$$

Such scores are useful for comparing the predictive skills of models. We derive them from the cross-validation used for model selection. They do not fully express the quality of the homogenization: we could simply increase them by skipping the variance matching, but this would result in bias in the temperature distribution.

For climatological analyses of the homogenized temperatures, the sampling error (2) is particularly relevant; it shows whether the set of parallel measurements is large enough for a reliable homogenization. The GAM fitting method (Wood, 2020) supplies an estimate of the sampling error standard deviation for each day, which needs to be rescaled by α; see Equations (6) and (7). However, the residuals are not independent, which violates the assumptions underlying these estimates. Therefore, we have used the cross-validation results to calibrate the rescaled GAM variances; see Appendix C.

2.2.4 | Replication

Code and data used for the project can be accessed on https://gitlab.com/cees.de.valk/homogenization-using-covariates-and-statistical-learning.

3 | RESULTS

The fitted GAM models use selected combinations of the variables or vectors (see Section 2.1 for further details):

- measured daily minimum or maximum temperature T_{src} to be homogenized,
- wind vector $U = (u_E, u_N)$ with eastward and northward components u_E and u_N,
- cloud cover N,
- specific humidity HUM,
- sea surface temperature SST (only for the stations Vlissingen and de Kooy near the coast),
- seasonality vector $SEAS = (\cos(\theta), \sin(\theta))$, with θ the day number in 1,..., 365 multiplied by $2\pi/365$ (Feb 29 is assigned the same day number as Feb 28).

Initial experiments showed that wind vector components offered as distinct variables are generally selected together. Therefore, the wind components are always added or skipped together as a vector, and so are the two sinusoidals representing the season. For computational reasons, models are restricted to have at most 4 variables or vectors. Furthermore, a single smooth involves at most 4 scalar variables (vectors counting for 2 here). The total number of models estimated and evaluated for each case varied from 42 to 58.

For TN, Table 2 shows for each station relocation the selected model, and some cross-validation statistics: *RMSE* (the square root of *MSE*), r^2 (see Equation 9), and furthermore the scores

$$S_1 = 1 - MSE/MSE_1, \quad S_2 = 1 - MSE/MSE_2,$$

with MSE_1 the mean squared error of the simplest model $T_{hom} \sim s(T_{src})$, and MSE_2 the mean squared error of the model $T_{hom} \sim s(T_{src}, SEAS)$ which also contains seasonality. The latter two statistics are very similar to r^2; they give the fraction of the residual mean squared error of a simple model explained by the selected model (hence,

TN	Model	*RMSE*	r^2	S_1	S_2
den Helder → de Kooy	$\{T_{src}\}+\{U,N\}+\{SST\}$	0.96	0.97	0.30	0.26
Souburg → Vlissingen	$\{T_{src},SST\}+\{U\}+\{N\}$	0.60	0.99	0.69	0.45
Maastricht → Beek	$\{T_{src}\}+\{U\}+\{HUM\}$	0.62	0.99	0.06	0.06
Groningen → Eelde	$\{T_{src}\}+\{U,N\}$	0.93	0.98	0.36	0.28

TABLE 2 Estimated/selected models for homogenization of daily minimum temperature TN with cross-validation root mean squared error *RMSE* (°C) and skill scores r^2, S_1 and S_2 (see Equation 8, for notation).

TX	Model	*RMSE*	r^2	S_1	S_2
den Helder → de Kooy	$\{T_{src},HUM\}+\{U,SEAS\}$	0.57	0.99	0.29	0.21
Souburg → Vlissingen	$\{T_{src},SST,U\}$	0.45	1.00	0.53	0.37
Maastricht → Beek	$\{T_{src}\}+\{HUM\}+\{U,N\}$	0.51	1.00	0.13	0.11
Groningen → Eelde	$\{T_{src},HUM\}+\{U\}+\{N\}$	0.50	1.00	0.26	0.24

TABLE 3 As Table 2 for TX.

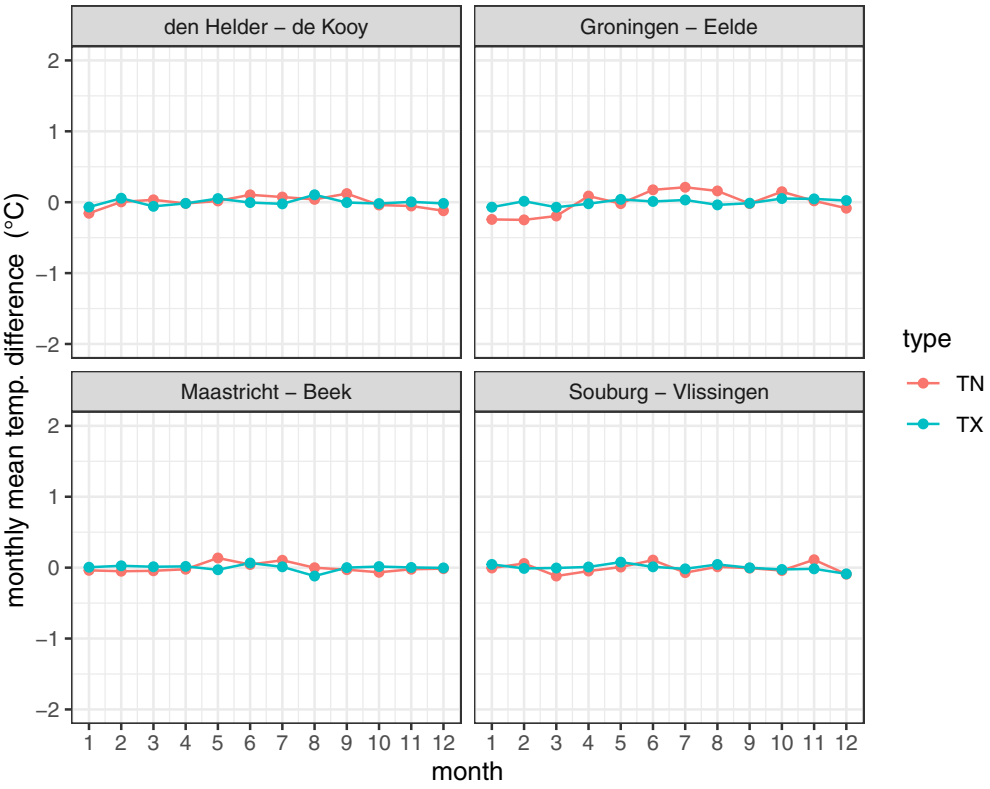

FIGURE 3 Monthly mean differences between the homogenized and reference temperatures (compare Figure 2). [Colour figure can be viewed at wileyonlinelibrary.com]

the relative improvement with respect to the simple model). Table 3 presents the selected models and their cross-validation statistics for TX.

T_{src} and wind vector are consistently selected as covariates. Seasonality (represented by the vector SEAS) is absent from all models except one (TX at de Kooy).

For the stations de Kooy en Vlissingen close to the coast, the sea surface temperature SST is selected instead; its smooth seasonal cycle (in combinations with other covariates) apparently captures the seasonal dependence effectively. However, seasonality is also missing from the models selected for the two stations further inland. Therefore, variables like wind, cloud cover, humidity and SST apparently capture the variation in the relation between y_{src} and y_{ref} over the year effectively. This is reinforced by comparing Figure 3, showing the mean temperature difference between homogenized and reference temperature per month, with Figure 2.

The r^2 values of the selected models are high, in particular for TX. The other skill scores S_1 and S_2, which express the fractional improvements relative to simpler models, are similar for TN and TX but vary strongly among the sites. The largest improvement is achieved for Vlissingen, and the smallest for Beek. Overall, the improvements are modest but not negligible.

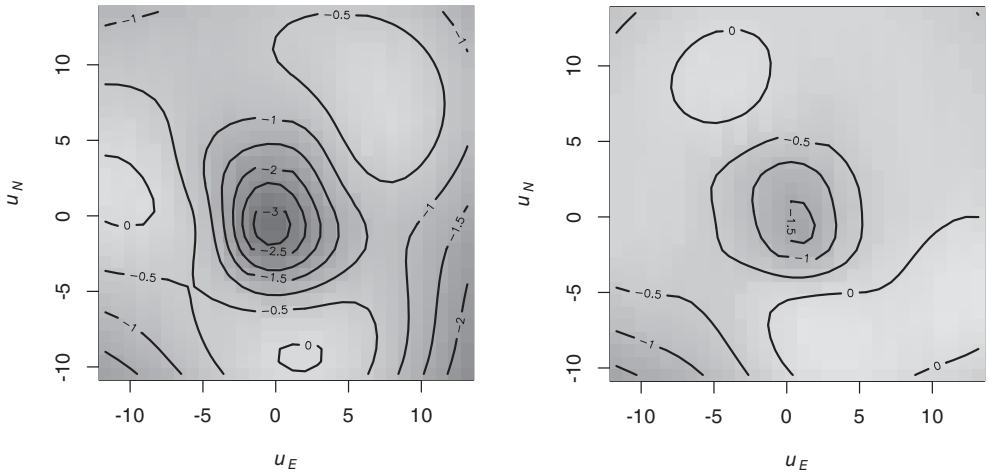

FIGURE 4 GAM model for daily minimum temperature TN at Eelde: response to wind components u_E and u_N (in m/s) for clear sky $N=0$ (left) and fully covered sky $N=10$ (right) for temperature at Groningen $T_{src}=0$ (the same brightness scale is used in both panels).

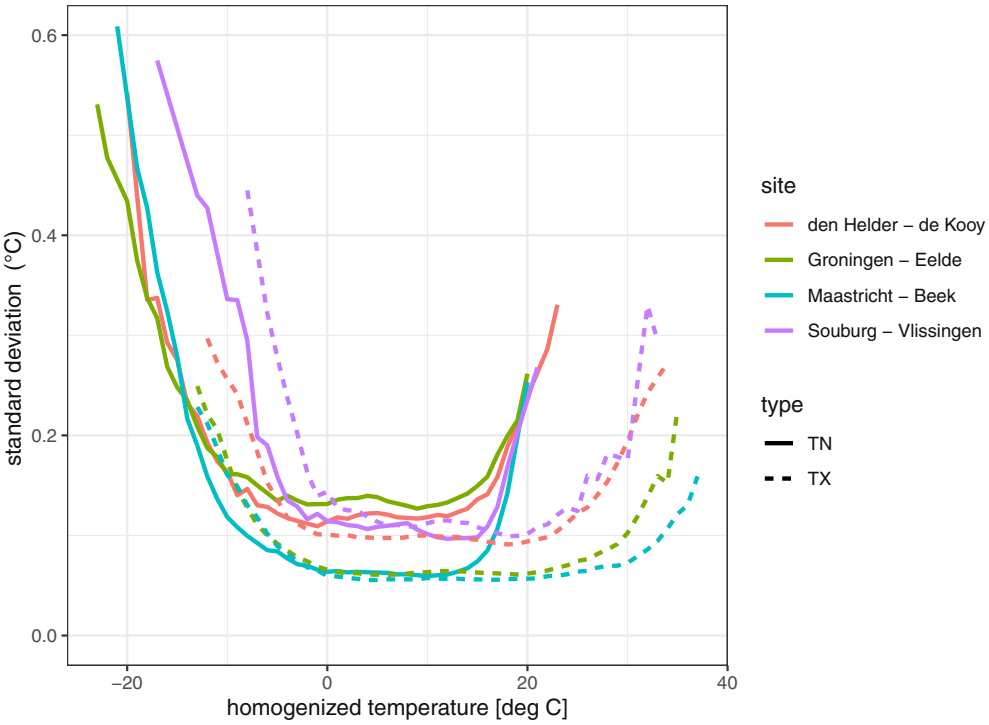

FIGURE 5 Median of the standard deviation of the sampling error of the homogenized temperature as a function of homogenized temperature per site and for TN and TX. [Colour figure can be viewed at wileyonlinelibrary.com]

The most frequently selected interaction is between wind and cloud cover (three cases). Because wind is a vector, this is in fact an interaction between three variables. A typical pattern for TN is shown Figure 4: at low wind speeds, the temperature at Eelde is considerably lower than at Groningen. This difference is modulated by cloud cover; it is stronger with lower cloud cover. This makes physical sense: with low wind speeds and clear sky, local radiative cooling can be strong and mixing weak, so spatial gradients in TN can be large. Due to mismatches between the exact times of the measurements of TN, wind and cloud cover, we may still underestimate this effect.

Figure 5 gives an overview of the estimated standard deviations of the sampling error of the homogenized temperature records. The estimates of sampling error are affected by the number of years of parallel measurements (small for Vlissingen) and associated positively with *MSE* and with the skill score S_1, which indicates how large the effect of covariates is. For example, sampling error is low for Beek, because the effects of covariates other than the temperature at Maastricht are small. The sampling error of midrange TN or TX is negligible compared to the *RMSE* values in Tables 2 and 3. For extreme TN and TX, the sampling error is more substantial.

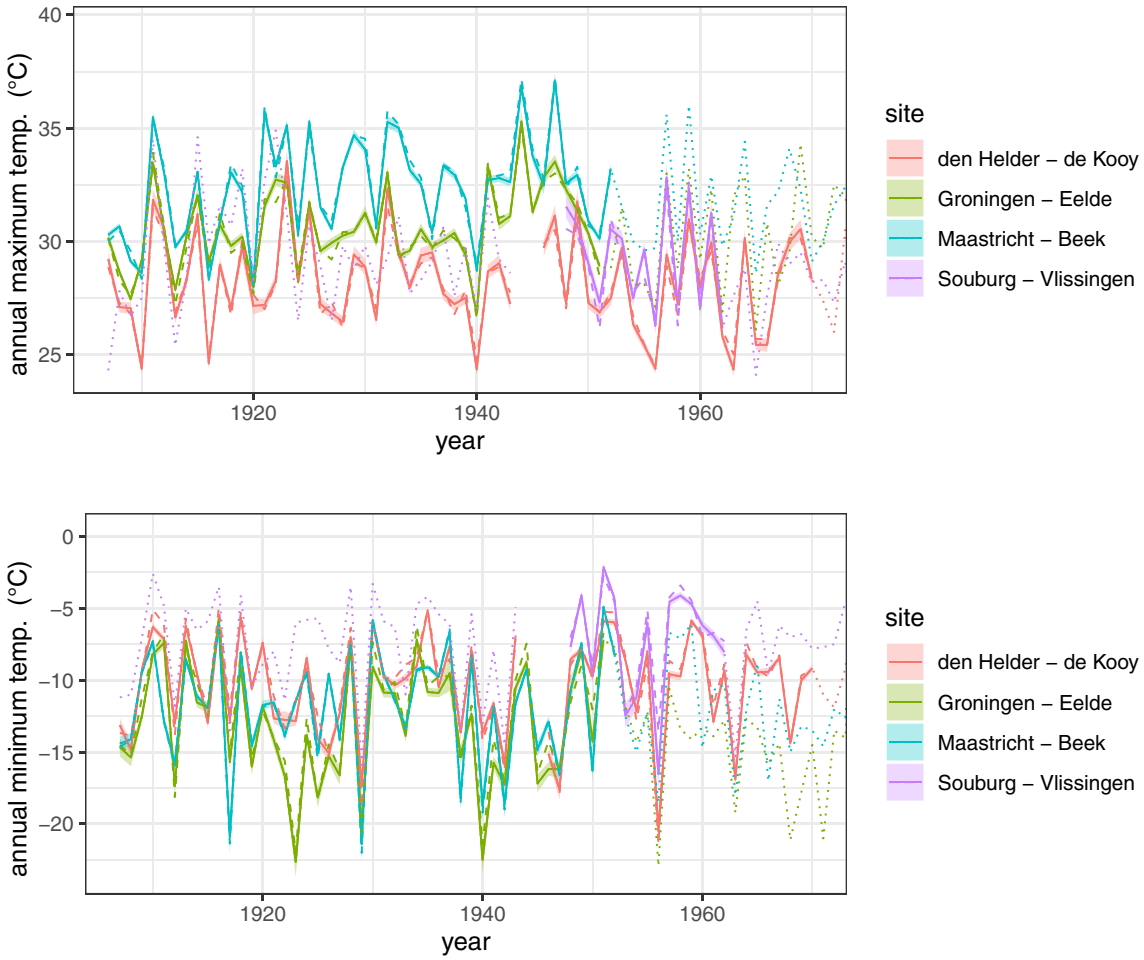

FIGURE 6 Homogenized annual maximum temperature (full line top) and minimum temperature (full line bottom); coloured shading indicates approximate 95% confidence intervals for the sampling error. Dashed: homogenized using only temperature and season as covariates. Dotted: data not needing homogenization. [Colour figure can be viewed at wileyonlinelibrary.com]

However, as shown in Figure 6, the sampling errors of the annual minima and maxima derived from the homogenized TN and TX, respectively, are still negligible compared to the large variability.

Figure 7 shows smooth nonlinear trend lines determined from the full homogenized time-series of annual minima and maxima over 1907–2022 (see de Valk, 2020, for the method). The uncertainty bands represent to the impact of year-to-year fluctuations on the trend lines. Also, similar trend lines determined from the non-homogenized data are shown for comparison. As expected, the impact of homogenization is substantial; the results are compatible with the earlier homogenization in Brandsma (2016a) and Brandsma (2016b).

4 | DISCUSSION AND CONCLUSIONS

The best performing models reduce the cross-validation *MSE* of the simplest models by only modest fractions: S_1 ranges from 0.06 to 0.69, S_2 from 0.06 to 0.45. This means that for most purposes, the use of covariates (even season) for homogenization is beneficial but not essential for the station relocations considered. However, the use of cross-validation for model selection ensures that the more complex models are not less robust.

Apart from y_{src}, the wind vector is the most important covariate. This makes sense, because larger wind speeds increase vertical mixing of air and advection of heat and thereby reducing spatial temperature gradients.

The estimated combined effect of wind speed and cloud cover on minimum temperature at Eelde, discussed in Section 3 (Figure 4), illustrates that without imposing physical constraints, the adopted statistical learning method is able to detect subtle but real effects. Therefore, the method offers good prospects also for homogenization in other settings, for example, without direct parallel measurements, and for other tasks involving the statistical modelling of complex relationships between weather variables.

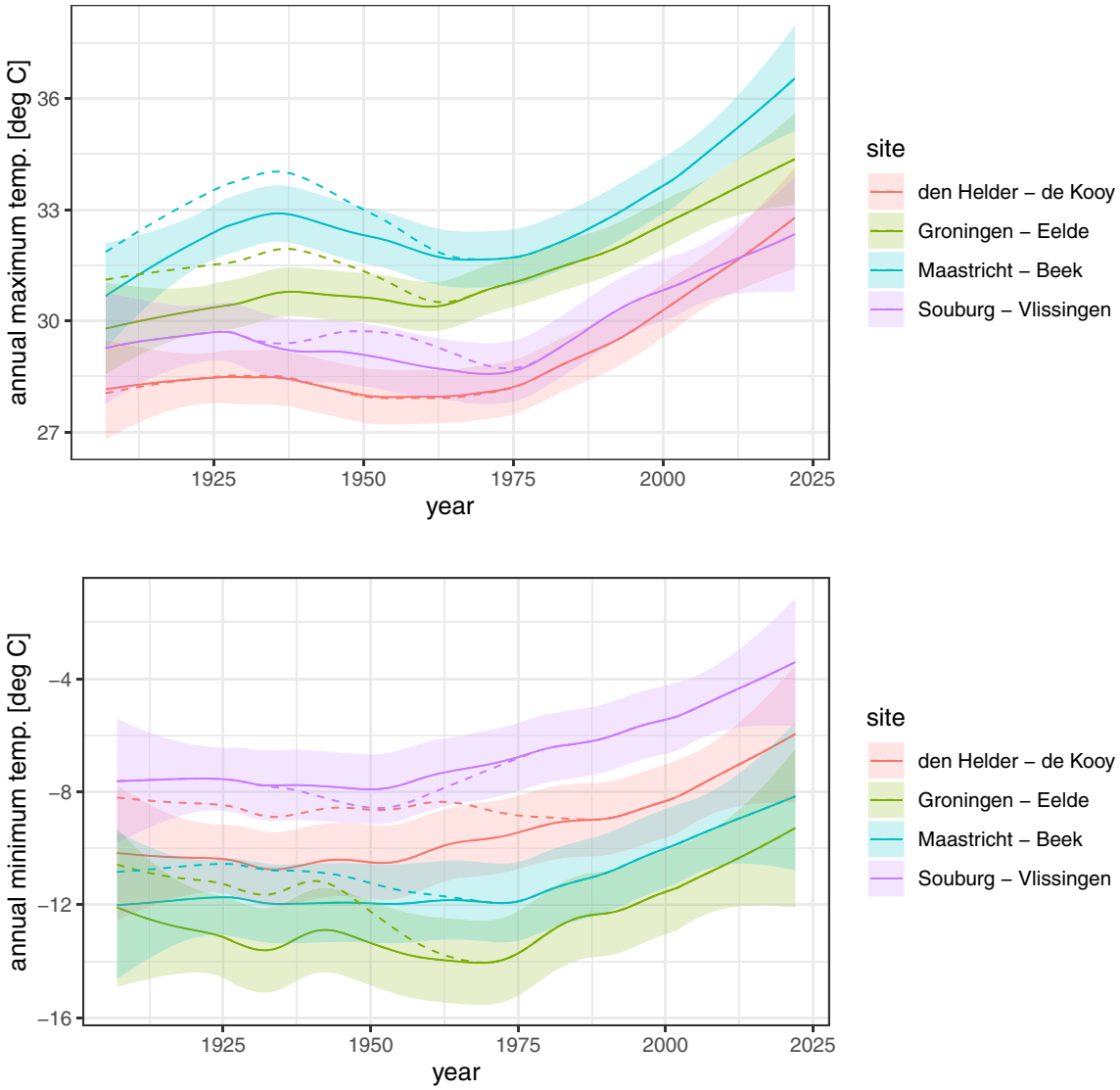

FIGURE 7 Nonlinear trends of homogenized annual maximum temperature (full lines top) and of homogenized minimum temperature (full line bottom); coloured shading indicates approximate 90% confidence intervals for the sampling error of the trendline. Dashed lines: same, but without homogenization. [Colour figure can be viewed at wileyonlinelibrary.com]

The absence of seasonality from almost all selected models is encouraging: as spatial differences in temperature are not causally related to the time of the year, a model based only on the values of physical quantities is to be preferred. All selected models outperform the model using only temperature and season (see S_2 statistics in the tables), so we may conclude that the seasonal dependence of the temperature difference between the sites does not need to be modelled explicitly if the right variables have been measured and included in the model.

The relatively modest improvements of the selected models with respect to simpler models indicates that we may not have the most relevant information present in our data sets. In particular, we needed to approximate the values of cloudiness and humidity at tTN and tTX as these were only measured three times a day (see Section 2.1). The effects of such temporal mismatch may be substantial. Even the hourly sampling may not be sufficient to fully resolve the physical relationships between covariates and the temperature differences between stations. In reality, the largest temperature differences occur for clear-sky and windless conditions (<1.0 m/s; Brandsma & Wolters, 2012). Especially accurate wind measurements at screen height are in practice not available for these conditions. Finally, daytime eddies and nighttime non-linear build up and break down of stable layers near the ground limit every homogenization of TN and TX.

For important climatological indicators such as annual extremes of TN and TX (see Figure 5), the effect of covariates is small in comparison to the natural year-to-year variation. The simpler models based on temperature and season only are already effective.

The differences between the selected models for TN and TX at the four stations reflect the different physical environments at these sites, but there is also considerable ambiguity in the model selection. For instance, in most cases, several models have almost the same skill as the selected model and they all produce very similar homogenized time-series. This is not a problem, because some covariates are correlated and therefore to some extent interchangeable. Furthermore, all these models are useful, as they are robustly estimated using cross-validation.

Several extensions of the method presented here are possible. First, the method may be adapted to correct inhomogeneities without having (direct) parallel measurements, for example, by fitting two models predicting the temperature before and after a known breakpoint having the same homogeneous temperature measurements from nearby weather stations as covariates (e.g., wind), and using the difference of the output of these models to perform the correction. Second, extending the approach to other variables such as daily rainfall, with heavily skewed distributions, is of interest. Finally, it may be of interest to embed the current method in a more general framework involving breakpoint detection in time-series from multiple sites.

Like all other homogenization methods, the present method cannot correct the effects of gradual changes in the sensor environment, which may result in ambiguity in the interpretation of observed local long-term changes.

AUTHOR CONTRIBUTIONS

Cees de Valk: Conceptualization; visualization; writing—original draft; methodology; investigation; software; validation; formal analysis; writing—review and editing. **Theo Brandsma**: Conceptualization; data curation; visualization; writing—original draft; investigation; writing—review and editing; validation.

FUNDING INFORMATION

KNMI project Klimatologische data-analyse.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

Cees de Valk https://orcid.org/0000-0001-8104-3210

REFERENCES

Auchmann, R. & Brönnimann, S. (2012) A physics-based correction model for homogenizing sub-daily temperature series. *Journal of Geophysical Research*, 117, D17119. Available from: https://doi.org/10.1029/2012JD018067

Brandsma, T. (2016a) Homogenisatie van dagelijkse temperaturen van de knmi hoofdstations. *Meteorologica*, 25, 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. & Van der Meulen, J. (2008) Thermometer screen intercomparison in De Bilt (the Netherlands)—Part II: description and modeling of mean temperature differences and extremes. *International Journal of Climatology*, 28, 389–400.

Brandsma, T. & Wolters, D. (2012) Measurement and statistical modeling of the urban heat Island of the city of Utrecht (The Netherlands). *Journal of Applied Meteorology and Climatology*, 51, 1046–1060.

Bürger, G. & Cubasch, U. (2005) Are multiproxy climate reconstructions robust? *Geophysical Research Letters*, 32, L23711.

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

Efron, B. & Tibshirani, R.J. (1994) *An introduction to the bootstrap*. New York: CRC Press.

Hastie, T. & Tibshirani, R. (1990) *Generalized additive models*. Monographs on statistics and applied probability, Chapman and Hall.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013) *An introduction to statistical learning*, Vol. 112. New York: Springer.

Maraun, D. (2013) Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. *Journal of Climate*, 26, 2137–2143.

Squintu, A., van der Schrier, G., Štěpánek, P. & Zahradníček, P. (2020) Comparison of homogenization methods for daily temperature series against an observation-based benchmark dataset. *Theoretical and Applied Climatology*, 140, 285–301.

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.

Venema, V.K.C., Mestre, O., Aguilar, E., Auer, I., Guijarro, J.A., Domonkos, P. et al. (2012) Benchmarking homogenization algorithms for monthly data. *Climate of the Past*, 8, 89–115.

Von Storch, H. (1999) On the use of "inflation" in statistical downscaling. *Journal of Climate*, 12, 3505–3506.

Wahba, G. (1990) *Spline models for observational data*. Philadelphia, PA: SIAM.

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, 307–339.

How to cite this article: de Valk, C., & 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. https://doi.org/10.1002/joc.8258

APPENDIX A

ESTIMATION DETAILS

The model (1) with (3) is estimated by numerical minimization of the function

$$\begin{aligned}&\mathcal{J}(s_0,s_1,.,.,.,s_d)\\&=\sum_{i=1}^{n}\left|\sum_{j=1}^{d}s_j\left(\mathbf{x}_j(t_i)\right)+s_0-y_{\mathrm{ref}}(t_i)\right|^2\\&+\sum_{j=1}^{d}\lambda_j P_j\left(s_j\right),\end{aligned}\tag{A1}$$

for the functions $s_0,s_1,...,s_d$ under the constraint that $\sum_{i=1}^{n}s_j\left(\mathbf{x}_j(t_i)\right)=0$ for $j=1,...,d$, so the intercept s_0 can be determined unambiguously. On the left, we see the sum of the squared deviations between the model predictions and the reference values. This sum is proportional to minus the logarithm of the likelihood of the data if $\epsilon(t)$ at successive t are Gaussian and independent and have the same variances and mean 0 (but independence and uniformity of the variances is not actually required for estimation; without these conditions, we still retain a valid pseudo-likelihood method).

Each term $\lambda_j P_j\left(s_j\right)$ of the sum on the right represents a so-called thin-plate spline penalty on the function s_j (Wood, 2020). It is the integral over the covariate space of a sum of squared higher-order partial derivatives of s_j Wahba, 1990. The positive factor λ_j determines the degree of smoothing to be applied to s_j; increasing λ_j gives a smoother s_j. The factors $\lambda_1,...,\lambda_d$ are jointly estimated by REML, a variant of marginal likelihood estimation (Wood, 2020).

APPENDIX B

VARIANCE MATCHING AS CONSTRAINED REGRESSION

Minimizing Equation (A1) under the additional constraint that the variance of the prediction $s_0+\sum_{j=1}^{d}s_j\left(\mathbf{x}_j\right)$ matches the variance $VAR\left(y_{ref}\right)$ of y_{ref} as in Equation (5) requires finding the real number ρ maximizing the minimum over $s_0,s_1,...,s_d$ of the Lagrangian

$$\begin{aligned}&\mathcal{L}(s_0,s_1,...,s_d,\rho)=\mathcal{J}(s_0,s_1,...,s_d)\\&+\rho\left(\sum_{i=1}^{n}\left|\sum_{j=1}^{d}s_j\left(\mathbf{x}_j(t_i)\right)\right|^2-n\mathrm{VAR}(y_{\mathrm{ref}})\right).\end{aligned}\tag{B1}$$

This should be done within the loop of estimating $\lambda_1,...,\lambda_d$ in Equation (A1). For practical reasons, we cannot do this. However, if we know how adding the constraint would affect the estimates of $\lambda_1,...,\lambda_d$, then we might be able to determine how the constraint affects the estimates of $s_0,s_1,...,s_d$. Note that the non-constant term

$$\rho\sum_{i=1}^{n}\left|\sum_{j=1}^{d}s_j\left(\mathbf{x}_j(t_i)\right)\right|^2$$

in Equation (B1) has the same form as the sum of squared residuals

$$\sum_{i=1}^{n}\left|y_{\mathrm{ref}}(t_i)-s_0-\sum_{j=1}^{d}s_j\left(\mathbf{x}_j(t_i)\right)\right|^2$$

in $\mathcal{J}$; see Equation (A1); only in this case, the 'observations' and intercept are all zero, and the term is weighted by ρ. This implies that it also needs to be compensated in the thin-plate spline penalty term $\sum_{j=1}^{d}\lambda_j P_j\left(s_j\right)$ in $\mathcal{J}$ (see Equation A1), by multiplying it by $1+\rho$. In fact, since $\rho<0$, the penalty is relaxed, as it should. The result is that the solution $s_1,...,s_d$ corresponds to a stationary point of the adjusted Lagrangian

$$\begin{aligned}\mathscr{L}(s_0,s_1,...,s_d,\rho)=&\sum_{i=1}^{n}\left|\sum_{j=1}^{d}s_j\left(\mathbf{x}_j(t_i)\right)+s_0-y_{ref}(t_i)\right|^2\\&+\sum_{j=1}^{d}(1+\rho)\lambda_j P_j\left(s_j\right)\\&+\rho\left(\sum_{i=1}^{n}\left|\sum_{j=1}^{d}s_j\left(\mathbf{x}_j(t_i)\right)\right|^2-nVAR\left(y_{ref}\right)\right).\end{aligned}\tag{B2}$$

Inserting Equation (4) for the functions $s_1,...,s_d$ and equating the partial derivatives of the resulting function to the spline coefficients $\widehat{\beta}_{k,j},k=1,...,N_j,j=1,...,d$ to zero can be shown to give the same spline coefficients as doing the same with $\mathcal{J}$ in Equation (A1), and then correcting the solution afterwards by dividing all spline coefficients by $1+\rho$. Therefore, the solution of the constraint problem is equivalent to variance matching with $\alpha=1/(1+\rho)>1$ in Equation (7).

APPENDIX C

ESTIMATION OF THE SAMPLING ERROR

A prediction with the fitted GAM model provides standard deviations of the model output $s(x_1(t),...,x_m(t))$ for every t. These represent the sampling variability in the predictions due to the limited size of the sample on which it is trained, assuming that disturbances are uncorrelated. However, the residuals of the fitted GAM models

are correlated, so the estimated standard deviations are biased.

For the m years with parallel measurements, we can also make another, crude, estimate of the standard deviations: from the cross-validation (Section 2.2), we have for every year Y with parallel measurements a model estimated from the data of all years except Y. From these estimates, we can use the delete-1-year jackknife method to estimate the variance of the estimate of a model parameter overall m years (see e.g., section 11.7 of Efron & Tibshirani, 1994). Although such estimates are crude, they are hardly affected by serial dependence. In the same manner, we can estimate the variance of the homogenized values $y_{hom}(t)$ for every day t by using the prediction $y_{hom}Y(t)$ from the model estimated from all data except the data from year Y:

$$\widehat{\sigma}^2_{jack}(y_{hom}(t)) = \frac{m-1}{m}\sum_{j=Y_1}^{Y_m}\left|y^j_{hom}(t) - y^0_{hom}(t)\right|^2,$$
$$y^0_{hom}(t) = \frac{1}{m}\sum_{j=Y_1}^{Y_m} y^j_{hom}(t). \tag{C1}$$

These variance estimates for individual instants of time t are likely too crude to be useful, but their mean or median over all days should give a reasonable estimate of the true mean or median variance. To make the distribution of the variance estimates from the GAM fit compatible with the distribution of the jackknife estimates, the former are first randomized by multiplying them by Chi-squared random variables with $m-1$ degrees of freedom. Then we compute the medians of $\widehat{\sigma}^2_{jack}(y_{hom}(t))$ and of the randomized $\widehat{\sigma}^2_{gam}(y_{hom}(t))$ over all days and compute their ratio. The GAM estimates of variance (not randomized) for all days are then scaled by this variance ratio. A practical advantage of this approach is that if the cross-validation *MSE* is large because of incompatibility of the data from different years, it automatically increases the estimates of sampling error variance.