Bencharking the performance of homogenization algorithms on synthetic daily temperature data

Rachel E. Killick\textsuperscript{1,2} | Ian T. Jolliffe\textsuperscript{2*} | Kate M. Willett\textsuperscript{1}

\textsuperscript{1}Met Office Hadley Centre, Exeter, UK
\textsuperscript{2}College of Engineering, Maths and Physical Sciences, University of Exeter, Exeter, UK

Correspondence
Rachel E. Killick, Met Office Hadley Centre, Met Office, FitzRoy Road, Exeter EX1 3PB, UK.
Email: rachel.killick@metoffice.gov.uk

Funding information
BEIS and Defra; Engineering and Physical Sciences Research Council, Grant/Award Number: 1199873

Abstract
This paper describes a new homogenization algorithm validation methodology, and its use to assess the skill of eight different algorithms, when applied to synthetic daily temperature time series. These algorithms were ACMANT, Climatol (in both daily and monthly configurations), DAP, HOM, MAC-D, MASH and SpliDHOM. Algorithms were tested on benchmark data replicating daily temperature variability in four regions in North America: Wyoming, the South East, the North East and the South West. These benchmarks contained plausible spatial and temporal correlation, differing station densities and both abrupt and gradual inhomogeneities. Algorithm ability was assessed both according to detection of inhomogeneities and correction of their effects, investigating bias, root-mean-square error (RMSE), linear trend recovery, station extremes and variability recovery. Inhomogeneities with a magnitude greater than 1°C were those most commonly detected with smaller inhomogeneities, and those that were not constant over time, proving harder to identify. Regional RMSE was always reduced by all algorithms and regional bias was reduced in over half of the region/scenario pairs. Trend recovery was variable, but the correct sign of regional trends was retained by all algorithms. Areas for future algorithm improvement include working with autocorrelated data and correcting moments higher than the mean. The data are available from https://www.metoffice.gov.uk/hadobs/benchmarks and the validation code from https://github.com/RachelKillick/Daily_benchmarks allowing the extension of this work and its application to new algorithms.

KEYWORDS
benchmarking, daily temperature, homogenization, synthetic data

1 | INTRODUCTION

To be homogeneous, a temperature time series should only contain variability arising from weather and climate (Conrad, 1946). However, in the real world a number of sources can cause spurious variability to enter a time series and corrupt its homogeneity. These changes can be as subtle as the grass height near the thermometer or as noticeable as a long-distance station relocation (Menne et al., 2009). The result of these changes can be a nonclimatic artefact, or inhomogeneity, of similar magnitude to the true climate signal,
confounding efforts to quantify the rate of temperature change (Williams et al., 2012).

Many homogenization algorithms exist that seek to identify and remove inhomogeneities from time series to make them suitable for climate change assessment, but quantifying their effectiveness is challenging when the truth being sought is unknown. For this reason, making use of known benchmarks is a valuable tool in algorithm assessment. Such benchmarks consist of time series, with known properties, such that the performance of the algorithms run on them, and the quality of the data returned, can be assessed.

Notable past benchmarking studies include the work of Venema et al. (2012) assessing 25 different algorithm contributions and Williams et al. (2012) assessing variants of a single algorithm. Both these studies worked with data at a monthly time resolution, whereas the current paper summarizes the results from the first study to assess multiple homogenization algorithms at the level of daily mean temperature data (Killick, 2016). Subsequent work by Squintu et al. (2020) has now also assessed variants of four algorithms on daily minimum and maximum temperature benchmarks.

A description of the data used in this study is in section 2.1, followed by an overview of the algorithms applied. An explanation of the validation measures used to assess algorithm performance is in section 2.3. Section 3 analyses the performance of the algorithms based on each validation measure, followed by summaries of the strengths and weaknesses of each algorithm. Conclusions are presented in section 4.

2 DATA AND METHODS

2.1 Benchmark data

The data used in this study were created using a Gamma Generalized Additive Model (GAM) with inputs from climate observations and downscaled reanalysis data used to produce outputs mimicking daily mean temperature time series from January 1970 to December 2011. Full details of the formulation and justification of this model as well as the strengths and weaknesses of the outputs are given by Killick et al. (2019).

The data mimic weather station measurements for four regions of the United States (Figure 1), namely Wyoming, a land-locked state with varying topography and distinct seasons; the South East, incorporating parts of multiple states with subtropical and tropical climates, sea borders and indistinct seasons; the North East, experiencing influence from the coast and a snow climate; and the

![Figure 1](wileyonlinelibrary.com)
South West, incorporating five different climatic regimes from deserts to mountains. Using a statistical model allowed the production of homogeneous data (clean data), to which known inhomogeneities were added to produce the released data. These inhomogeneities were created in two different ways; using constant offsets (on approximately 30% of occasions) and by perturbing the inputs to the GAM (approximately 70% of occasions). Inhomogeneities created by perturbing the GAM inputs are referred to as explanatory variable changes and create seasonally varying inhomogeneities affecting moments higher than the mean. Explanatory variable changes are typically smaller than constant offsets.

Four different scenarios were created to allow algorithm assessment under different circumstances. These four scenarios incorporated a best guess (for the real world), increased station density, step changes only and increased autocorrelations. The contents of these scenarios are summarized in Table 1 with further details of their characteristics given by Killick et al. (2019). Clean data and released data are available from https://www.metoffice.gov.uk/hadobs/benchmarks.

2.2 Homogenization algorithms

Eight algorithms available at the time of this work (2014/2015) were run on the released data. The validation principles, benchmarking principles and data described below are designed to be re-used. All algorithms run in this study can be considered to be automated and their references can be found in Table 2. Where specific settings were provided by the homogenisors, they are specified in the Supporting Information.

2.2.1 MAC-D

MAC-D searches for Multiple Abrupt Changes in the mean value of Daily time series. MAC-D was applied by deseasonalizing the time series and splitting the station network into subsets of stations whose deseasonalized daily time series are highly correlated with the regional time series of that subset (calculated as the median of the deseasonalized series in that region). Stations then have the climatic signal removed (weighted by the correlation of their deseasonalized data with this signal) and their seasonal cycle removed. These anomaly series will still traditionally be autocorrelated, so they undergo a calibration process of a linear filter and inhomogeneity detection and removal until they exhibit no autocorrelation at lags one or two and no remaining inhomogeneities.

The inhomogeneities are sought using the Standard Normal Homogenization Composite Method (Rienzner and Gandolfi, 2011). Once this iterative process is complete, pairwise comparisons are used to verify inhomogeneity locations, and those not exhibited in a user-defined proportion of pairs (usually 0.5) are disregarded. As MAC-D is designed primarily as a detection algorithm, inhomogeneity adjustments were calculated separately by the creator as constant offsets.

2.2.2 Climatol

Climatol is a climate analysis package, containing automatic homogenization procedures, designed for the programming language R. Climatol searches for step change inhomogeneities by iteratively applying the Standard Normal Homogeneity Test (SNHT) (Alexandersson, 1986) and splitting the series at the inhomogeneities it detects. For the detection process, target series are compared to composite reference series.

Once all the time series have been split sufficiently that every segment is deemed homogeneous, adjustments are made. These are made by either adjusting the split data of the same series or, if there were missing data, using a weighted average of the closest available data at each time step. Adjustments were constant offsets at the daily scale and variable at the monthly

### Table 1 Summary of the characteristics of each test scenario

| Scenario                  | Focus regions | Number of stations relative to real world | Step inhomogeneities | Gradual inhomogeneities | Altered autocorrelations |
|---------------------------|---------------|------------------------------------------|----------------------|-------------------------|--------------------------|
| Best guess                | All           | Same                                     | Yes                  | Yes                     | No                       |
| Increased station density | All           | Increased                                | Yes                  | Yes                     | No                       |
| Step changes only         | All           | Increased                                | Yes                  | No                      | No                       |
| Increased autocorrelations| Wyoming       | Same                                     | Yes                  | Yes                     | Yes                      |
scale. Monthly adjustments were assigned to the 15th of each month and linearly interpolated to the daily data. In more recent versions of Climatol, monthly homogenization for daily data is only used to provide change point dates which are then used to split the daily series and adjust them independently of the monthly corrections.

### 2.2.3 | MASH

The Multiple Analysis of Series for Homogenization (MASH) method can be run interactively or automatically on daily data. This algorithm aggregates the daily series to monthly. These monthly series are then compared and adjusted in an iterative process until no more inhomogeneities are found. Each series in turn is considered the candidate series and difference series between this candidate and the weighted reference series are used to identify inhomogeneity location confidence intervals and magnitudes. MASH can search for multiple inhomogeneities at once during this iterative process and those inhomogeneities common to all difference series can be identified as belonging to the candidate series. The difference between the average in the reference series and the candidate series is the inhomogeneity magnitude. Adjustments are made by smoothing the monthly corrections to the daily level. MASH also quality controls and infills the daily data.

### 2.2.4 | ACMANT

ACMANT is the Adapted Caussinus-Mestre Algorithm for Homogenizing Networks of Climatic Time series (Domonkos, 2011; Domonkos and Coll, 2017). ACMANT searches only for step change inhomogeneities, identified using optimal step function fitting and the Caussinus–Lyazrhi criterion (Caussinus and Lyazhri, 1997). Once the inhomogeneities have been located their magnitudes are calculated by the ANOVA correction model (Caussinus and Mestre, 2004; Lindau and Venema, 2018). In this model adjustments are constant.

The applied subversion differs from ACMANTv2 in some details of the reference series creation. Reference series were created using the weighted averages of time series from surrounding stations. The weights were based on spatial ordinary kriging with some modifications.

Since the benchmarking study reported here, ACMANTv3 (Domonkos and Coll, 2017) and ACMANTv4 (https://github.com/dpeterfree/ACMANT) have been released; both performed favourably in the recent comparison study of Domonkos et al. (2021).

### 2.2.5 | HOM, SpliDHOM and DAP

A common inhomogeneity detection method was applied for the Higher Order Moments (HOM), Spline Daily
Homogenization (SplIDHOM) and Distribution Adjusted by Percentiles (DAP) methods. This detection method sought step change inhomogeneities at the annual, seasonal and monthly levels using three different detection tests at each level: the SNHT, the Maronna and Yohai bivariate test (Potter, 1981) and the Easterling and Peterson test (Easterling and Peterson, 1995). If a proportion of these tests agreed (this proportion is dependent on the presence or absence of metadata) then an inhomogeneity was deemed to have taken place. It was recommended that a gap of at least 4 or 5 years between change points was necessary for these algorithms to perform at their best, although a daily implementation could in theory cope with gaps as small as half a month. For these methods a single reference series was used, selected to be the neighbour most highly correlated with the target station. DAP has since been adapted to work with a composite reference series from five neighbours (Squintu et al., 2020).

The adjustment methods of these algorithms all start with the target series split into homogeneous subperiods according to the inhomogeneities identified in the detection step. Each period is homogenized in turn, beginning with the period before the most recent inhomogeneity. The process continues iteratively until the series is deemed homogeneous. All three methods adjust at the percentile level, thus correcting higher order moments of the distribution.

For HOM, a nonlinear local regression model, LOESS, is used to identify the relationship between the target and reference station before and after the focus inhomogeneity. This model is used to predict the temperatures at the target station after the inhomogeneity. These predictions are then compared to the observations and the differences between them are binned into the deciles of the probability distribution of the observed data after the inhomogeneity. A smoothed function is then fitted to the deciles to obtain percentile differences for each point. The points before the inhomogeneity are also assigned to percentiles of their probability distribution and each point is then corrected by the defined percentile amount.

For SplIDHOM, a cubic spline approach is used to determine the relationship between the target and reference station before and after the inhomogeneity; correction then proceeds in a similar manner to HOM.

For DAP, the percentiles of the target and reference series, before and after the inhomogeneity, are estimated empirically. Differences in the percentiles are obtained between the target and the reference series in these two time periods, creating two difference series. The difference between these series is then smoothed to provide the adjustments. These steps are illustrated in Stepanek et al. (2013, fig. 5). This process is applied iteratively.

### 2.3 | Homogenization algorithm assessment

To be homogenized, a time series must have change point locations detected and any resulting inhomogeneous periods adjusted. Assessing detection ability is important for identifying the type of change points algorithms may be prone to detecting or missing in different scenarios. Assessing adjustment ability allows the quantification of improvements made to time series. A range of assessment measures were implemented as different aspects will be of interest to different users.

#### 2.3.1 | Change point detection ability

The first step in homogenization is to identify the location of change points. Any change point can either be found (hit) or not (miss). Likewise, any location where there is not a change point can be identified as clean (correct rejection) or have a change point “identified” (false alarm). These four possibilities are often summarized in a contingency table, as in Table 3.

The measures used to assess detection ability of the algorithms were the hit rate $H$, $H = \frac{a}{a+c} \in [0,1]$, (1) where $a$ is the number of hits and $c$ is the number of misses, and the false alarm rate (FAR), $\text{FAR} = \frac{b}{b+d} \in [0,1]$, (2) where $b$ is the number of false alarms and $d$ is the number of correct rejections (allowing for the caveats explained in the following paragraphs).

Some adaptation of these measures, using a windowing approach, was necessary to ensure that $d$ was not orders of magnitude larger than the other three quantities when working with daily data.

In the windowing approach a series is divided into homogeneous sub-period windows (HSPs) and change point windows (CPs). An HSP window can either be a false alarm or a correct rejection, while a CP window can either be a hit or a miss. With such a classification the magnitude of all four quantities is similar. In order to keep classification as simple as possible, each window
was only deemed to contain one hit, miss, false alarm or correct rejection. That is, if multiple false alarms were assigned in a single HSP it still only increased by one.

The change point window length was 180 days, allowing a detection to be counted as a hit in the 90 days either side of the true change point location. HSPs were of variable length as they spanned the time between two CPs, or between the end of the series and a change point window.

As the values of \(a\), \(b\), \(c\) and \(d\) are small for individual stations the detection ability results will always be presented for regions as a whole.

2.3.2 | Adjustment ability

The adjustment ability of an algorithm is its ability to provide data closer to the clean data on return than it was on release. Adjustment ability was assessed in five areas: bias, root-mean-square error (RMSE), linear trend recovery, variability and extremes.

The measure of percentage recovery (PR) was used to assess bias, RMSE and trend recovery (Willett et al., 2014; Killick, 2016). PR allows the percentage improvement in a property to be seen and is calculated as

\[
PR = \left( \frac{\text{property}_{\text{released}} - \text{property}_{\text{returned}}}{\text{property}_{\text{released}} - \text{property}_{\text{clean}}} \right) \times 100. \quad (3)
\]

The important range of values for PR is from 0 (no improvement in the returned series compared to the released series) to 100 (the returned series reproduces the clean series). For some quantities, such as bias and trend, a change can be made too far in the right direction and PR can exceed 100, with values greater than 200 indicating deteriorating performance. A negative PR indicates that the homogenization process has adjusted a series in the wrong direction, for example, it has increased the bias. A variation on PR where values greater than 100 are subtracted from 200 to allow comparisons between overcorrecting and undercorrecting is considered in the Supporting Information.

PR should not be considered in isolation. For example, two stations with no change in their RMSE between released and returned series would have the same (zero) value of PR, even though their RMSEs may be very different. This situation can be addressed by providing value recovery plots, showing the size of the quantity being assessed, alongside PR plots, as illustrated in section 3.2. At the station level it was not practical to include value recovery plots for all regions, scenarios and algorithms in this paper. Therefore, although conclusions about the number of stations improved, unchanged and made worse are included as guidance for some measures, further investigation on the impacts of accepting some small changes to a station without penalization would be beneficial future work. Detailed station breakdowns and plots were provided to the algorithm developers and can be found at https://www.metoffice.gov.uk/hadobs/benchmarks/.

Bias in the mean

This was calculated on a station by station basis as the difference in station means between clean and returned time series. Ideally, an algorithm would adjust the series such that the bias between the released and clean data was completely removed. To get a regional mean value of the bias, the mean was taken across the station bias values. Bias is used as an adjustment ability measure as it is important to know if an algorithm is prone to making stations consistently warmer or cooler (Killick, 2016). If a station bias was smaller than 0.05°C, then its time series is referred to as minimally biased.

Root-mean-square error

RMSE was calculated on a station by station basis, but the regional RMSE is the focus here. To create a regional RMSE, all the time series in a region were joined end to end as one long series and the RMSE of this series was calculated as

\[
\sqrt{\frac{1}{N} \sum_{i=1}^{N} (\text{clean}_i - \text{returned}_i)^2}, \quad (4)
\]
where \( N \) is the total number of points in the region, \( \text{clean}_i \) are the clean data and \( \text{returned}_i \) are the returned data.

RMSE was used as an adjustment ability measure in addition to bias as it can be shown to incorporate mean, standard deviation and correlation components (Murphy, 1988).

**Linear trends**
Linear trends were assessed both because they are fundamental measures of climate change and because trends due to inhomogeneities can be similar in magnitude to true climatic trends (Williams et al., 2012). Linear trends were calculated using least squares regression on deseasonalized data that had been aggregated to the annual level in order to reduce autocorrelation. Focus was on both station trends and regional trends. Regional trends were calculated by first averaging each day across all stations to get a regional time series and then aggregating to the yearly level for the trend calculation.

**Variability**
Maintaining the true variance of a dataset is important for studying extremes. The ratios of clean to released and clean to returned time series’ standard deviations were calculated in order to examine if the homogenization process had recovered the variability in the clean data or whether it made the time series too uniform or too variable. This allowed assessment of whether more than just the mean of a series had been addressed by the homogenization procedure, which is important in daily data where inhomogeneities are known to affect higher order moments (Della-Marta and Wanner, 2006). The explanatory variable inhomogeneities were created to affect the variability as well as the mean in order to create the test bed for this part of the assessment. This measure was assessed on a station by station basis.

Given that this variability assessment was carried out on non-deseasonalized data it will not be capturing all aspects of variability. The variability in the seasonal cycle will dominate that found day-to-day. Looking instead at the standard deviation of anomaly series would reduce this dominance, but not eliminate it, as the day-to-day variability also varies across the year. If day-to-day variability with influence from the seasonal cycle minimized is to be assessed the authors recommend analysing the anomaly standard deviations separately in “summer” and “winter,” but this was outside the scope of the present study.

**Extremes**
Extreme value recovery is relevant as extremes are often where the effects of climate change are felt (Gross et al., 2018). Algorithms had to return the correct clean extreme value on the correct day at the correct station for it to be counted as a success, as smoothing one extreme while adding another is not desirable behaviour. This measure was assessed at the station level.

When comparing extreme values on like for like days, a measurement uncertainty value of ±0.14°C was allowed, following Brohan et al. (2006). These comparisons focused on single events, so measurement uncertainty is important to include.

3 | RESULTS AND DISCUSSION
The majority of the algorithms described in section 2.2 were applied to all regions and scenarios. The exceptions were MAC-D, which was run only on Wyoming, and DAP, HOM and SplIDHOM which were not run on the increased station density or step changes only scenarios in the North East. Climatol-monthly adjustment ability was not assessed for the Wyoming best guess or increased station density scenario as the authors were made aware of a mistake made in the application of the adjustments. MASH did not have its detection ability assessed because of its provision of change point locations as intervals instead of single locations.

All participants were asked to homogenize to the most recent time period, acknowledging that this should be the most reliable. Only the inhomogeneous data were provided to the participants, thereby ensuring a blind test. The clean data have since been released and further investigation into the performance of these and other algorithms is encouraged.

3.1 | Detection ability assessment
In almost every case the HR, was greater than the FAR (Figure 2). This illustrates that the algorithms can detect change points in a varied set of scenarios, but that all algorithms miss some change points and falsely detect others.

For all algorithms and regions, apart from Climatol-Monthly in Wyoming, the HR was highest in the step changes only scenario, suggesting that this was the easiest scenario in which to identify inhomogeneities. This is logical as, with detection based on comparisons to nearby stations, sudden changes are easier to identify than gradual changes. Gradual changes may start at different locations relative to different neighbouring stations owing to their different local environments.

The lowest HR was almost always found in the increased autocorrelation scenario in Wyoming. DAP,
HOM and SpliDHOM (DHS) instead showed its lowest HR in the best guess scenario in the North East, where it also showed its highest FAR. Climatol-Daily and ACMANT also showed their highest FARs in best guess scenarios (the South West and South East, respectively). MAC-D’s highest FAR was paired with the low HR in the Wyoming increased autocorrelation scenario while Climatol-Monthly’s was in the increased station density scenario of the South West. This shows little pattern in FARs and demonstrates that different algorithms have different strengths and weaknesses when detecting inhomogeneities.

Comparing the four algorithms that were run on all four geographical regions, each one showed its highest HR in a different region: South West for Climatol-Daily (0.38), Wyoming for Climatol-Monthly (0.44), North East for ACMANT (0.47) and South East for DHS (0.20). This suggests that there is sufficient variation in the created data to evaluate algorithms over a range of network and climate types.

Constant offset inhomogeneities were consistently better detected than explanatory variable inhomogeneities, with a maximum of 79% of constant offset inhomogeneities being found (by ACMANT in the North East step changes only scenario). A maximum of 35% of explanatory variable inhomogeneities were found (by ACMANT in the North East increased station density scenario). ACMANT was always best at detecting explanatory variable changes, but for constant offset changes Climatol-Monthly and Climatol-Daily were also high performers (Table S1, Supporting Information).

There were three magnitude classifications of inhomogeneities: large (>1°C), medium (0.2–1°C) and small (≤0.2°C). The greatest proportion of inhomogeneities found were large, with Climatol-Daily locating 100% of these inhomogeneities in both the increased station density and the step changes only scenario of the South East. This compares to a maximum of 79% of the medium inhomogeneities being found (by Climatol-Daily in the step changes only scenario of the South East) and a maximum of 25% of the small inhomogeneities being found (by ACMANT in the best guess scenario of the South East) (Table S2).

In summary, ACMANT was the best algorithm at detecting medium, small, and explanatory variable inhomogeneities, but also had the highest FAR in all scenarios apart from the Wyoming increased autocorrelations scenario. Climatol-Daily was best at finding large inhomogeneities.

### 3.2 | Adjustment ability assessment

Assessing algorithm adjustment ability was done with the caveat that lower HRs or higher FARs are likely to be accompanied by less favourable adjustment ability as adjustments are being based on incorrect inhomogeneity locations. It should also be noted that all returned data were masked to be of the same level of completeness as the released data, to ensure fair comparisons between contributions where some had infilled missing data and others had not.
A further caveat when assessing adjustment ability is that MASH did not always homogenize to the most recent time period. If the algorithm deemed that this period was not homogeneous, or if there was only one inhomogeneity and it was close to the end of the series then MASH would choose a different reference period. This affected fewer than 10% of stations and therefore this differing reference period was not considered in the adjustment assessment. Although these stations returned by MASH will compare less well with the clean data in many of the statistics in this study, there will be circumstances in the real world where not correcting to the most recent period is the best course of action.

All adjustment ability measures were calculated from data rounded to the precision of the released data (0.1°C), but the measures themselves are shown at higher precision where appropriate.

### 3.2.1 Bias and RMSE

In the released data, four of the region/scenario pairs were positively biased relative to the clean data, by at most 0.04°C in the Wyoming increased autocorrelation scenario. The remaining nine were negatively biased, by at most −0.10°C in the North East best guess scenario (Table 4). All algorithms achieved reductions in the biases for the majority of region/scenario pairs. However, all algorithms apart from Climatol-Monthly changed the sign of the bias in at least one region, indicated by a PR greater than 100 (Table 4). Predominantly this was changing a positive released bias to a negative returned bias. In the best guess in the South West, Climatol-Daily, ACMANT and MASH all changed the sign and increased the magnitude of the regional bias. ACMANT also did this in the increased autocorrelation scenario for Wyoming. The South West step changes only scenario started with a regional bias of −0.003°C which all algorithms increased in magnitude, with Climatol-Daily, MASH, DAP, HOM and SpliDHOM all changing its sign as well.

At the station level, most region/scenario pairs contained a greater number of negatively biased stations than positively biased stations in the released data (Table 5). ACMANT returned more negatively biased stations than positively biased stations in all apart from the South East step changes only scenario. Climatol-Daily returned the most minimally biased stations in over half of the region/scenario pairs, with ACMANT, MASH and Climatol-Monthly being the other algorithms to return most minimally biased stations. In the increased autocorrelation scenario, DAP, HOM and SpliDHOM returned fewer minimally biased stations than were present in the released data. However, they were also the only algorithms not to bias the two stations that were completely biased for the majority of region/scenario pairs. However, all algorithms apart from Climatol-Monthly changed the sign of the bias in at least one region, indicated by a PR greater than 100 (Table 4). Predominantly this was changing a positive released bias to a negative returned bias. In the best guess in the South West, Climatol-Daily, ACMANT and MASH all changed the sign and increased the magnitude of the regional bias. ACMANT also did this in the increased autocorrelation scenario for Wyoming. The South West step changes only scenario started with a regional bias of −0.003°C which all algorithms increased in magnitude, with Climatol-Daily, MASH, DAP, HOM and SpliDHOM all changing its sign as well.

### TABLE 4 Regional bias Percentage Recovery (PR).

| Region | Scenario          | Released bias | PR          |
|--------|-------------------|---------------|-------------|
| WY     | Best guess        | 0.05          | 130 166 NA 123 170 −26 −17 −20 |
|        | Increased density | −0.07         | 24 21 NA 33 33 5 5 8 |
|        | Step changes only | −0.08         | 34 21 68 83 68 12 20 21 |
|        | Increased autocorrelation | 0.04         | −5 100 49 46 207 44 49 37 |
| SE     | Best guess        | −0.08         | NA 49 31 94 49 −8 −17 −9 |
|        | Increased density | −0.02         | NA −4 −38 −17 0 25 21 29 |
|        | Step changes only | 0.01          | NA 54 8 8 62 138 154 138 |
| NE     | Best guess        | −0.10         | NA 58 67 43 47 28 0 29 |
|        | Increased density | −0.08         | NA 67 65 47 51 NA NA NA |
|        | Step changes only | −0.02         | NA 96 71 104 83 NA NA NA |
| SW     | Best guess        | 0.02          | NA 220 73 213 433 27 47 33 |
|        | Increased density | −0.06         | NA 56 7 13 0 5 −3 7 |
|        | Step changes only | −0.003        | NA 1,000 −133 633 −567 500 333 500 |
| Median PR over all region/scenario pairs | NA | 29 58 49 47 51 25 20 29 |

Note: A PR > 100 indicates the sign of the bias has been changed. Changing the sign and increasing the magnitude of the bias is a PR > 200. Shaded cells indicate a reduced regional bias. The median PR over all region/scenario pairs adjusted is also shown. Column headings for algorithms are as follows: MACD = MAC-D, CD = Climatol-Daily, CM = Climatol-Monthly, MASH = MASH, AC = ACMANT, D = DAP, H = HOM, S = SpliDHOM.
### Table 5  Station biases in the released data and each of the returned contributions

| Region/scenario          | Released | MACD | CD | CM | MASH | AC | D | H | S |
|--------------------------|----------|------|----|----|------|----|---|---|---|
|                          | P        | N    | M  | P  | N    | M  | P | N | M |
| WY best guess            | 30       | 27   | 18 | 25 | 23   | 27 | 19| 27| 29|
| WY increased density     | 55       | 67   | 36 | 33 | 64   | 61 | 27| 61| 70|
| WY step change only      | 51       | 77   | 30 | 26 | 69   | 63 | 18| 72| 68|
| WY increased autocorrelation | 31 | 32  | 12 | 39 | 21   | 15 | 30| 27| 18|
| SE best guess            | 49       | 60   | 44 | 12 | 36   | 105| 12| 61| 80|
| SE increased density     | 65       | 78   | 67 | 12 | 42   | 156| 9 | 50| 151|
| SE step change only      | 75       | 55   | 80 | 19 | 11   | 180| 19| 17| 174|
| NE best guess            | 48       | 73   | 25 | 16 | 42   | 88 | 19| 40| 87|
| NE increased density     | 67       | 95   | 45 | 30 | 65   | 112| 13| 46| 148|
| NE step change only      | 77       | 76   | 54 | 32 | 31   | 134| 22| 28| 157|
| SW best guess            | 49       | 49   | 53 | 20 | 33   | 98 | 24| 40| 87|
| SW increased density     | 59       | 98   | 65 | 20 | 53   | 149| 34| 89| 99|
| SW step change only      | 77       | 80   | 65 | 50 | 17   | 155| 65| 45| 112|
| Total                    | 594      | 1359 |     | 1329 |     | 1252 |       |    |   |

**Note:** Here minimally biased is used for any station with a bias of less than 0.05 °C. Total number of minimally biased stations is provided only for algorithms that homogenized all region/scenario pairs to ensure fair comparisons. Pale grey indicates a greater number of minimally biased stations on return than release, with dark grey indicating which algorithm returned the greatest number of minimally biased stations in each region/scenario pair. Column headings for algorithms are as in Table 4.

Abbreviations: P, positively biased; N, negatively biased; M, minimally biased.
clean in this scenario. It should be noted that the timing of the inhomogeneities will affect the returned biases. If a large inhomogeneity is missed near the beginning of the series, only a small proportion of the data will be affected; if it is missed near the end then the majority of the series will remain biased. Whether algorithms were better at detecting inhomogeneities in specific sections of the time series was not assessed in this study.

Climatol-Daily is among the algorithms making fewest station biases worse, although it is also often near the top for returning stations unchanged (Table S3). MASH is always the top algorithm for reducing station biases, although it is also almost always top for making station biases worse, likely as it rarely leaves any stations unchanged. MASH is the top performing algorithm in the increased autocorrelation scenario, making 55 stations better, and only 20 stations worse. ACMANT predominantly comes second to MASH in number of stations improved, number of stations made worse, and number of stations left unchanged.

Regional RMSE was reduced by all algorithms in all region/scenario pairs. Climatol-Daily is consistently among the top algorithms for overall improvement (Figure 3). In three of the step changes only scenarios, Climatol-Daily produces the greatest reduction in RMSE and it lags MASH and ACMANT by less than 0.005°C in the North East step changes only scenario. Climatol-Daily also shows the greatest RMSE reduction in the increased autocorrelation scenario. Where Climatol-Daily did not take the lead in RMSE reduction, this was predominantly taken by ACMANT although Climatol-Monthly was best in the increased station density scenario of the North East. DAP, HOM and SpliDHOM reduced the RMSEs by the smallest amount.

3.2.2 | Linear trend recovery

Looking at the observational data from the Global Historical Climatology Network Daily (GHCND) database

**Figure 3** (a) Value recovery in °C and (b) PR plots for regional RMSE relative to the clean data. X-axis labels are as follows: 1–4 = Wyoming scenarios (best guess, increased density, step changes only, increased autocorrelation); 5–7 = South East scenarios (best guess, increased density, step changes only); 8–10 = North East scenarios (best guess, increased density, step changes only); 11–13 = South West (best guess, increased density, step changes only). Black crosses represent clean data (always an RMSE of 0°C) and blue/grey crosses represent the released data relative to the clean data. MAC-D is represented by circles, Climatol-Daily by downward triangles, Climatol-Monthly by upward triangles, MASH by triangles pointing right, ACMANT by triangles pointing left, DAP by squares, HOM by addition signs and SpliDHOM by diamonds [Colour figure can be viewed at wileyonlinelibrary.com]
(Menne et al., 2012), which were the inputs to the GAM used to create the clean data, both the North East and South West regions showed significant linear trends over 1970–2011 at the 5% level. When the clean data were created, the North East still showed significant positive trends, but none of the other regions did. On adding inhomogeneities all the North East scenarios kept significant trends, and the South West increased station density scenario also showed a significant positive trend.

Algorithms returned improved regional trends, closer to those in the clean data, in most cases. However, the increased station density scenario of Wyoming saw the majority of algorithms make the regional trend worse and in the South West best guess scenario ACMANT overshoot the true trend sufficiently to make the returned trend more dissimilar to the clean trend than the released trend had been.

All regional trends that were significant in the released data were also significant, and of the same sign, for the returned data (Figure 4). No regional trends were made significant by any of the algorithms’ adjustments. No regional trends were negative in the clean, released or returned data. There were also always more positive than negative trends at the station level for the clean, released and returned data. The number of negative station trends was always correctly reduced from release to return. All algorithms improved the spatial coherence of the trends in all region/scenario pairs, correctly reducing the standard deviation among station trends in released compared to returned data.

ACMANT, MASH, Climatol-Monthly and Climatol-Daily all have instances of attaining a near 100% PR for the regional trend (Figure 4b). However, the best PR in the South West, the most climatologically varied region, was a 78% trend recovery (Climatol-Monthly, step changes only). ACMANT and Climatol-Daily exhibit the greatest tendency to overshoot and move the trend too far in the right direction, but this is true for less than half the region/scenario pairs. Only in the increased station density scenario of the South East (and once in the best guess for the South West) do algorithms adjust so far in the right direction that they make the trend more dissimilar to the clean data on return than release. However, in the case of the South East increased station density scenario, the linear trends in the clean and released data were almost identical, leading to percentage recoveries in Figure 4b off the scale, as a change in trend of −0.002°C resulted in a PR of −800%. This illustrates why it is important to look at values of trends (Figure 4a) and not just the percentage recoveries (Figure 4b). There was no

![Figure 4](wileyonlinelibrary.com)
best algorithm when it came to regional trend recovery; Climatol-Daily, Climatol-Monthly and MASH all had instances of showing the greatest trend improvement. Of these, MASH was top most frequently and was never the worst for trend recovery in any region/scenario pairs.

3.2.3 | Variability and extremes assessment

More stations had their variability increased than decreased by the addition of inhomogeneities. Algorithm performance in terms of variability is mixed. Generally, MAC-D, Climatol-Daily, Climatol-Monthly and ACMANT improved the variability more often than they worsened it. DAP, HOM and SpliDHOM tended to make more station variabilities worse and MASH improved the station variabilities in the majority of North East and South West scenarios, while making more worse in Wyoming and showing no consistent pattern in the South East. This highlights the fact that adjusting for changes in variance caused by inhomogeneities is difficult.

Looking at where algorithms increased station variabilities, DAP, HOM and SpliDHOM always degraded more stations than they improved, in terms of the statistics assessed. MASH behaved the same apart from in the best guess and increased density scenario of the South West, as did ACMANT apart from for the best guess scenario in the North East. In contrast, variability increases made by Climatol-Daily were predominantly for the better in the increased density and increased autocorrelation scenarios for Wyoming and all three North East and South West scenarios. Climatol-Monthly did the same for all the North East scenarios and all bar the increased station density scenario in the South West (Figures 5 and S1–S3).

Where algorithms decreased the variability of stations from released to returned data this was for the better in all scenarios for MAC-D, ACMANT, MASH, Climatol-Daily and Climatol-Monthly. DAP, HOM and SpliDHOM returned improved stations in Wyoming, the best guess scenario in the South East and the step changes only scenario in the South East for DAP and SpliDHOM. All three also made more improvements than degradations in the step changes only scenario for the South West. DAP, HOM and SpliDHOM were the only algorithms designed to homogenize higher order moments, but in part owing to more inhomogeneities being missed, this did not appear to improve their skill relative to other algorithms.

Looking at the retention and recovery of extreme values there is no algorithm that always performs best (Table 6). However, all algorithms return more extremes exact to within measurement precision than were in the released data in the majority of the scenarios. This finding is positive as the retention and recovery of extremes is crucial for reliable long term climate analysis.

Climatol-Daily, MASH and ACMANT always return more cold extremes exact to within measurement precision than were in the released data. Climatol-Daily also returns more hot extremes exact to measurement precision in all scenarios apart from the increased autocorrelations scenario of Wyoming.

HOM loses one hot extreme relative to the released data in the best guess in Wyoming, while DAP and SpliDHOM lose a maximum of two extremes. MASH loses the hot extremes in the South West best guess and increased station density scenarios as well as the step changes only scenario in Wyoming. In the increased autocorrelations scenario in Wyoming, MASH moves 57 hot extremes further from the truth, while the other algorithms corrupt a maximum of 20. ACMANT also returns fewer matched hot extremes for the best guess and increased station density scenarios of the South West than were found in the released data. Climatol-Monthly does the same, but only for the increased station density scenario.

3.3 | Algorithm performance summaries

A ranking of algorithm detection and adjustment performance according to measures discussed in this paper is provided in Table 7. It highlights that different algorithms have different strengths and weaknesses and is designed as an overview of this paper’s findings as are the following summaries.

3.3.1 | MAC-D

Run only in Wyoming, MAC-D shows better HRs for large inhomogeneities than all other algorithms apart from Climatol-Daily, but displays higher FARs than Climatol-Daily. MAC-D shows a lower tendency to leave station biases unchanged than Climatol-Daily, DAP, HOM and SpliDHOM, resulting in it having more improved, but also more made worse. Its regional performance was lower than Climatol, ACMANT and MASH in terms of bias, RMSE and trend PR, but there was a smaller sample size for MAC-D.

MAC-D’s performance was degraded in the increased autocorrelation scenario, suggesting working with such data could be an area for development. Other changes in scenario characteristics in Wyoming did not show a consistent impact on MAC-D’s performance.
### Table 6

Number of stations with hot (H) and cold (C) extremes exact to measurement precision.

| Region/scenario     | Released H | Released C | MACD H | MACD C | CD H | CD C | CM H | CM C | MASH H | MASH C | AC H | AC C | D H | D C | H H | S H |
|---------------------|------------|------------|--------|--------|------|------|------|------|--------|--------|------|------|-----|----|-----|-----|
| WY best guess       | 56         | 26         | 59     | 42     | 62   | 44   | 56   | 38   | 62     | 48     | 57   | 25   | 56  | 24 |
| WY increased density| 113        | 69         | 126    | 91     | 127  | 101  | 129  | 87   | 140    | 121    | 113  | 70   | 113 | 72 |
| WY step change only | 125        | 68         | 131    | 89     | 132  | 100  | 138  | 99   | 124    | 101    | 139  | 121  | 126 | 72 |
| WY increased autocorrelation | 63 | 22 | 55 | 21 | 59 | 36 | 48 | 33 | 15 | 29 | 56 | 41 | 63 | 22 |
| SE best guess       | 101        | 67         | 134    | 107    | 126  | 95   | 124  | 98   | 134    | 104    | 106  | 77   | 108 | 74 |
| SE increased density| 150        | 95         | 184    | 145    | 183  | 145  | 183  | 127  | 186    | 139    | 165  | 118  | 162 | 113 |
| SE step change only | 154        | 106        | 197    | 163    | 191  | 149  | 187  | 136  | 199    | 155    | 173  | 130  | 169 | 127 |
| NE best guess       | 80         | 56         | 101    | 104    | 113  | 116  | 114  | 100  | 116    | 117    | 96   | 70   | 96  | 73 |
| NE increased density| 123        | 98         | 159    | 149    | 172  | 160  | 168  | 138  | 174    | 156    | 174  | 156  | 172 | 126 |
| NE step change only | 119        | 95         | 151    | 167    | 167  | 168  | 168  | 147  | 169    | 175    | 176  | 148  | 172 | 124 |
| SW best guess       | 127        | 94         | 134    | 116    | 131  | 109  | 119  | 120  | 123    | 108    | 128  | 99  | 128 | 100 |
| SW increased density| 187        | 120        | 196    | 176    | 181  | 144  | 181  | 171  | 184    | 164    | 192  | 128  | 191 | 127 |
| SW step change only | 160        | 117        | 180    | 181    | 171  | 162  | 170  | 170  | 179    | 172    | 162  | 128  | 163 | 129 |
| Total               | 1558       | 1033       | 1816   | 1589   | 1570 | 1462 | 1861 | 1621 |

Note: Cells highlighted in pale grey indicate there are a greater number of extremes exact to measurement precision in the returned than the released data. Cells highlighted in dark grey indicate the best performing algorithm for that region/scenario pair. Total values are only provided for algorithms that homogenized all region/scenario pairs to ensure fair comparisons. Column headings for algorithms are as in Table 4.
TABLE 7 | Rank of each algorithm according to a number of assessment measures (with one being the best performing algorithm)

| Statistic | Algorithm | MACD | CD | CM | MASH | AC | D | H | S |
|-----------|-----------|------|----|----|------|----|---|---|---|
| Percentage of constant offset inhomogeneities found | 4 | 2 | 3 | | 1 | | | 5 |
| Percentage of explanatory variable inhomogeneities found | 3 | 4 | 2 | | 1 | | | 5 |
| Percentage of large inhomogeneities found | 2 | 1 | 3 | | 4 | | | 5 |
| Percentage of medium inhomogeneities found | 4 | 3 | 2 | | 1 | | | 5 |
| Percentage of small inhomogeneities found | 5 | 3 | 2 | | 1 | | | 4 |
| Median false alarm rate | 2 | 4 | 1 | | 5 | | | 3 |
| Regional bias median PR | 5 | 1 | 3 | | 4 | | | 2 | 7 | 8 | 5 |
| Number of unbiased stations | | | | | | | | 1 | 2 | 3 |
| Regional RMSE PR | 5 | 1 | 4 | | 3 | | | 2 | 6 | 8 | 6 |
| Regional trend median PR | 5 | 2 | 3 | | 4 | | | 1 | 6 | 8 | 7 |
| Exact cold extremes | | | | | | | | 2 | 3 | 4 | 1 |
| Exact hot extremes | | | | | | | | 2 | 3 | 4 | 1 |

Note: Not included in these rankings is the number of stations improved/made worse for bias, linear trends and variability as these could be subjective based on acceptable thresholds and will already be reflected in the regional values to some extent. Algorithms that did not homogenize all regions are not included in measures that involve counts. Column headings for algorithms are as in Table 4. DAP, HOM and SpliDHOM cells are merged for detection ability rows as these algorithms share the same detection method and will therefore always be ranked the same.
3.3.2 | Climatol

Both Climatol-Daily and Climatol-Monthly show good homogenization performance. Climatol-Daily has the lowest FAR of any algorithm, while rarely having the lowest HR, although it is normally lower than Climatol-Monthly. Climatol-Daily is the only algorithm to find 100% of any inhomogeneity type (large inhomogeneities) and also the only algorithm to homogenize any stations to perfection. Climatol-Monthly ranks second to ACMANT in finding small, medium, and explanatory variable inhomogeneities.

Climatol-Daily is more cautious, leaving more stations unchanged than some of its contemporaries, but almost always degrading fewer as well. Climatol-Daily is the top performing algorithm when looking at regional bias and RMSE reduction and the total number of stations returned minimally biased. Climatol-Daily generally does better than Climatol-Monthly for recovery of extremes, except in the North East, where Climatol-Monthly also surpasses it for bias and RMSE reduction. For regional trends neither is consistently better than the other.

For Climatol-Daily the increased station density relative to the best guess generally did not result in a difference in algorithm performance apart from in the South West, while for Climatol-Monthly the increased station density appears to have improved algorithm performance. For the step changes only scenario relative to the increased station density scenario Climatol-Monthly's performance appears to be less affected apart from in the South West. Both algorithms saw their detection ability degraded in the increased autocorrelation scenario, although with Climatol-Monthly's FAR also reduced.

3.3.3 | MASH

MASH improves the homogeneity of large numbers of stations, while leaving a non-negligible number made worse. The tendency to change all stations means that MASH is always the top performing algorithm for number of station biases reduced, but when median PRs are considered over all region/scenario pairs for bias, RMSE and trends it is never top.

For variability, MASH is the algorithm most prone to making stations worse, but also improves most in five region/scenario pairs. For extreme recovery MASH is the best algorithm for cold extremes in the increased station density scenario of the South East and always returns a greater number of cold extremes exact to measurement precision than in the released data. However, it does return fewer exact hot extremes in three Wyoming scenarios and two South West scenarios, though the choice of different homogenization periods may have impacted MASH here. MASH showed independence of scenario characteristics, with no consistent change in algorithm performance because of a lack of gradual inhomogeneities or an increased station density and a lower change in performance than most other algorithms in the presence of increased autocorrelations.

3.3.4 | ACMANT

ACMANT is the top performing algorithm for finding small, medium, and explanatory variable inhomogeneities, although it also predominantly has the highest FAR. It is commended for its adjustment ability where it ranks first for regional trend recovery and total number of extremes returned exact to measurement precision across all region/scenario pairs (although it was not always best in individual scenarios and we note that the spread of regional trend recoveries is large). ACMANT reduces over two thirds of the station biases in all but the Wyoming increased autocorrelation scenario and the best guess and increased density scenarios of the South West. However, this is caveated by its tendency to corrupt some station biases, with more than a fifth made worse in the previously mentioned scenarios and in the South West step changes only scenario. ACMANT also returned more negatively than positively biased stations in 12 region/scenario pairs and this should be investigated in its later versions. The different scenario characteristics did not have consistent impacts on ACMANT’s adjustment ability, though its detection ability was lower in the increased autocorrelation scenario than the equivalent best guess scenario in Wyoming (Killick, 2016).

3.3.5 | HOM, SpliDHOM and DAP

In part owing to a lower detection ability, these three algorithms show the most room for improvement as their missing of inhomogeneities limits their ability to perform well based on our adjustment metrics. However, the detection algorithm, common to all three methods, is being updated in 2021 (Petr Stepanek, personal communication).

These algorithms leave large numbers of stations unchanged, with HOM returning most unchanged station biases in every scenario it homogenized. Although these algorithms make some station biases worse, it is generally to a lesser extent than other algorithms, apart from Climatol-Daily. In the Wyoming increased autocorrelation scenario, they increase the station biases of just
two stations, where all other algorithms increase 15 or more. For variability, if these algorithms increase the variability of returned stations it is for the worst more than for the better, more so than most of their counterparts, although their counterparts do not try to homogenize moments higher than the mean and, therefore, have less room for error.

The change in station density did not have a consistent effect on these algorithms and nor did the lack of trend inhomogeneities in the \textit{step changes only} scenario. The \textit{increased autocorrelations} in Wyoming did cause a poorer algorithm performance, likely as the hit rate was further decreased, although the regional bias and linear trend recovery were better in the \textit{increased autocorrelations} scenario than in the \textit{best guess} scenario.

4 CONCLUSIONS

This study evaluated the performance of eight homogenization algorithms according to both their ability to detect inhomogeneities and to correct for the impacts of these inhomogeneities. The data used were created from a statistical model, allowing the benchmarking of algorithm performance because the truths about the underlying data were known completely.

All eight homogenization algorithms improved the data in some way, indicating that the investment in developing daily homogenization algorithms is important. RMSE was consistently reduced in each region/scenario pair, although on a station by station basis some were increased. However, this benchmarking exercise has indicated that more work needs to be done to allow effective recovery of station variability and extremes in homogenization. Trend recovery was variable, but no regional trends experienced a sign change during homogenization. Regional trends significant at the 5% level on release were always significant on return as well, including in the South West \textit{increased station density} scenario where the trend in the clean data had not been significant.

Venema et al. (2012) also assessed ACMANT, MASH and Climatol-Monthly and found ACMANT and Climatol-Monthly to have a good HR. Their study found that ACMANT performed well at the station level using CRMSE, a similar measure to RMSE, but less well at the regional level, similar to what was found in this study for MASH, but not to the same extent for ACMANT. Their study did not show these algorithms degrading station quality, although this could be an artefact of the monthly resolution. MASH outperformed ACMANT in network trend recovery in Venema et al. (2012), which was true in some region/scenario pairs in this study, but not consistently.

A daily humidity homogenization study detailed in Chimani et al. (2018) made use of ACMANT, MASH and SpliDHOM. Their findings showed similarities of SpliDHOM homogenizing fewer stations than MASH or ACMANT. In their study over 50% of RMSEs and trends were improved for all three homogenization algorithms, which agrees with what is seen in this paper for MASH and ACMANT, but for SpliDHOM it was not uncommon for over 50% of RMSEs and trends to be unchanged by homogenization, although over 50% were improved in the \textit{increased station density} and \textit{step changes only} scenarios in the South East and the \textit{best guess} in the North East.

Both Venema et al. (2012) and Chimani et al. (2018) commend ACMANT in their algorithm comparisons and results from the present study support this commendation. ACMANT also shows the highest HRs for the smallest, and traditionally hardest to find, inhomogeneities, although this still results in almost 75% of inhomogeneities less than or equal to 0.2°C in magnitude being missed. Climatol-Daily should also be commended. It is a far more cautious algorithm, leaving more stations un-homogenized, but also making fewer worse than most others, in part owing to its consistently low FAR.

The variation in the detection ability of inhomogeneities of different sizes indicates that a variety of inhomogeneity structures are beneficial in any benchmark data set. The evidence of differing algorithm performance in different regions and scenarios also indicates that underlying data structure can impact algorithm results, with \textit{increased autocorrelations} causing the most consistent degradation in algorithm performance.

When choosing an algorithm to apply to a data set, users should consider all the characteristics of the data that are known, for example, station density, missing data, autocorrelation. If metadata are available, then it is advisable to try multiple algorithms as an idea of HR can then be determined. FAR is harder to judge as some changes may not be recorded in the metadata. Whether regional or station level characteristics are of interest should also be considered. If regional, then Table 7 should be a good guide to the most reliable algorithms; if station level then some guidance can be gained from the number of stations improved, unchanged and made worse in this study, with the caveat that in this study no lenience was allowed for small changes.

The clean and released data used in this study are freely available and their use is encouraged to continue this benchmarking study as existing algorithms undergo development and new ones are produced. The code used to evaluate the returned algorithms is available on GitHub at https://github.com/RachelKillick/Daily_benchmarks.
ACKNOWLEDGEMENTS
The authors express their sincere thanks to Peter Domonkos, Jose Guijarro, Petr Stepanek, Michele Rienzner and Tamás Szentimrey for running their algorithms on the data for this study and allowing us to freely discuss the results. The authors also thank two anonymous reviewers for their helpful suggestions to improve this paper. REK (in part) and KMW were supported for this work by the Met Office Hadley Centre Climate Program funded by BEIS and Defra. This article resulted from work carried out under EPSRC (1199873) and National Meteorological Office UK CASE funding.

AUTHOR CONTRIBUTIONS
Rachel Killick: Conceptualization; data curation; formal analysis; investigation; methodology; project administration; validation; visualization. Ian Jolliffe: Conceptualization; methodology; project administration; supervision; validation. Kate Willett: Conceptualization; funding acquisition; methodology; project administration; supervision.

ORCID
Rachel E. Killick https://orcid.org/0000-0001-5335-4097
Kate M. Willett https://orcid.org/0000-0001-5151-0076

REFERENCES
Alexandersson, H. (1986) A homogeneity test applied to precipitation data. Journal of Climatology, 6, 661–675.
Brohan, P., Kennedy, J.J., Harris, I., Tett, S.F.B. and Jones, P.D. (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. Journal of Geophysical Research: Atmospheres, 111, 1–24.
Caussinus, H. and Lyazhri, F. (1997) Choosing a linear model with a random number of change-points and outliers. Annals of the Institute of Statistical Mathematics, 49, 761–775.
Caussinus, H. and Mestre, O. (2004) Detection and correction of artificial shifts in climate time series. Journal of the Royal Statistical Society. Series C: Applied Statistics, 53, 405–425.
Chimani, B., Venema, V., Lexer, A., Andre, K., Auer, I. and Nemec, J. (2018) Intercomparison of methods to homogenize daily relative humidity. International Journal of Climatology, 38, 3106–3122. https://doi.org/10.1002/joc.5488.
Conrad, V. (1946) Methods in Climatology, 2nd edition. Massachusetts: Harvard University Press.
Della-Marta, P.M. and Wanner, H. (2006) A method for homogenising the extremes of mean daily temperature measurements. Journal of Climate, 19, 838–853.
Domonkos, P. (2011) Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. Theoretical and Applied Climatology, 105, 455–467.
Domonkos, P. (2014) The acmant2 software package. In: Proceedings of the Eighth Seminar for Homogenisation and Quality Control in Climatological Databases and Third Conference on Spatial Interpolation Techniques in Climatology and Meteorology, Geneva: World Meteorological Organisation; pp. 46–72.
Domonkos, P. and Coll, J. (2017) Homogenisation of temperature and precipitation time series with ACMANT3: method description and efficiency tests. International Journal of Climatology, 37, 1910–1921.
Domonkos, P., Guijarro, J.A., Venema, V., Brunet, M. and Sigro, J. (2021) Efficiency of time series homogenization: method comparison with twelve monthly test datasets. Journal of Climate, 34, 2877–2891.
Easterling, D.R. and Peterson, T.C. (1995) A new method for detecting undocumented discontinuities in climatological time series. International Journal of Climatology, 15, 369–377.
Gross, M.H., Donat, M.G., Alexander, L.V. and Sisson, S.A. (2018) The sensitivity of daily temperature variability and extremes to dataset choice. Journal of Climate, 31, 1337–1359.
Killick, R.E. (2016) Benchmarking the performance of homogenisation algorithms on daily temperature data. PhD Thesis, Department of Mathematics, University of Exeter.
Killick, R.E., Jolliffe, I.T. and Willett, K.M. (2019) Use of generalized additive modelling techniques to create synthetic daily temperature networks for benchmarking homogenization algorithms. Dynamics and Statistics of the Climate System, 2019, 1–29.
Lindau, R. and Venema, V. (2018) On the reduction of trend errors by the ANOVA joint correction scheme used in homogenization of climate station records. International Journal of Climatology, 38, 5255–5271.
Menne, M.J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R.S., Gleason, B.E. and Houston, T.G. (2012) Global historical climatology network-daily (Ghcn-Daily), version 3.00. Washington: NOAA National Climatic Data Centre. Available at: http://doi.org/10.7289/V5D21VHZ [Accessed 19th November 2012].
Menne, M.J., Williams, C.N., Jr. and Vose, R.S. (2009) The US Historical Climatology Network Monthly Temperature Data, Version 2. Massachusetts: Bulletin of the American Meteorological Society, pp. 993–1007.
Mestre, O., Gruber, C., Prieur, C., Caussinus, H. and Jourdain, S. (2011) SPLIDHOM: a method for homogenisation of daily temperature observations. Journal of Applied Meteorology and Climatology, 50, 2343–2358.
Murphy, A.H. (1988) Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review, 116, 2417–2424.
Potter, K.W. (1981) Illustration of a new test for detecting a shift in mean precipitation series. Monthly Weather Review, 109, 2040–2045.
Rienzner, M. and Gandolfi, C. (2011) A composite statistical method for the detection of multiple undocumented abrupt changes in the mean value within a time series. International Journal of Climatology, 31, 742–755.
Rienzner, M. and Gandolfi, C. (2013) A procedure for the detection of undocumented multiple abrupt changes in the mean value of a daily temperature time series of a regional network. International Journal of Climatology, 33, 1107–1120.
Squitu, A.A., van der Schneider, G., Stepanek, P., Zahradnicek, P. and Klein Tank, A. (2020) Comparison of homogenization methods for daily temperature series against an observation-
based benchmark dataset. *Theoretical and Applied Climatology*, 140, 285–301.

Stepanek, P., Zahradnicek, P. and Farda, A. (2013) Experiences with data quality control and homogenisation of daily records of various meteorological elements in The Czech Republic in the period 1961–2010. *Idiojaras, Quarterly Journal of the Hungarian Meteorological Service*, 117, 123–141.

Szentimrey, T. (2008) Development of MASH homogenisation procedure for daily data. In: *Proceedings of the Fifth Seminar for Homogenisation and Quality Control in Climatological Databases*, WMO, Geneva. 116–125.

Venema, V.K.C., Mestre, O., Aguilar, E., Guijarro, J.A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Muller-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 homogenisation algorithms for monthly data. *Climate of the Past*, 8, 89–115.

Willett, K., Williams, C., Joliffe, I.T., Lund, R., Alexander, L.V., Broonimann, S., Vincent, I.A., Easterbrook, S., Venema, V.K.C., Berry, D., Warren, R.E., Lopardo, G., Auchmann, R., Aguilar, E., Menne, M.J., Gallagher, C., Hausfather, Z., Thorarinsdottir, T. and Thorne, P.W. (2014) A framework for benchmarking homogenisation algorithm performance on the global scale. *Geoscientific Instrumentation Methods and Data Systems*, 3, 187–200.

Williams, C.N., Menne, M.J. and Thorne, P.W. (2012) Benchmarking the performance of pairwise homogenisation of surface temperatures in the United States. *Journal of Geophysical Research: Atmospheres*, 117, 1–16.

**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of the article at the publisher’s website.

*How to cite this article:* Killick, R. E., Joliffe, I. T., & Willett, K. M. (2022). Benchmarking the performance of homogenization algorithms on synthetic daily temperature data. *International Journal of Climatology*, 42(7), 3968–3986. [https://doi.org/10.1002/joc.7462](https://doi.org/10.1002/joc.7462)