University of Exeter
Department of Mathematics

Benchmarking the Performance of Homogenisation Algorithms on Daily Temperature Data

Rachel Elizabeth Killick

March 2016

Supervised by Professor Trevor Bailey, Professor Ian Jolliffe and Dr Kate Willett

Submitted by Rachel Elizabeth Killick, to the University of Exeter as a thesis for the degree of Doctor of Philosophy in Mathematics , March 2016.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

(signature) .....

Abstract

Reliable temperature time series are necessary to quantify how our world is changing. Unfortunately many non-climatic artefacts, known as inhomogeneities, affect these time series. When looking at real world data it is often not possible to distinguish between these non-climatic artefacts and true climatic variations that are naturally found in our world. Therefore, trying to remove the non-climatic artefacts with complete confidence is problematic, but leaving them in could lead to misinterpretation of climate variations. In creating realistic, homogeneous, synthetic, daily temperature series the truth can be known about the data completely. Known, created inhomogeneity structures can be added to these series, allowing the distinguishing between true and artificial artefacts. The application of homogenisation algorithms to these created inhomogeneous data allows the assessment of algorithm performance, as their returned contributions are being compared to a known standard or benchmark, the clean data.

In this work a Generalised Additive Model (GAM) was used to create synthetic, clean, daily temperature series. Daily data pose new challenges compared to monthly or annual data owing to their increased variability and quantity. This is the first intercomparison study to assess homogenisation algorithm performance on temperature data at the daily level. The inhomogeneity structures added to the clean data were created by perturbing the inputs to the GAM, which created seasonally varying inhomogeneities, and by adding constant offsets, which created constant inhomogeneities. Four different regions in the United States were modelled, these four regions are climatically diverse which allowed for the exploration of the impact of this on homogenisation algorithm performance. Four different data scenarios, incorporating three different inhomogeneity structures, were added and evaluations also investigated how these impacted algorithm performance. Eight homogenisation algorithms were contributed to this study and their performance was assessed according to both their ability to detect change points and their ability to return series that were closer to the clean data than they were on release. These evaluations sought to aid the improvement of these algorithms and enable a quantification of the uncertainty remaining in daily temperature data even after homogenisation has taken place. Evaluations were also made of the benchmarks as it was important that benchmark weaknesses were taken into account. It was found that more climatologically diverse regions were harder to model and less climatologically diverse regions were easier to homogenise. Station density in a network and the presence of artificial trend inhomogeneities did not impact algorithm performance as much as changes in autocorrelations did, and the latter area was an area that most algorithms could improve on.

This work feeds into the larger project of the International Surface Temperature Initiative

which is working on a wider scale and with monthly instead of daily data.

For my parents, Janet and Mike Warren, for always encouraging me to do my best and for my husband Peter Killick who has supported me through this PhD.

My thanks also go to all those who have enabled this PhD to come to fruition. To my supervisors Trevor Bailey, Ian Jolliffe and Kate Willett for their guidance and comments. To those who contributed algorithms to this work to enable it to be a comparison study; Peter Domonkos, José Guijarro, Michele Rienzner, Petr Stepanek and Tamas Szentimrey. To those who provided code for this PhD, Mark Cherrie, Robert Dunn and Peter Killick. To the friends who did PhDs before me and were able to offer their advice, John Allsup, Phil Sansom and Dave Singeisen. Also to the family and friends who haven't done PhDs, but kept me sane during the process, Sophie Warren, Vincent Jarman and Sarah Blane. Last, but by no means least, to all those friends and family who have offered prayerful support through the entirety of the PhD process, you are too many to thank, but your support has been invaluable.

Contents

List of Tables

List of Figures

1. Introduction

This chapter explains the motivations and aims of the research described in this thesis and gives an overview of the content which follows in the rest of the thesis. It is designed to aid the reader's understanding by providing them with a brief introduction to the areas of homogenisation and benchmarking, which will be expanded on in the literature review in chapter two.

1.1. Motivations

Attempts are increasingly being made to quantify how the world is changing and what is causing these changes. Changes in climate are frequently attributed to anthropogenic (man made) or natural drivers and these changes are detected using in situ observations, amongst other sources [IPCC, 2014]. However, in situ observations have long since been known not to be error free. Widely quoted in the homogenisation literature is Viktor Conrad's definition of a homogeneous series as being one that has variations "caused by and only by variations of weather and climate" [Conrad, 1946]. This quote comes after his comment that a manifold of causes exist to stop a time series being homogeneous, just two of which are changes in instrumentation and changes in station surroundings.

Artefacts that stop a time series from being homogeneous are known as inhomogeneities. These inhomogeneities confound attempts to draw conclusions from in situ data because their magnitudes are often similar to the magnitudes of true climate artefacts [Williams et al., 2012]. Therefore, it is necessary to remove these inhomogeneities in order to create time series' that can be relied upon in climate studies; this process is known as homogenisation. Multiple algorithms have been designed to homogenise data and comparison studies have been undertaken in the past to compare their strengths and weaknesses, Easterling and Peterson [1995]; Reeves et al. [2007]; Venema et al. [2012], to name a few.

If algorithms are being run on real world data then reliably assessing their performance is problematic as the correct answer is not known. Therefore, it is common practice to instead create synthetic data, where known inhomogeneities are inserted into known homogeneous series', meaning that the truth about the data is known a priori. This process is known as benchmarking, and if the truth of the homogeneous data, often called clean data, is not revealed to the homogenisers until after the algorithms have been run then it is known as blind benchmarking. The largest comparison study to date, Venema et al. [2012], also known as COST HOME, was a blind benchmarking study. It compared

25 variants of 13 different algorithms by assessing their performance on a variety of inhomogeneous scenarios which were created by the first author. A larger comparison study still is currently being planned. The International Surface Temperature Initiative plan to compare multiple homogenisation algorithms on a range of global, monthly benchmark datasets. The temporal and spatial scale of this project will vastly exceed all previous work in this area. The current project is a part of this larger project, but looking at daily data on a smaller scale.

At the beginning of the write up of COST HOME one motivation for the project given was that, although there were many homogenisation algorithms available for monthly data, previous comparisons had been carried out at the annual level. Therefore, Venema et al. [2012] created monthly benchmark datasets for their analysis. Now that COST HOME has been completed the next logical step is to carry out a benchmarking study at the daily scale. There are not a plethora of algorithms available which are designed to work with daily data, therefore, this thesis has dual motivation. Firstly, to assess the existing algorithms that can be applied to daily data and secondly, to encourage the development of further algorithms by providing a tool with which to develop them.

Daily data are of interest because it is often at this level that societal impacts of climate change are felt. Heat waves that last a few days will be aggregated out of data when time series are looked at on an annual or monthly scale, but it is important to know how often these events are occurring as they are events that can claim lives. Thus, being able to look at trends in extremes of data is incredibly valuable, but should only be carried out on reliable daily time series.

Daily data are more complex to work with than monthly or annual data. They vary over shorter space scales, thus making comparisons between stations more difficult. They are typically not normally distributed, thus making creation of synthetic data more difficult. Finally, their extra variability internally confounds attempts to distinguish true inhomogeneities from natural temperature variations. Given this extra variability that is present in daily data it is not sufficient to assess homogenisation algorithms based only on their adjustments to the mean of the data. Instead, this study will look at a variety of algorithm assessment measures, including analyses of returned station variabilities and extremes.

Although temperature is not the only climate variable in need of homogenisation it will be the sole variable focussed on in this project. This is because it was one of two variables identified as being of most interest to the homogenisation community in 2012 [Venema et al., 2012]. The other variable identified was precipitation, which is arguably still more complex than daily temperature data as the spatial correlations are greatly reduced.

1.2. Data and Analysis

No global, daily, homogenised dataset exists. However, there are sources of daily, quality controlled data. Quality control differs from homogenisation in that it looks for effects with

random consequences e.g. it might seek to identify a value recorded a factor of ten out, whereas homogenisation looks for systematic effects e.g. effects on temperatures arising due to a change in instrument. In this study, the data are created in such a manner that participating homogenisers can assume that no effects with random consequences are present, and thus, no quality control needs to take place.

The choice of data used to build the model for creating synthetic temperature series in this project will be expanded on in chapter three, but it had to meet certain requirements. The temperature data had to be daily in nature and available at the station level in order to be able to investigate the spatial and temporal correlations that are present in data at this scale so that these could be reproduced and the quality of their incorporation in the benchmarks could be assessed. Given the modelling nature of this project, it was also desirable that data for more than just temperature were available for all stations so that these other variables could be used to enhance the model. It was also known that benchmarks should be seeking to be realistic in terms of observed longer-term variability such as the El-Niño Southern Oscillation.

All data analysis was carried out using freely available software; predominantly R, but with a few small scripts written for Python. All scripts are available on request and it is the author's hope that these may be used to further the work begun in this project.

1.3. Aims

High quality daily data are necessary for the climate research community, but inhomogeneities are all too frequent in the majority of these series. Daily homogenisation is still in its infancy and no previous benchmarking comparison studies for daily data have taken place. Therefore, this thesis has sought to aid the homogenisation and climate communities by providing daily mean temperature benchmarks that explore multiple different regions and scenarios. It also provides a validation framework that has been implemented to assess the performance of existing homogenisation methods and aid the development of new ones. This framework was implemented by the author in collaboration with the homogenisation community. The overall aims of this study were as follows:

1. To design a model capable of creating realistic, clean, daily data that could act as benchmarks. This model had to be capable of reproducing true data autocorrelations and inter-station correlations and be able to be easily generalised to multiple different regions. It also had to be able to produce large (> 100 stations) networks and ideally be able to incorporate other climatic variables.

2. To design a realistic range of inhomogeneity structures that could be added on to the clean data. These were the released data scenarios. These data needed to explore a range of inhomogeneity types known to affect daily temperature data and also explore the impacts of changing the station and network characteristics themselves.

3. To engage the homogenisation community and encourage their involvement in this work by providing them with the released data to homogenise and keeping them blind to the clean data until after the algorithm assessment.

4. To assess the homogenised contributions and provide an analysis of algorithm strengths and weaknesses and the uncertainty remaining in the homogenised data. This analysis was fed back to the homogenisers to aid further development of their algorithms.

5. To assess the quality of the created benchmarks and identify areas for improvement in a future iteration of this project.

1.4. Thesis Overview

The rest of the thesis proceeds as follows. Chapter two provides a more in depth description of previous homogenisation and benchmarking studies. It explains previous approaches to clean and inhomogeneous data creation and also looks at previous methods of assessing algorithm performance.

Chapter three introduces the data used for this project and the necessary pre-processing steps to be able to use them. It explains the motivations behind the choices of these data sources and locations where the reader can access the same data.

Chapter four looks at the modelling of daily temperature data. It gives an introduction to the statistical models considered when creating the benchmark clean data before expanding on the reasons for the final model choice and the variables contained within it. This chapter concludes with an assessment of the similarity between the created clean data and the real world data.

Chapter five introduces the inhomogeneity structures used in this benchmarking study. These form the released data scenarios that were made available to the homogenisation community. The reasoning behind the different scenarios created is also explained in this chapter.

Chapter six lays out the validation framework that was used to assess the returned data produced by homogenisers running their algorithms on the released data. This chapter introduces the different measures for assessing an algorithm's adjustment or detection ability.

Chapter seven introduces the algorithms evaluated in this project and then proceeds to explain the results of implementing the validation framework of chapter six on the returned data from them. It first gives an overview of how all algorithms performed for each validation measure before going on to quantify the strengths and weaknesses of each contributed algorithm. The chapter concludes with an assessment of the uncertainty remaining in the data after homogenisation.

Chapter eight provides a summary of the accomplishments of this thesis and highlights areas for future work.

2. Literature Review

Chapter one has introduced the motivation and aims for this project and given an overview of what follows in the rest of the thesis. This chapter explains in more detail previous work carried out in the areas of homogenisation and benchmarking and gives an overview of previous validation measures used when assessing homogenisation algorithm performance.

2.1. Homogenisation

Homogenisation, in the climate context, refers to the act of removing the non-climatic artefacts, inhomogeneities, from a time series. Multiple studies have illustrated the need for homogenisation of climate data, whether that is by illustrating the effects inhomogeneities can have, e.g. figure 2.1 taken from Menne et al. [2009]; highlighting that they can confound our attempts to quantify how our climate is changing because of their similar magnitudes to true climate signals [Della-Marta and Wanner, 2006; Venema et al., 2012; Williams et al., 2012]; or by highlighting the causes of the inhomogeneities themselves [Trewin, 2010; Parker, 1994; Hubbard and Lin, 2006].

Often, inhomogeneities are divided into two categories, step and trend changes [Menne and Williams JR., 2009]. The former is usually caused by a specific event in time, e.g., a station relocation [Xu et al., 2013], or a change in instrumentation [Willett et al., 2014]. Trend inhomogeneities are commonly caused by a change in station surroundings, the most common cause of which is urbanisation [Trewin, 2010]. Another cause of a trend inhomogeneity could be the deterioration of an instrument or shelter [Lopardo et al., 2014]. The effects of inhomogeneities could be largely constant e.g., because of a thermometer error, or could vary diurnally e.g., shelter change impacts [Parker, 1994], or seasonally, e.g., because of seasonal changes in surrounding vegetation at a new site (Blair Trewin, personal communication). More focus is given to non-constant inhomogeneities in this work, as constant inhomogeneities have been studied more in the past, see section 2.2 of this chapter. An overview of the inhomogeneities specifically focused on in this work and the reason for their choice is given in section one of chapter five.

As this project is focused on temperature homogenisation, this is also the area of homogenisation that this review focuses on. Temperature homogenisation has been the focal point of numerous studies in the past (see Yozgatligil and Yazici [2015] for an extensive list). It is one of the main variables to be homogenised because it has become the variable most commonly associated with the quantification of climate change. It is also

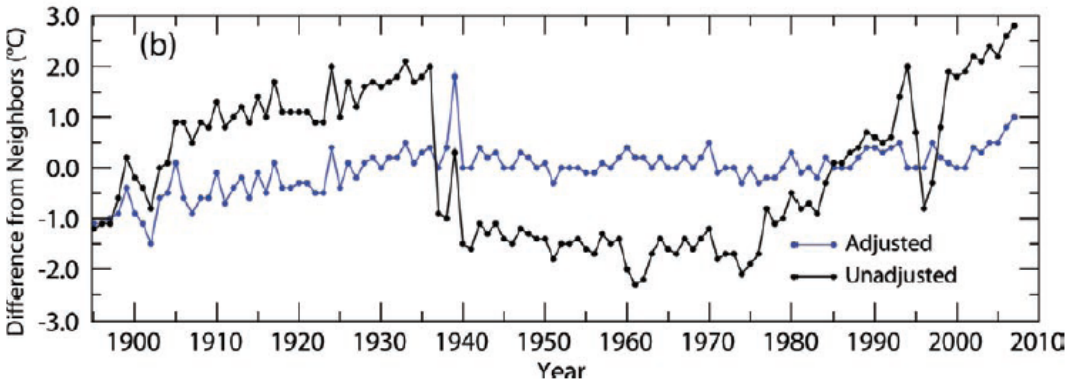

Figure 2.1. Figure 8b from [Menne et al., 2009]. This figure illustrates the difference in minimum adjusted (blue) and unadjusted (black) temperatures between Reno, Nevada and the mean of its ten nearest neighbours. The steps in this series were caused by station relocations and the trend was caused by urbanisation of the surrounding area.

commonly homogenised because there are typically long records available for temperature data and temperature is measured near globally, though the spatial coverage varies dramatically. The second most commonly homogenised variable is precipitation; this too is directly relevant to society, which is why acquiring more reliable records through homogenisation is so valuable. Venema et al. [2012], which has been one of the largest studies assessing homogenisation algorithms to date, looked at the homogenisation of precipitation data, and therefore references to precipitation homogenisation studies can be found in section one of Venema et al. [2012].

A great number of homogenisation algorithms are available and they have been being developed for many years, although many of these algorithms should correctly be referred to as change point detection algorithms as they seek only to find the inhomogeneities and not to also correct for them. As stated in section one of the introduction, one of the earliest references to homogenisation is from Viktor Conrad in the mid twentieth century. Conrad [1946] defined a series to be homogeneous if its variations were caused only by variations in weather and climate. He went on to define relative homogeneity; ”A climatological series is relatively homogeneous with respect to a synchronous series at another place if the differences (ratios) of the pairs of the homologous averages represent a series of random numbers which satisfies the law of errors”, and absolute homogeneity, when a series itself, i.e., not a difference series, is deemed to contain no inhomogeneities. These definitions have spawned the groupings of homogenisation algorithms today: absolute homogenisation methods are applied to each station separately, whereas relative methods use reference stations to assess the homogeneity of a candidate station [Costa and Soares, 2009].

As demonstrated by Venema et al. [2012], relative homogenisation algorithms are, in general, more reliable than absolute homogenisation algorithms, and the majority of algorithms do now fall into the former category. However, as Costa and Soares [2009] point out, both methods have disadvantages. Relative homogenisation methods cannot cope with simultaneous changes across a network as they will assume it to be a true climatic event, and absolute methods are overly reliant on metadata (data about data) to

distinguish between true and artificial changes, but metadata is all too often not available [Costa and Soares, 2009; Menne et al., 2009]. All algorithms applied in this study were relative homogenisation algorithms, though they differed in whether they used a single station as a reference series or multiple stations. For a review of absolute homogenisation algorithms the reader is referred to Reeves et al. [2007].

It is commonly acknowledged that a perfect algorithm does not exist [Thorne et al., 2011b; Williams et al., 2012]. However, knowing this has spurred on many comparison studies between different algorithms. The studies of Reeves et al. [2007] and Venema et al. [2012] have already been mentioned, but there are many others. Easterling and Peterson [1995] compared relatively early homogenisation methods whilst developing their own; Ducre-Robitaille et al. [2003] and DeGaetano [2006] compared the abilities of multiple algorithms for detecting step changes and Domonkos [2011] took the comparisons one step further by also looking at the correction abilities of algorithms. This is not an exhaustive list. The most recent comparison studies have been those of Venema et al. [2012] and Yozgatligil and Yazici [2015], though the latter was once more only assessing change point detection and not correction also. The upcoming study eluded to in Willett et al. [2014] should be by far the largest homogenisation comparison study undertaken as it will look at multiple algorithms' detection and adjustment abilities on a global scale.

The existence of so many comparison studies already does not remove the need for a further study. Venema et al. [2012] was the first study to compare multiple algorithms at the monthly level and the current study is the first to compare multiple algorithms at the daily level. However, even once multiple studies are available at the monthly and daily levels there will still be the need for more studies to enable different inhomogeneity scenarios, underlying data structures and new algorithms to be further investigated. This is why the International Surface Temperature Initiative plans on having a cycle of homogenisation comparison studies, [Willett et al., 2014], and why the author encourages a future iteration of this daily project.

2.2. Benchmarking

The Penguin English Dictionary defines a benchmark as 'Something that serves as a standard by which others may be measured' [Allen, 2003]. For homogenisation the benchmark is the known truth, usually created clean data, and the 'others being measured' are the returned data obtained by having run homogenisation algorithms on the released data. Such benchmarks are necessary as when the truth that is being aimed for is known, a reliable quantification of errors is possible, but when the truth is not known a priori no such reliable quantification is possible. In the area of homogenisation there have been some well known benchmarking studies, most notably the COST HOME action detailed in Venema et al. [2012] and the study comparing variants of the Pairwise Homogenisation Algorithm carried out by Williams et al. [2012]. The International Surface Temperature Initiative's current project, [Willett et al., 2014], will be the first global

benchmarking study. All of these studies look at monthly temperature data, and, in the case of Venema et al. [2012], monthly precipitation data as well.

The process of benchmarking homogenisation algorithms first requires the creation of the benchmark data. This creation process can be split into two stages; the creation of homogeneous (clean) data and the creation of error structures used to corrupt this data in a known manner to produce the released data. What follows is a brief review of some of the methods used at both these stages in studies to date.

2.2.1. Clean data creation

As stated previously most homogenisation studies to date have focussed on monthly or annual data. Therefore, this review will focus on these time scales. Also, as already stated, many studies only focus on change point detection and not the subsequent adjustment of the data. These change point detection studies will still be mentioned as they do involve data creation. However, the reader is cautioned that the focus of a full homogenisation study is on more than just the correct identification of change point locations (see section 2.2.3).

It is important to create realistic benchmarks to get a true idea of algorithm performance. The key elements that should be well replicated are station auto-correlations and cross-correlations, realistic trends, long- and short-term variability and realistic station level climatology [Willett et al., 2014]. The study of Willett et al. [2014] seeks to do this using a combination of interpolated Global Climate Model (GCM) output for long term trend and regional variability; analysis of the stations being modelled to get true seasonal cycles and station variabilities; and a vector autoregressive model to ensure realistic correlations. Their work is global (32 000 stations) and includes stations of various record lengths and qualities. The approach used is well suited to their work. However, for daily data added complexities arise with increasing variability in the data because of the higher time resolution and, therefore, this approach is not being adopted for this study.

The study of Williams et al. [2012] compares variants of the same algorithm and also uses interpolated GCM data to create their clean benchmarks. These data are interpolated to the station level over the contiguous United States. The stations created then had climatological offsets and noise added in a manner that allowed inter-station correlations and autocorrelations to be approximately equal to those of the observed network. This study determined the necessary correlations from previously homogenised data. As such data are often not available this approach was not adopted in the current study.

A different approach to creating clean data was taken by the COST HOME initiative and is detailed in both Venema et al. [2011] and Venema et al. [2012]. They base their created clean data on small networks of 5, 9 or 15 stations in Europe and their created data series are 100 years in length. Some of their test data comes from the real world and is therefore naturally not perfect. They also create 'surrogate' station networks. The time series' for these networks are created using the Iterative Amplitude Fast Fourier Trans-

form (IAFFT) algorithm (see Venema et al. [2006]) and reproduce true network spatial and temporal correlations. However, the IAFFT algorithm requires homogeneous data as an input and is therefore impractical for the present study where such data are rarely available, especially at the daily level necessary here. Venema et al. [2012] also create 'synthetic' station networks based on their surrogate station networks. Each synthetic network is paired with a surrogate station network and mimics its cross-correlations, means and standard deviations. However, it does not mimic the auto-correlations; the difference series between pairs of stations in the synthetic networks are temporally uncorrelated Gaussian white noise which is a simplification of reality. Venema et al. [2012] point out that the assumption of difference series between stations being white noise is a common one and that is why they chose to have synthetic and surrogate station networks - to investigate the impact of this assumption. The present study also investigates the impact of this assumption, as detailed in chapters four and five.

Many studies focus on modelling standardised anomaly series (those that have had the seasonal cycle removed and been divided by the series standard deviation). This allows generation of time series from relatively simple statistical models with a mean of zero, a standard deviation of one and usually a small auto-regressive parameter to ensure some low level of autocorrelation enters the model, for example Ducre-Robitaille et al. [2003]. DeGaetano [2006] built on this work to also incorporate correlations between stations using a multivariate normal model and observed climate time series to base their station information on. Menne and Williams JR. [2005] also modelled anomaly series and take information about the autocorrelations and inter-station correlations from observed annual temperature series in the United States.

A problem that can arise from working with standardised anomaly series is that, to then generate non-standardised and non-deseasonalised data, realistic means and variances must be determined. The means could be decided as a set value and then added back on as was done in Easterling and Peterson [1995], or by taking them from real world series, as will be done in Willett et al. [2014]. However, many homogenisation algorithms will only work with anomaly series anyway, thus eliminating the problem of adding variation back on to statistically created series. Some studies even focus on creating the difference series between two anomaly series given that many homogenisation methods now work with relative time series (candidate minus reference) anyway. For example, Domonkos [2008b] and Domonkos [2011] create synthetic difference series with reference to observed difference series.

As already mentioned, a very nice overview of many homogenisation studies can be found at the beginning of Yozgatligil and Yazici [2015]. These authors then go on to create data using a similar method to Yozgatligil et al. [2011] and Yazici et al. [2012] which takes into account the known autocorrelations in monthly temperature series by using a time series model. This time series model estimates a mean temperature and then adds a monthly offset to this value according to the time of year. A normally distributed error term with constant variance is also included. Highly correlated reference series from this method were created by slightly perturbing the seasonal offsets and error terms

generated from a multivariate normal distribution. The study by Yozgatligil et al. [2011] also investigated the impact of changing the variance of the series by generating three series with increasing variabilities.

Titchner et al. [2009] state the importance of basing clean data creation on realistic models instead of simple, randomly generated series, as it allows the capturing of real variations in climate such as inter-annual modes of variability, for example, the El Niño Southern Oscillation (ENSO). They adopt this approach when creating synthetic radiosonde data and then add noise on to the grid box values when downscaling to the station level.

A potential limit when basing created data on observed data is the lack of observing sites available. Studies that generate series using models (statistical or climatological) are less likely to face this problem as thousands of simulations can be run based broadly on real world properties. The present study uses observations and reanalysis data to capture the behaviour of real world temperature series, but it uses these in a statistical model that is able to create homogeneous stations at any location in the study area. It also models full time series with realistic seasonal cycles and variability, thus avoiding the potential problem raised above of having to model or determine these aspects separately.

2.2.2. Corrupted data creation

After synthetic clean data have been created, error structures need to be added on to allow the assessment of homogenisation algorithm performance. These structures differ in their complexity and content depending on the primary focus of the study in question.

The study of Reeves et al. [2007] assessed algorithm performance in the presence of at most one change point (AMOC). The location and magnitude of this change point was allowed to vary between three values and could also be accompanied by a trend change. A similar study of AMOC was carried out by Lund et al. [2007], but they investigated the impact of autocorrelation and periodicity in the data, still with a random inhomogeneity time allocation, but without any trend changes.

An increase in the complexity of the structure could be to add more than one change point into a series or to allow more variation in the size of these inhomogeneities. These aspects were assessed in varying degrees by Easterling and Peterson [1995], Ducre-Robitaille et al. [2003] and DeGaetano [2006]. All of these studies drew the time and size of inhomogeneities from a pre-specified range of values (discrete or continuous), which allowed the assessment of algorithm performance according to the time between change points [DeGaetano, 2006].

Another study also looking at a wider range of inhomogeneity sizes and locations was that of Menne and Williams JR. [2005] where step inhomogeneities had magnitudes drawn from a Normal distribution with mean 0 and variance 1. This distribution was chosen as their focus was on US temperature series where they showed a Normal(0,1) to be a reasonable proxy for observed inhomogeneity sizes (after they have been standardised).

They also didn't restrict the time when an inhomogeneity could occur, though they only assessed detection of inhomogeneities separated by at least 5 time points. The current study also imposes no lower limit on time between change points, but employs a windowing approach when validating algorithm performance that accounts for non-exact change point detection and near simultaneous change points.

Structures of varying complexity were created by Williams et al. [2012] when assessing the performance of the Pairwise Homogenisation Algorithm (PHA).These structures always contained step changes that did not vary seasonally, but the characteristics of these step changes were broad. In four scenarios they investigated the impact of various combinations of inhomogeneities that were: small or large; frequent or sparse; clustered or isolated; prone to bias or with an average size of zero and supported or unsupported by metadata. They also assessed whether algorithm performance was dependent on the underlying climate signal by recreating the same scenario with underlying data from four different climate models. They drew information in part from their knowledge of the US temperatures, which Karl and Williams JR. [1987] also did in an earlier study.

The COST HOME initiative also drew information from the real world and they assessed a real world section of the benchmark so they could evaluate how realistic their added inhomogeneity structures were [Venema et al., 2012]. They didn't limit the added inhomogeneities to just step changes, but also added trend inhomogeneities; these were between 30 and 60 years in length and had a magnitude drawn from a Normal(0,0.8) distribution. An underlying trend was also added to simulate climate change that should not be treated as an inhomogeneity. The step changes they added also had magnitudes drawn from a Normal(0,0.8) distribution and were seasonally varying with a seasonal cycle that had a variance of 0.4°C. Clustered change points were also allowed to simulate network wide station changes. The change point locations were modelled using a Poisson process, which is the method also employed in this thesis.

Many studies have been carried out recently by Peter Domonkos exploring various inhomogeneity structures [Domonkos, 2008a,b, 2011, 2013]. He also explored real world data, from Hungary, and his findings revealed that small inhomogeneities were more frequent than large inhomogeneities and inhomogeneities that only affect a short time period are more common than inhomogeneities that affect a longer time period [Domonkos, 2008b]. The scenarios that these Domonkos papers created explored these characteristics, pointing out that small and short term inhomogeneities shouldn't just be included to see if algorithms can detect them, but because they will have an impact on the detection of more substantial or pervasive inhomogeneities [Domonkos, 2011, 2013]. The present study also captures and investigates these inhomogeneity characteristics in its created error structures.

The study of Della-Marta and Wanner [2006] explored daily homogenisation extremes, acknowledging that inhomogeneities likely affect more than just the mean of a temperature distribution. Therefore, they added multiple change points, but not always in the same way. Five different change points were possible, two affected only the mean, two also caused a change in variance and one was implemented as a change in skew-

ness. Although the current study does not explicitly alter higher order moments, these will be changed by the nature of having seasonally varying inhomogeneities that are implemented by changing inputs to the models.

The current project of the International Surface Temperature Initiative is likely to produce the most comprehensive range of error structures to date. This benchmarking study will incorporate error structures similar to those already investigated, whilst also expanding the range of investigation to incorporate inhomogeneities affected by other climate variables and a wider range of inhomogeneity sizes, frequencies and real world similarities [Willett et al., 2014]. It will also surpass the spatial scale of any previous study by quite some margin, thus enabling the assessment of algorithm performance in quite diverse climatic areas. The present study also assesses the interaction of temperature with other climate variables and will feed back into the work of the International Surface Temperature Initiative.

Studies have, of course, been carried out on other climatic variables too. Two of these which investigate homogenisation problems similar to those encountered in temperature data are Young [1993] and Titchner et al. [2009]. The first of these looked at sea level pressure data and investigated the impacts on inhomogeneity detection arising from increasing the variability of the series, altering series length and altering the location of the inhomogeneity. The second study looked at the temperature in the upper atmospheric layer and created four error models. These four error models were: a 'best guess' structure; many small change points; change points added in such a way as to remove the underlying climate change signal; and few large change points. They were designed to be as diverse as possible to avoid algorithm tuning. The present study has not sought diversity as its main focus, but instead changes only one main aspect of the data or error structure per scenario. The reason for this choice was that it allows the investigation of whether changing only one aspect can markedly change algorithm performance, which is beneficial when feeding back to the creators of algorithms.

2.3. Validation

The validation of homogenisation algorithm performance can be split broadly into two categories; change point detection ability and inhomogeneity adjustment ability, where the inhomogeneity is the effect of the change point. These two insights into a homogenisation algorithm's performance should be considered as complementary and not competitive. Studies have highlighted that good performance in one of these areas does not guarantee good performance in the other [Domonkos, 2011; Venema et al., 2012]. However, it is only recently that investigation has begun to be carried out into adjustment ability as well as detection ability. Assessing adjustment ability poses more problems for the benchmark creator as knowledge of exactly how different inhomogeneities affect temperature series is far from perfect. The same inhomogeneity will not have the same effects everywhere, for example station relocation effects will vary with topography and uniformity of climate.

2.3.1. Detection ability

The quantification of detection ability has been carried out using various measures in the past, commonly formed from contingency tables that comprise information on hits, false alarms, misses and correct rejections [Menne and Williams JR., 2005]. These four terms refer to the correct allocation of a change point, the false allocation of a change point, failing to allocate a change point where one should have existed and correctly not allocating a change point where one didn't exist; they are commonly represented by the letters a, b, c and d respectively. Arguably the most common measure used when assessing detection ability is the hit rate, $H = \frac{a}{a+c}$ [Hogan and Mason, 2012]. This measure is also referred to as the probability of detection [Menne and Williams JR., 2005] and the proportion of discontinuities identified [DeGaetano, 2006] in different studies. There are also occasions where it is wrongly referred to as the percentage correct, which in fact credits hits and correct rejections [Easterling and Peterson, 1995].

The hit rate credits algorithms for correctly locating change points and penalises them for missing true change points. A closely related measure to the hit rate is the correct change point power statistic (CRC), which also credits algorithms for not falsely inserting change points into truly homogeneous series [Menne and Williams JR., 2009]. A commonly used measure that is the opposite to the CRC in a sense is the type I error rate of an algorithm. The type I error rate is the proportion of truly homogeneous series that are made corrupt by the homogenisation process [DeGaetano, 2006; Menne and Williams JR., 2009; Yozgatligil and Yazici, 2015].

Hits are usually counted if a change point has been allocated within a certain window of the true change point. This window varies in length depending on the time scale of the study, but for monthly or annual series it is common to have it as around ± 2 time steps [Menne and Williams JR., 2005, 2009]. In the current study, at the daily level, the longest window considered for an allocated change point to be classified as a hit is 180 days, 90 days either side of the true change point; this is noticeably longer than ± 2 time steps in the knowledge that homogenisation algorithms are unlikely to be accurate within a time scale of four days.

Some studies define hits slightly differently, insisting that for an allocated change point to not count as a false detection it must have a similar magnitude and the same sign to the true change point and not just a similar location [Domonkos, 2011]. A further distinction was made by Ducre-Robitaille et al. [2003] who distinguished between a correctly identified change point (close in magnitude and exact in time) and a well identified change point (relatively close in both magnitude and time). A few studies have also assessed an algorithm's ability to identify the correct type of inhomogeneity out of a range of possible models. This was done by both Reeves et al. [2007] and Menne and Williams JR. [2009].

Given that the hit rate rewards correct change point allocation it is helpful to also have a measure that penalises an algorithm for inserting false change points. Various such measures exist, the two most common of these are the false alarm rate, $F = \frac{b}{b+d}$, also known as the probability of false detection, and the measure that it is consistently con-

fused with in the literature, the false alarm ratio, $FAR = \frac{b}{a+b}$ [Venema et al., 2012]. A review of 26 papers published in American Meteorological Society journals found that the terms false alarm rate and false alarm ratio were used incorrectly on 38% of occasions [Barnes et al., 2009]. Although the study of Barnes et al. [2009] did not look at hit rate terminology as well, this is also often incorrectly used as mentioned previously. The FAR is also sometimes referred to as the error rate [Easterling and Peterson, 1995]. A good algorithm should have a low false alarm rate and a high hit rate.

It is known that there are various confounding factors when assessing an algorithm's detection ability and some of these are commonly assessed in the literature. Many studies analyse the influence of a change point's size on the hit rate and conclude unsurprisingly that smaller change points are harder to detect [Ducre-Robitaille et al., 2003; DeGaetano, 2006]. Given that small change points are harder to detect some studies focus on larger change points only [Domonkos, 2011]. The variability of the series is also known to affect the detection rate, with more variable series being more difficult to homogenise [Young, 1993; Yozgatligil et al., 2011]. Change points closer to the ends of time series are also found to be harder to locate in general [Young, 1993; Yozgatligil et al., 2011] and as the frequency of change points increases detection also tends to become harder [McCarthy et al., 2008]. Lund et al. [2007] also look at the impacts of autocorrelation and periodicity on change point detection and find that both degrade an algorithm's detection ability, but autocorrelations have the far more noticeable effect. In the present study the impacts of series autocorrelation and inhomogeneity size will be explicitly assessed; by their nature, daily time series exhibit periodical (seasonal) behaviour so this can also be taken into account. Focus is not given to the location of change points within a series or their frequency, but further study in this area could be of interest. Another aspect that is assessed in the present study and was mentioned in DeGaetano [2006] is the effect of changing station density on change point detections.

Other measures relating to an algorithm's detection ability include the frequency bias of an algorithm, $B = \frac{a+b}{a+c}$, though it's debatable to what extent this analyses detection ability and not just rate of insertion [Menne and Williams JR., 2005] and also various skill scores [Hogan and Mason, 2012]. Skill scores compare the performance of an algorithm with that of some reference algorithm of no skill, e.g. one that allocates change points randomly [Venema et al., 2012]. Certain skill scores and measures do not involve the term d which is advantageous when this is difficult to define owing to the number of correct rejections far outweighing the hits, false alarms and misses. In the present study, the chosen 'non-d' measures are the frequency bias and the Critical Success Index, $CSI = \frac{a}{a+b+c}$, also known as the Jaccard coefficient or the threat score [Warrens, 2008; Hogan and Mason, 2012].

2.3.2. Adjustment ability

As one of the primary aims of homogenisation is to improve the temporal consistency of time series [Venema et al., 2012] it is valuable to examine the ability of homogenisation

algorithms to recover long term trends. This is the primary focus of the studies by Thorne et al. [2011a] and Williams et al. [2012], but is also highlighted in many other studies [Titchner et al., 2009; Domonkos, 2011; Venema et al., 2012]. It is commonly linear trends that are assessed, calculated using a least squares regression model on annual data. Linear trends are used as they are simple to compare and give an answer as to whether temperatures are going up or down overall at a glance. These trends are compared using measures such as the root mean squared error between clean and returned and clean and released trend coefficients [Menne and Williams JR., 2009; Venema et al., 2012]. Other measures looking at the improvement that homogenisation has created in the long term data trends are the percentage trend recovery [Willett et al., 2014] or skill of linear trend estimation [Domonkos, 2008a, 2011]. Focus has also been on the statistical significance of trends in the past, whether they were significant before or after corruption or homogenisation, and whether or not these significances were legitimate [Domonkos, 2008b].

Adjustment ability can also focus on general similarity measures between the clean and returned data. One of the earliest studies to assess adjustment ability in this way was that by Ducre-Robitaille et al. [2003] who looked at the sum of squared errors between the adjusted series and the known mean of the clean series and could therefore compare results to a known target value. An even earlier study, that assessed performance by analysing whether confidence intervals associated with adjustments incorporated the clean data, was that by Karl and Williams JR. [1987]. More recent studies have used error metrics to assess the similarities, these have included the RMSE [Della-Marta and Wanner, 2006] and the centred RMSE (CRMSE) [Venema et al., 2012]. The CRMSE is simply the RMSE computed on series that have first been centred by the subtraction of their means.

One final aspect of homogenisation algorithm quality that has thus far received relatively little attention is the homogenisation of moments higher than the mean. When shifts have been artificially applied to only affect the mean of a series this is almost acceptable, but in the real world variability and skewness of distributions could also be affected by the presence of inhomogeneities. Della-Marta and Wanner [2006] compared trends in extreme measures as they stated that mean focus is not sufficient. Communication within the homogenisation community has also highlighted that depending on the goal of homogenisation, different approaches should be considered as different parties are interested in different things. ”When trying to compute monthly means or trends, you do not worry about losing variance, but this would be highly inconvenient if you are interested in extremes” (pers. comm. J. Guijarro). Also, measures should be considered in tandem as, ”For daily data RMSE is likely not the best measure. RMSE quickly ’rewards’ methods that have too little variability around the mean” (pers. comm. V. Venema). This is one reason why RMSE was considered as only one of several measures in the present study and, as with all the other measures, an algorithm’s performance classification according to it shouldn’t be taken out of context.

As well as looking at different measures it is advantageous to look at the same mea-

sures aggregated over different spatial and temporal scales. COST HOME revealed that algorithms that did well on a station by station basis were not necessarily as effective when looking at regional series [Venema et al., 2012]. Equally, measures on different time scales can highlight different algorithm capabilities [Domonkos, 2013]. In the current study regional and station by station similarity measures were compared across algorithms. Longer term variability was also compared at different temporal scales.

When considering the evaluation of algorithm performances the quality of the benchmark data they are being applied to should also be taken into account. Test datasets that are more realistic should give a truer picture of how algorithms will perform in the real world, but fully knowing the characteristics of the real world in their entirety is not possible [Domonkos, 2013]. However, studies can gather much information from their focus region before creating synthetic data and it is hoped that these will then bear relatively good resemblance to reality [Domonkos, 2008b; Venema et al., 2012]. It is also advantageous to create blind studies where algorithm users do not know the properties of the underlying data, which reduces the chances of algorithms being tweaked to only perform well in specific circumstances [Willett et al., 2014]. The present study was a blind study that took information from real world stations and existing homogenisation literature.

2.4. Summary

This chapter has provided an overview of past studies that have sought to assess homogenisation algorithm performance, albeit on different spatial and temporal scales to the present study. It has highlighted both advantages and disadvantages of previous work and has hinted at how the present study will go about dealing with these issues. The following chapters all contain a brief review of the main points covered here that are relevant to the specific chapter, before proceeding to explain how the chosen methodology was implemented.

3. Data Analysis and Pre-Processing

Chapters one and two have given a brief overview of the motivation for this project and previous work in this area. They have explained that this project worked with daily data as opposed to the more commonly used monthly time series. This chapter begins with further information on the data that were used, where they were sourced from and how to access them. Details are provided on the data themselves, their spatial and temporal resolution and any necessary manipulations that took place in order to get them to the daily station level. Also included in this chapter is an explanation of why the four specific focus regions in North America were chosen.

3.1. Data Sources

Real world data were sought to be used in the modelling process to produce realistic, clean data benchmarks, against which returned data from homogenisers were evaluated. The benchmarks needed to closely replicate real world climate characteristics at the daily scale. Specifically, they needed realistic autocorrelations, cross-correlations, high and low frequency variabilities and long term trends. The data used had to be freely available, so that the work was reproducible; well documented, so that their source was known; and of a reasonably high quality, that is, having undergone quality control.

As stated in chapter one section two, daily mean temperature data at the station level were desirable as they allowed the reproduction of real station networks. Also, daily mean temperature station data provided a framework against which to broadly assess the created data's properties, such as their inter-station correlations and auto-correlations. In addition to the temperature data, variables that could aid the modelling of these data were sought. These variables included other climatic variables known to be related to daily temperature variations, and available at a similar scale, such as precipitation. Variables that could represent larger space or time-scale variability, such as El Niño Southern Oscillation (ENSO) events, were also sought in order to best match the real world data. A full list of the variables used in this model can be found in section 1.2 of chapter four.

The final data used came from the following three sources:

1. The Global Historical Climatology Network Daily (GHCND) Database - `https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/` [Menne et al., 2012a].

2. The National Oceanic and Atmospheric Administration (NOAA) 20th Century Re-

analysis - `http://www.esrl.noaa.gov/psd/data/gridded/data.20thC_ReanV2.monolevel.html` [NOAA, 2014].

3. The Australian Bureau of Meteorology - `ftp://ftp.bom.gov.au/anon/home/ncc/www/sco/soi/soiplaintext.html`.

3.1.1. GHCND

The GHCND database is from the National Oceanic and Atmospheric Administration's (NOAA's) National Climatic Data Center (now the National Center for Environmental Information). It fulfils the criteria of being at the daily station level, allowing the mimicking of real station networks. It has near global coverage and has been quality controlled, with those data that fail any quality control checks given flags allowing database users to identify them. A detailed description of this database including its coverage, sources, creation and quality control can be found in Menne et al. [2012b]. The main points of interest are summarised below.

Data characteristics

GHCND covers 180 countries and contains over 80 000 stations, though only one third of these contain temperature information, which is the variable of interest in this study. The database is formed from daily data with records varying in length from less than a year to over 200 years. The number of stations peaks in the 1960s, but remains at a relatively high level for temperature records, though it drops for precipitation data. The spatial coverage of stations that have any temperature records can be seen in figure 3.1. This figure shows that the coverage is good over most of Europe, North America and Australia; it is reasonable over Asia; and is relatively low over South America and Africa.

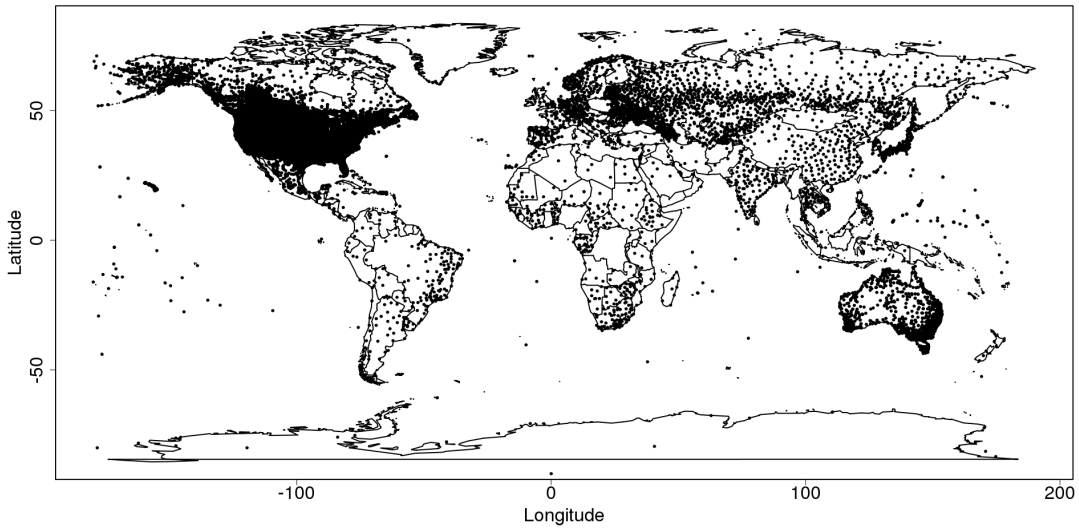

Figure 3.1. Location of GHCND stations with temperature records. No time constraint has been put on the length of these records and some may therefore be very short.

The database is regularly updated to ensure that records cover the present time period where possible and incorporate newly available data. This updating process is subject to

automated checks to ensure that any new stations that are replicas of existing stations are treated accordingly, so that the maximum amount of information can be used. All stations have their format checked to ensure that it complies with that of GHCND, and the data themselves are subject to 19 quality assurance checks detailed in Durre et al. [2010]. Examples of these checks include identifying cases where the maximum temperature has been recorded as being lower than the minimum temperature, or where the temperature recorded at one station falls significantly outside the range of temperatures at neighbouring stations. Occasionally additional checks are implemented that test the integrity of the records, these assess the records for artefacts such as repeated failure of earlier quality assurance checks [Menne et al., 2012b]. The data are not homogenised, but the creators hope they will be in a future version of the database.

Variables available

As stated above, daily precipitation is the most commonly archived GHCND variable. The other core variables which the dataset seeks to provide are minimum and maximum temperature, snowfall and snow depth. There are around fifty more variables reported only at certain stations; these include multiday maximum and minimum temperatures, average cloudiness and average daily wind speed.

For this work interest lies in mean temperatures. Therefore, as minimum and maximum temperatures are recorded in GHCND, the following formula for the calculation of the midrange, henceforth referred to as the mean, of daily temperature was used by the author:

$$TMEAN = \frac{TMIN+TMAX}{2}.$$

For any days where only $TMIN$ or $TMAX$ was recorded $TMEAN$ was set to be missing.

3.1.2. 20th Century Reanalysis

A reanalysis dataset is one that has been produced by combining past observations with a forecasting model to produce realisations of multiple climate variables that are physically consistent. Reanalyses offer data that are spatially and temporally complete and also allow a range of climatic variables to be obtained from a single source for long periods of time. These characteristics can make them preferable to observational datasets which can be short and incomplete. The disadvantage of reanalysis datasets is that they are gridded and are therefore at a much worse spatial resolution than observational datasets, which is one reason why the 20th century reanalysis (20CR) was used in conjunction with the other data sources specified in the introduction of section 3.1. Further information on 20CR can be found in Compo et al. [2011], the main points of which are summarised below.

Data Characteristics

The 20CR version 2 dataset, created by NOAA, contains data at the daily and sub-daily

scales from the 1st January 1871 to the 31st December 2012. The data are on an irregular Gaussian grid with a resolution fractionally better than two by two degrees. This dataset is created using surface pressure and sea level pressure records as these elements have been documented well since the late nineteenth century, or even earlier in places. A numerical weather prediction (NWP) model is used with 56 ensemble members to provide inputs for an Ensemble Kalman Filter Data Assimilation method that allows the creation of an analysis every 6 hours over the time period. This NWP model is fully parametrised and has boundary conditions defined using SST and sea ice fields from the Met Office Hadley Centre's HadISST dataset [Rayner et al., 2003].

In addition to these analyses the 20CR provides gridded forecasts every 3 hours for the same time period and it is predominantly these forecasts that were used in this thesis. The exception is the sea level pressure data which were taken from the analysis itself owing to them being one of the components used to create the rest of the data. The mean values of the ensemble forecasts were used in this study.

Given that 20CR is sourced from just pressure variables more can be assumed about its homogeneity than is possible for other reanalysis products. The creators even suggest it could be used to investigate inhomogeneities in observed time series showing their confidence in its homogeneity. Although, it should be pointed out that, Ferguson and Villarini [2012] do show that there is in fact an inhomogeneity present in 20CR for the United States region around 1950 owing to an increase in assimilated surface pressure observations, and they do not rule out further inhomogeneities. Ferguson and Villarini [2012] therefore recommend that climate studies should only use the data from this source from 1960.

Variables available

There are 35 variables available from the 20CR forecast dataset and many of these are available at different pressure levels. As this study is looking at station based observations the data from the surface level were used where possible. Data from the surface level were available for temperature, precipitation rate and downward solar radiation flux of the variables included in this model. Three variables incorporated into the model that were not available at the surface were wind speed (eastward and northward) which had to be taken at the 10m level (standard observing height for wind) and precipitable water content which is only considered in the atmosphere as a whole in the 20CR.

3.1.3. Southern Oscillation Index from the Australian Bureau of Meteorology

There are various phenomena in the climate system that are known to impact temperatures, but act over a longer time period than days. One of these phenomena is the El-Niño Southern Oscillation (ENSO). This is an ocean-atmosphere coupling which has its origins in the tropical Pacific, but can have impacts on variables including temperature on a much wider scale, see Marshall and Plumb [2008]. As such it is helpful to include a

measure of this phenomenon.

One such measure is the Southern Oscillation Index (SOI). This is a monthly index that can track the progression of ENSO events. There are various slightly different ways of presenting this index, but the one used in this study is that given by the Australian Bureau of Meteorology, as this is where the records of this index were sourced from. The index is defined as:

$$SOI = \frac{Pdiff+Pdiffav}{SD(Pdiff)},$$

where $Pdiff$ is the (average Tahiti mean sea level pressure (MSLP) for the month) - (average Darwin MSLP for the month), $Pdiffav$ is the long term average of $Pdiff$ for the month in question, and $SD(Pdiff)$ is the long term standard deviation of $Pdiff$ for the month in question. When the SOI has sustained negative values below -8 it is an El-Niño phase, when there are sustained positive values above 8 it is a La Niña phase. El-Niño is traditionally referred to as the warm phase of ENSO, but can still have cooling effects on temperatures. For example, in the South Western United States an El-Niño phase is generally associated with cooler than normal temperature anomalies [Wang et al., 2013].

The Bureau caution that daily values of the SOI should not be trusted as they fluctuate too much because of weather conditions, as such the SOI data are provided at the monthly level. These trustworthy values were then interpolated to the daily level for this study using the method detailed in section 3.2.2.

There are other phenomena that vary on a longer time scale and affect the atmosphere, however it was deemed sufficient to have only one index of such events, especially when the chosen phenomenon can have impacts on such a global scale.

3.2. Source Data Processing for Benchmark Data Creation

The sparsity of data in some regions is one reason why this study is not global. Other reasons include the computational and climate complexity involved in global modelling and the lack of algorithms that could comfortably cope with this amount of data. This section details the selection of the regions that were focussed on and also the focus time frame. It also details the temporal and spatial interpolation of the source data necessary to provide the variables that are to be used in the modelling process at the daily station level.

3.2.1. Temporal and regional focus areas

Focus time frame

GHCND station coverage is best in the 1960s and remains at approximately this level for the rest of the record for temperature stations [Menne et al., 2012b]. Reanalysis quality is best where most pressure observations are available, which is also in the more recent

time period [Compo et al., 2011]. It is therefore sensible to make the focus time frame the later half of the 20th century and the beginning of the 21st century.

The exact time frame selected was 1st January 1970 to the 31st December 2011. This time frame is long enough that inter-annual and inter-decadal artefacts will be incorporated in the data, but short enough that data manipulation is relatively straight forward. The record stops in 2011 as this was the last complete year available when this study began.

Within this time frame it was decided that for a station to be considered for inclusion in the modelling process it must be at least 75% complete. This restriction was imposed as a cautionary measure as those stations with large amounts of missing data could also be less reliable, which could in turn bias the modelling process. Also, many algorithms aggregate data to the monthly level and the WMO advises that this should only be done if months are at least 83% complete [WMO, 1989]. A desired 75% completeness for the entire series was therefore a tighter restriction overall, though individual months could still be less complete than this, but it was hoped to ensure that on average no benchmark data released would have to be discarded because of insufficient records. The modelling framework itself, detailed in section one of chapter four, is capable of handling missing data and of reproducing realistic and complete stations regardless of whether they were present in the model building process. Therefore, these data restrictions are acceptable and should not be detrimental to the study.

As high quality data are preferred for this project and GHCND provide quality flags for data that have failed any of the quality assurance procedures detailed in Durre et al. [2010] any flagged data were made missing. Therefore, from this point 'missing data' refers to both data that were missing in the original dataset and data that were made missing when the data were being processed for use in this research.

Focus regions

Not all stations shown in figure 3.1 contain temperature records for the time period 1970 to 2011. Stations that do contain any data between 1970 and 2011 are highlighted in red in figure 3.2 and this shows that the regions with the best station coverage are the contiguous United States and Japan. However, calculations on the station data revealed that in Japan only 5 out of 136 stations satisfied the criteria of being at least 75% complete.

Figure 3.3 shows all GHCND temperature stations in the contiguous United States, with stations that are at least 75% complete over 1970 to 2011 shown in red or blue. The blue stations are those in the focus regions for this project. This figure shows that even with temporal data restrictions there are still many stations available for use in the modelling process. The exact number is 3373 out of 3984 stations. It would have been possible to create a model using all available stations, but the United States incorporates multiple different climate regimes and focusing on certain regions in more detail allowed exploitation of this fact. For this reason four regions were chosen; Wyoming, the South East, the North East and the South West, these are the regions with their stations highlighted in blue in figure 3.3. The climates of these regions are outlined below, and these outlines

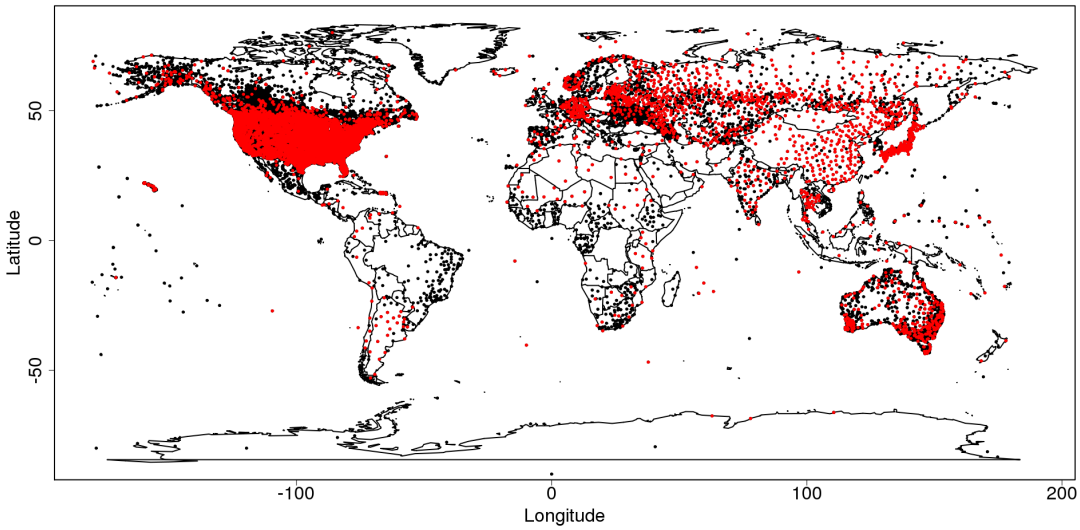

Figure 3.2. Location of GHCND stations with temperature records. Stations in red indicate stations with any temperature records in the period 1970 to 2011.

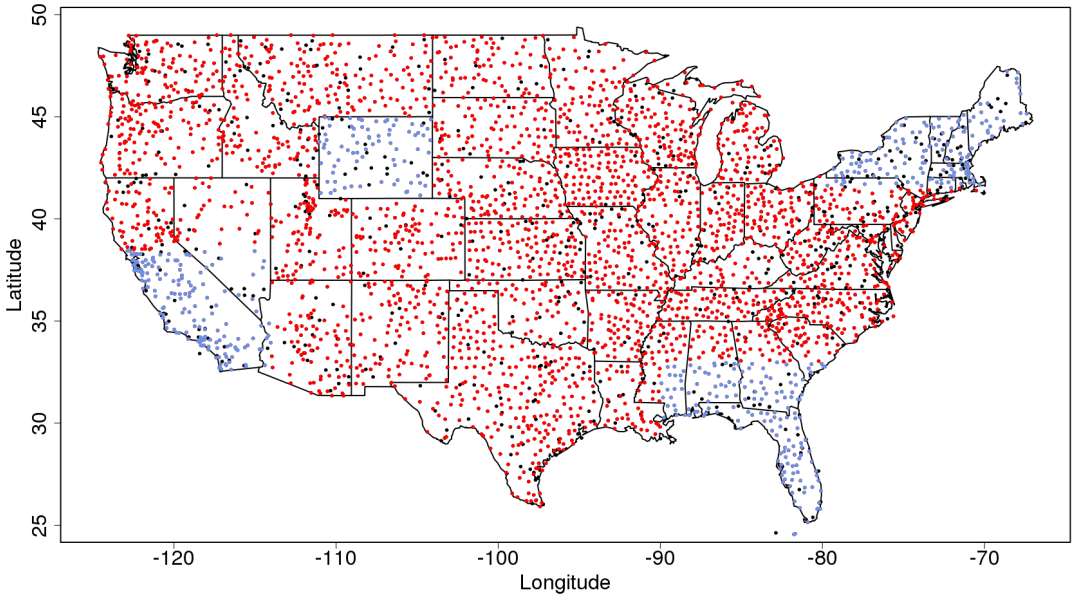

Figure 3.3. Location of GHCND stations with temperature records for the contiguous United States, those which are at least 75% complete in the period 1970 to 2011 are highlighted in red. Those which are in the focus regions, and at least 75% complete, are highlighted in blue instead.

were largely sourced from the North America Climate entry in the Encyclopedia of World Climatology [Corcoran and Johnson, 2005].

Wyoming is a mid western state, characterised by distinct seasons and a relatively dry climate with the majority of its precipitation occurring in summer through variably distributed convective showers. It was chosen as a focus region particularly because it has no sea borders, thus eliminating one aspect of variability. It is relatively diverse topographically, with the Rocky Mountains in the west of the State offering a mountain climate where extra variability is added by rapidly changing station altitudes and mountains imposing barriers to wind and precipitation. The distribution of temperatures in the stations used for modelling in this region can be seen in figure 3.4a and their standard deviations can be seen in figure 3.5a. These plots illustrate the seasonality of temperatures in Wyoming and the

increased temperature variability in winter.

The South East incorporates all or part of the following states: Florida, South Carolina, Georgia, Alabama, Mississippi and Louisiana. The very tip of Florida could be classed as having a tropical climate while the rest of this region is classified as subtropical. As this region is closer to the equator than Wyoming the seasons are much less distinct. It is a hot and humid region, though this is less pronounced in the east owing to a greater amount of solar energy going into evaporation and not heating. Summers are long and hot and also the time of most precipitation, which is mainly in the form of convective showers. Winters with mean monthly temperatures above freezing are common. This region's temperature distribution can be seen in figure 3.4b, its lack of bi-modality illustrates the lack of distinct seasons in this area, but figure 3.5b shows that there is still a strong seasonal cycle in temperature variability.

The North East incorporates all or part of the following states: Maine, New Hampshire, Vermont, Massachusetts, New York, Rhode Island, Connecticut and Pennsylvania. These states have a snow climate, and monthly average temperatures below freezing are not uncommon [Corcoran and Johnson, 2005]. The region is not too dissimilar to Wyoming, having distinct seasons and a wide annual temperature range, the widest in the United States. The biggest difference from Wyoming is the coastal border that the North East has with the North Atlantic Ocean meaning that although in the west of the North East region the majority of precipitation falls in the summer this is not the case in the east of the North East region. For the entire North East region frontal precipitation is the most common form, which is not the case in Wyoming, though convective showers do occur more towards the south of the North East. The North East's similarity to Wyoming can be further seen by comparing figures 3.4a and c and 3.5a and c. These show that the temperature distributions and variability are very similar in these two regions.

The South West incorporates the southern half of California and Nevada and the very western most stations of Arizona. Unlike the other focus regions, which are a mixture of at most two climates, the South West could be argued to incorporate five different climates even in this relatively small region. There is a strip in its centre that shares a similar climate with Wyoming; both the dry climate and the mountain climate, but it also incorporates desert climates in the east where high temperatures and low humidities are the norm and timings and amounts of precipitation are highly variable. Here cyclonic storms give winter precipitation while monsoon systems give the summer precipitation and the deserts themselves see convective showers. In the west it has two coastal climates; the north west of the region has cooler, damper summers than the rest of the west of the region which has warmer, drier summers, with only about 5% of the year's precipitation falling in this time. Winter rains are predominantly caused by frontal systems. Figure 3.4d illustrates the regional temperature distribution here, but investigation of individual station distributions reveals a lot of variability in their shape. The yearly variability in temperatures for the region also differs from other regions as can be seen in figure 3.5d.

Figure 3.4 reinforces the differences between the four focus regions. It can be seen that plots (a) and (c), for Wyoming and the North East respectively, exhibit wide temperature

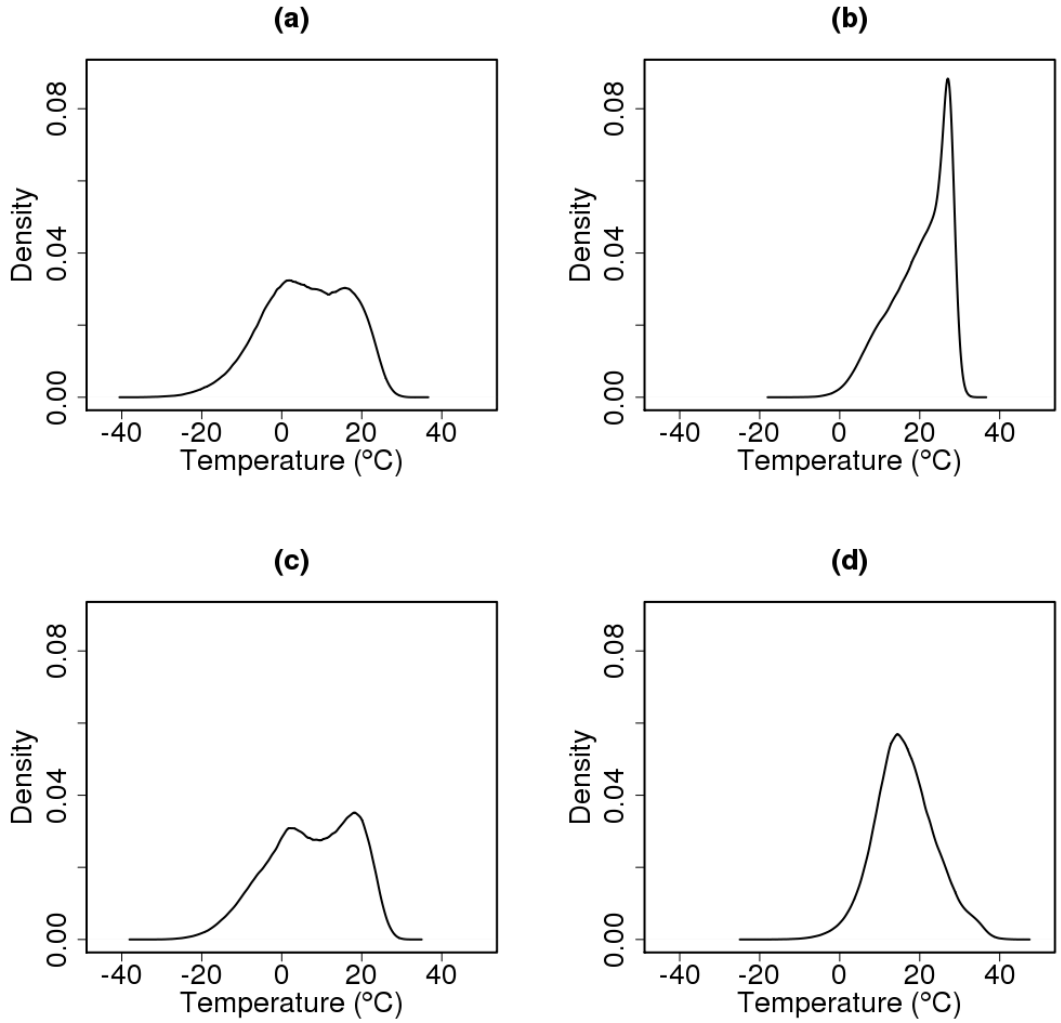

Figure 3.4. Density distributions of calculated mean temperatures in each of the four focus regions. (a) Wyoming, (b) South East, (c) North East and (d) South West. Axes were constrained to be the same for all regions to allow direct comparisons. All stations that were 75% complete over 1970 to 2011 contributed data to these plots.

ranges and are almost bimodal indicating distinct seasons. The two southern regions are unimodal and have narrower temperature ranges. All four distributions exhibit negative skew (long negative tails) indicating the non-Gaussianity of daily temperature data in these regions.

Figure 3.5 shows how the standard deviation (the square root of the variance) of temperature varies throughout the year in each of the four focus regions. Clearly the variability of temperature is non-constant with larger variances being seen in the winter for Wyoming, the South East and the North East and in the summer and the winter for the South West.

Interest in stations extends beyond their means and variabilities. In homogenisation inter-station correlations, station autocorrelations and difference series autocorrelations are all of interest too.

Inter-station correlations are important as many homogenisation algorithms will use neighbouring stations to determine the location and magnitude of inhomogeneities. The higher inter-station correlations are, the easier it should be to find inhomogeneities and the lower

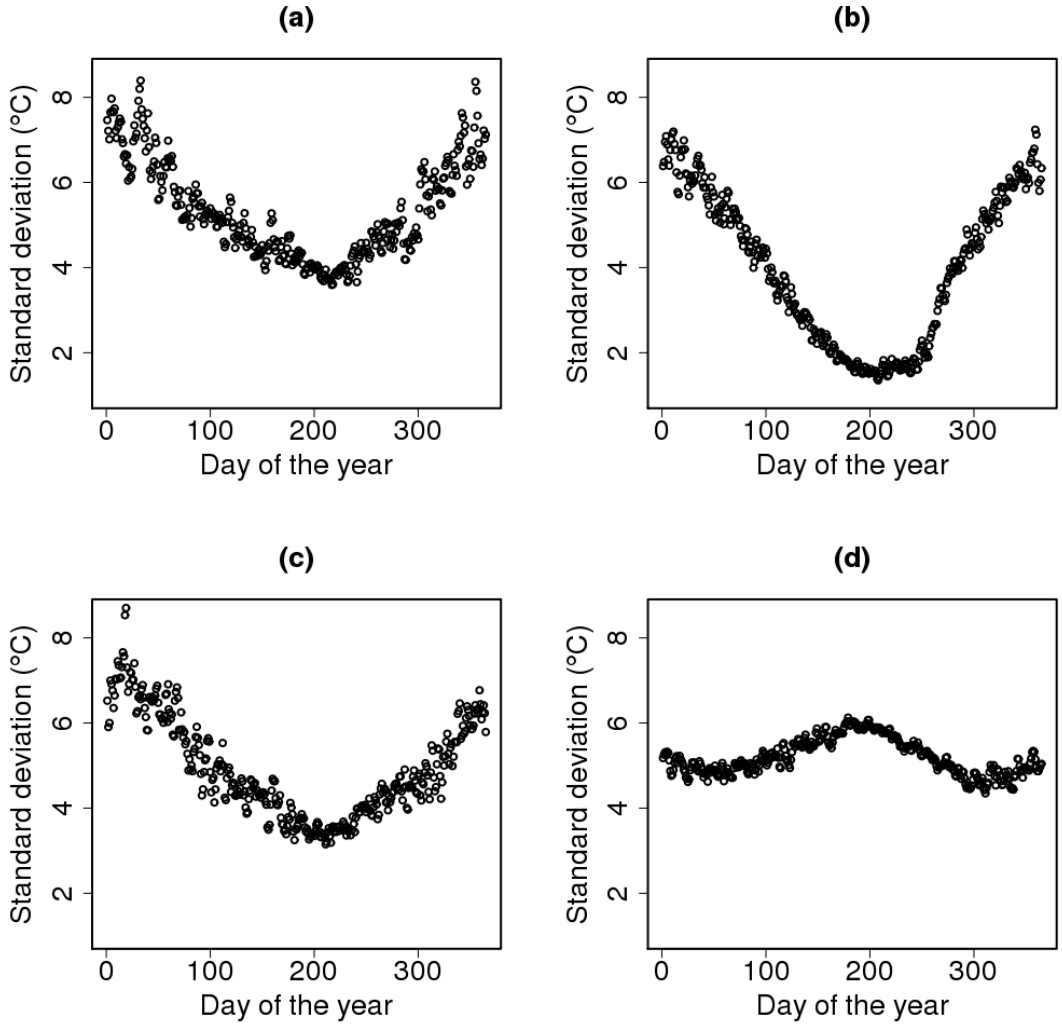

Figure 3.5. Scatter plots to show how the standard deviation of mean temperatures varies over the year in each of the four focus regions. (a) Wyoming, (b) South East, (c) North East and (d) South West. Axes were constrained to be the same for all regions to allow direct comparisons. All stations that were 75% complete over 1970 to 2011 contributed data to these plots.

they are the harder it is expected to be [Williams et al., 2012]. Figure 3.6 shows the density distribution of observed inter-station correlations in each of the four focus regions. It can be seen that stations in Wyoming show fewest low inter-station correlations, influenced undoubtedly by the fact that the maximum separation of stations in this region is not as large as the maximum separation between stations in other regions. More low inter-station correlations are seen in the South West, which, as stated above is the most climatologically complex region. Inter-station correlation density distributions in the North East and South East are relatively similar to each other. From these figures it would be expected that, in the real world, Wyoming would be the easiest region to homogenise and the South West would be the hardest. These levels of difficulty in terms of inter-station correlations are reproduced in the created data, as can be seen in figures 4.14 and 5.9.

Autocorrelations are of interest in benchmarking studies as the stations being created need to look and feel sufficiently like the real world for the conclusions drawn from the analysis to be generalisable to observed climate series. Before calculating autocorrelations stations should be deseasonalised to ensure that the seasonal cycle is not domi-

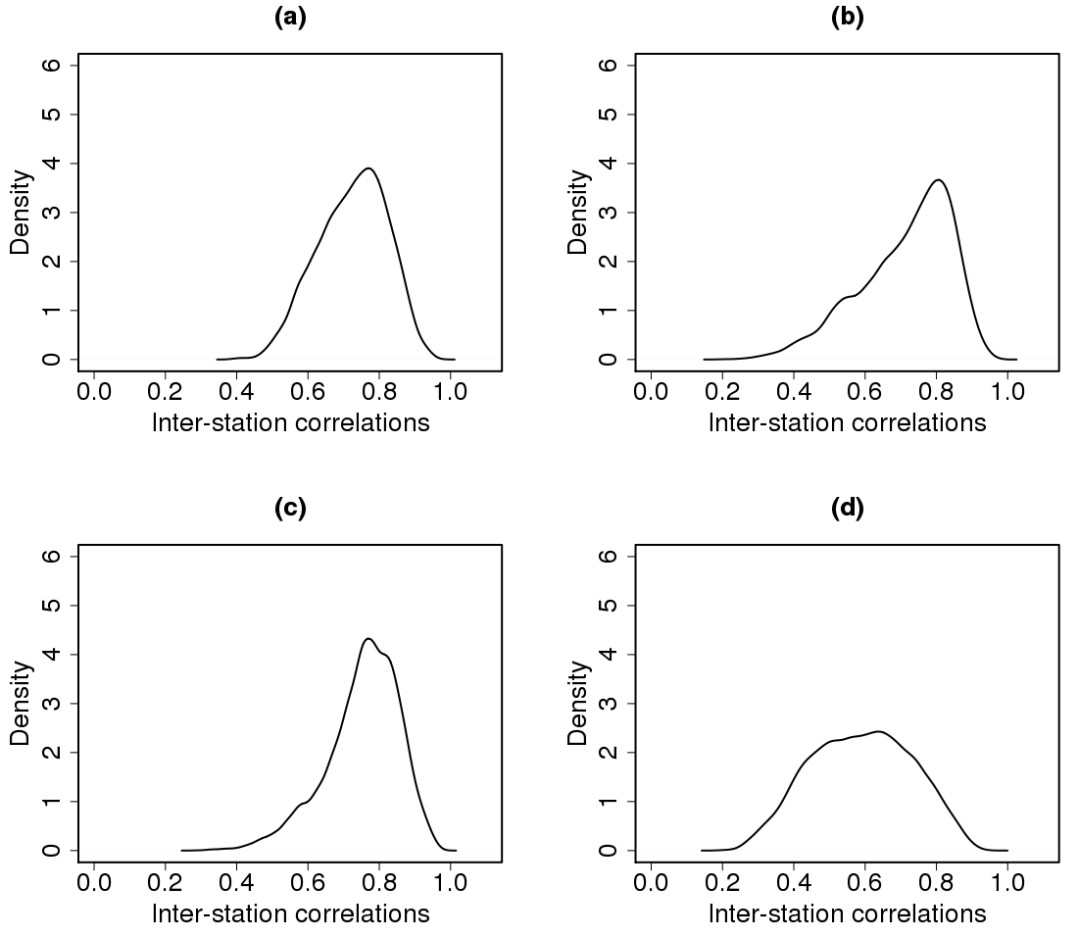

Figure 3.6. Density plots of the inter-station correlations found in the observed temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Axes were constrained to be the same for all regions to allow direct comparisons.

nating the results.

Figure 3.7 shows the autocorrelations in regional average series for Wyoming, the South East, the North East and the South West. These regional average series were created by taking the mean of all deseasonalised values for each day of the time series for each region separately. It can be seen that in all regions the autocorrelation drops below 0.1 relatively quickly; this value was chosen relatively arbitrarily as the autocorrelation of interest cut off point in this study. The similarity in shapes of the regional autocorrelation plots in Wyoming (a) and the North East (c) again emphasises that these two regions are the most similar. The autocorrelations tail off a little quicker in these two regions than in the South East or the South West. The author believes this to be due to the prevailing winds in the United States. For the majority of the United States the prevailing winds blow from the west [Ahrens, 2000]; for the South West and the South East winds from the west often mean winds from the coast, whereas for the North East winds from the west come from the land. As oceans have a longer memory of temperatures than land this can explain the resultant increased autocorrelations in the South East and South West relative to those that are seen in the North East or Wyoming.

Often, instead of working just with deseasonalised series algorithms will work with deseasonalised difference series. That is, a series created by subtracting one deseasonalised

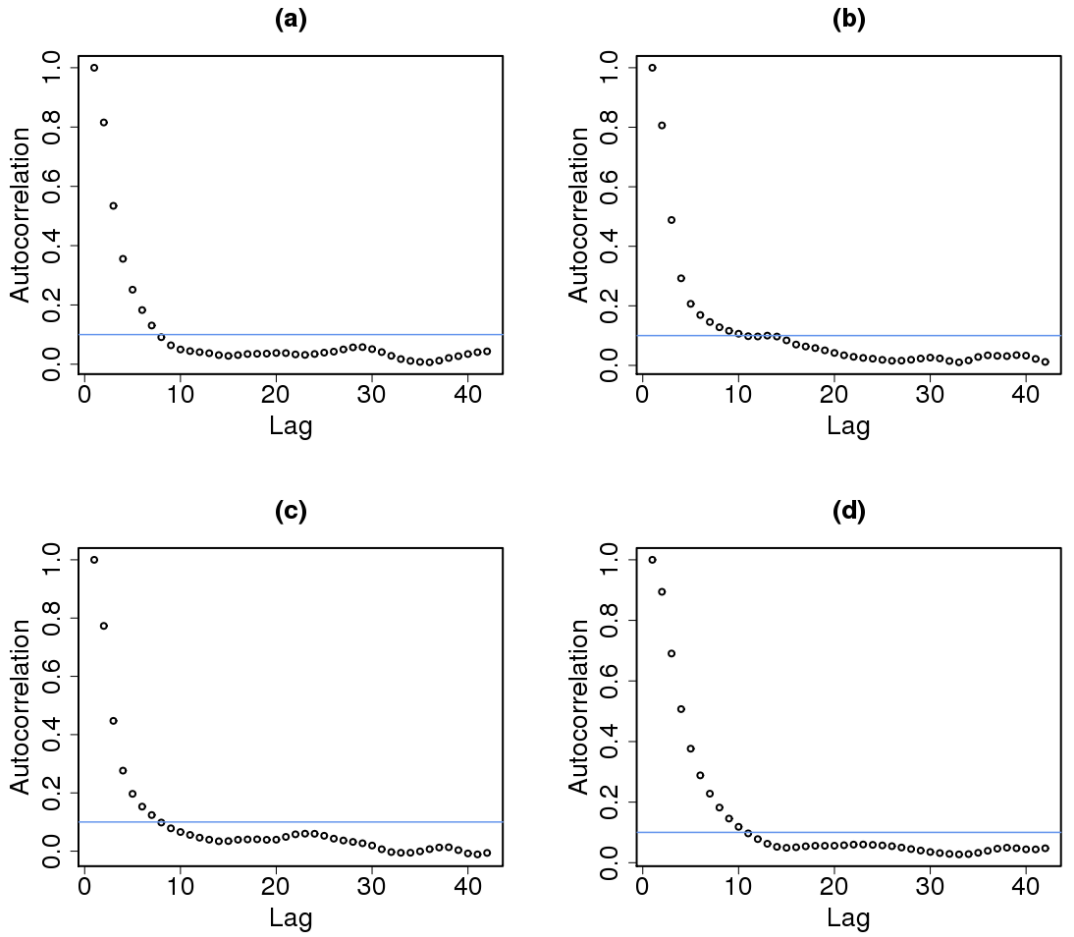

Figure 3.7. Plots to illustrate the autocorrelations found in the regional average series of observed temperatures for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West.

series from its most highly correlated neighbour, which has also been deseasonalised. The primary reason for doing this is that it removes another level of variability from the series being assessed, based on the assumption that nearby stations will show similar climate variations. Removing more variability from a series should make detecting inhomogeneities easier as the signal (inhomogeneity) to noise (background variability) ratio should be greater.

Figure 3.8 shows the average autocorrelations at each lag for deseasonalised difference series in each of the four regions. The average autocorrelation at each lag has been determined by working out the autocorrelation at each lag in all difference series, these are the difference series which have been created by differencing each station and its most highly correlated neighbour, and then taking the mean of these autocorrelations at lag one, lag two and so on. It can be seen that even after differencing there is still a noticeable amount of autocorrelation in the observations. This is of interest as many algorithms assume that deseasonalised difference series will be white noise. Also note that, once more, the autocorrelations are more persistent in the South East and South West than in Wyoming and the North East.

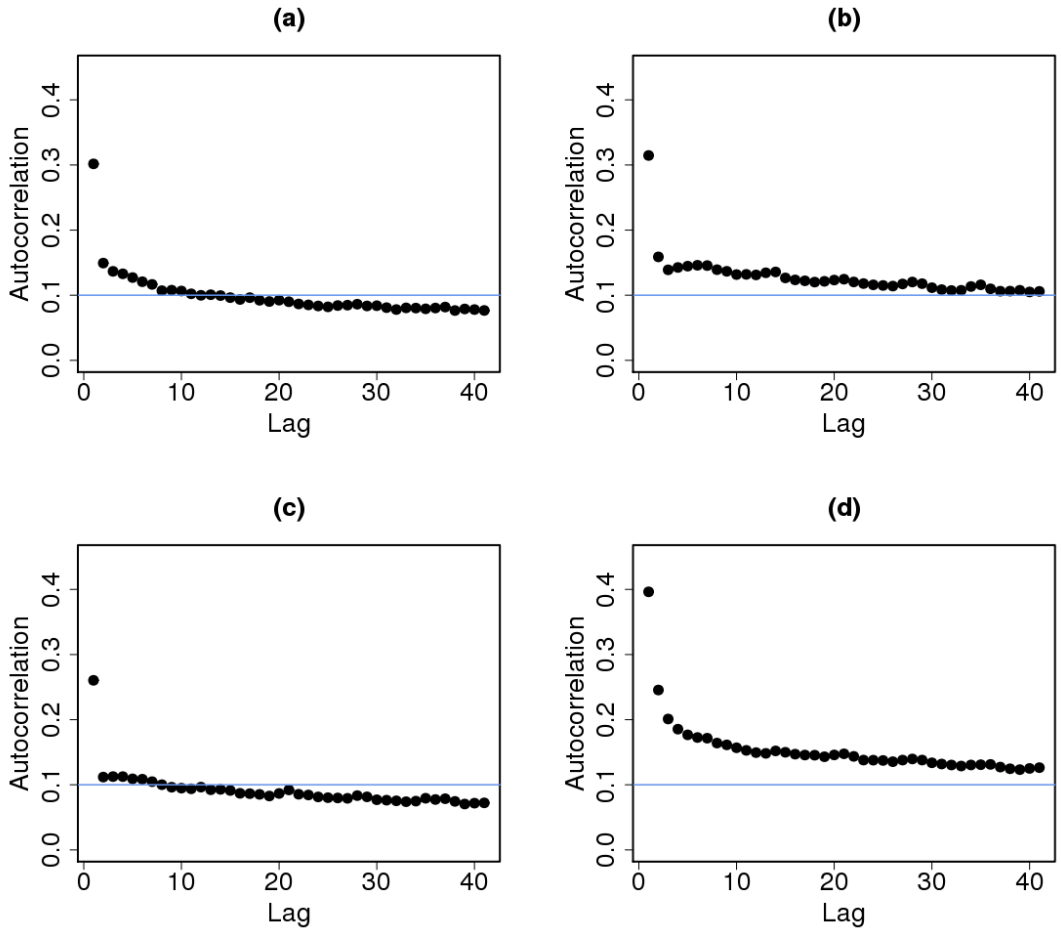

Figure 3.8. Plots to illustrate the average autocorrelations found in the deseasonalised difference series of observed temperatures for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Each series in a region was differenced according to its most highly correlated neighbour and the mean autocorrelation at each lag was then calculated across all stations in each region.

3.2.2. Interpolation

As stated in section 3.1 not all the data are available at the daily station level.The 20CR data downloaded are daily gridded data and the SOI is a monthly index and is the same everywhere in space. This section therefore gives the methods of spatial and temporal interpolation used for this project.

Spatial Interpolation

The `interpp` function in the R package `akima` was used for the spatial interpolation of 20CR data from the gridded level to the station level. More information on this function can be found in the R documentation on `akima`, or through the original paper about the function [Akima, 1978]. `interpp` is an interpolation method able to take regularly or irregularly gridded data and interpolate to irregular points. It is not able to cope with missing values, but this is not a hindrance when the data are sourced from 20CR as they are given as complete fields. The command to use this function in the current study took the following form:

call=interpp(ReanalysisLats,ReanalysisLons,ReanalysisVar,PredLats,PredLons,

linear=TRUE, extrap=FALSE, duplicate = "error").

Here, $ReanalysisLats, ReanalysisLons$ and $ReanalysisVar$ were the values of the latitudes, longitudes and variable being interpolated from the 20CR grid. $PredLats$ and $PredLons$ were the values being predicted to and the command $call\$z$ gave the values of the variable of interest at the specified prediction locations. A buffer zone was included around the edge of each of the regions to minimise the edge effects of the interpolation and some example plots were created to ensure that the method was behaving as expected. No extrapolation beyond the reanalysis points was allowed and this is indicated by $extrap = FALSE$ in the above line of code. $linear = TRUE$ indicates that linear interpolation was used. If $linear = FALSE$ then spline interpolation would be used. Spline interpolation would be a valid approach, and splines were employed for interpolation elsewhere in this study, as detailed in section 4.2.2. However, linear interpolation was deemed appropriate here; it constrains extremes to only occur at given data points, which could be considered an unnecessary restriction, but does ensure that no unrealistic values can be produced from this process. Splines would still be unlikely to produce unrealistic values and, therefore, future work could investigate if better interpolations were achieved when using splines, but this investigation was not carried out in the present study. The final option defined in the above code was $duplicate = "error"$. No duplicate prediction points should have occurred in this work as the interpolation was to distinct stations, but this was a valid safety net to protect against input data errors. If duplicate data were encountered then the command would report an error.

Interpolating gridded data will not achieve the same level of inter-station variability as is found in true station data because there is less information available. However, interpolated data do provide more variability than data that are just taken from the nearest grid point. Interpolation outputs will be better for variables that are known to have higher spatial correlations, such as temperature, than lower spatial correlations, such as precipitation. This is especially true in areas where precipitation is known to fall as convective showers which are much smaller than the near two by two degree grid boxes of the 20CR.

This spatial interpolation method was used as the code was simple to implement and able to take gridded data and interpolate to irregularly spaced points. Alternative methods of spatial interpolation are available, such as the commonly used geostatistical method, kriging, see Cressie and Wikle [2011], but the capabilities of `interpp` were deemed sufficient for this project.

Temporal Interpolation

SOI is interpolated from monthly data to daily data using a simple linear interpolation. This takes the values of SOI for each pair of consecutive months, fits a line between the two points and then reports the values at daily intervals on that line. This method of interpolation was deemed preferable to reporting a constant value of SOI for each month, as a constant value could change suddenly at the beginning of the next month and, depending on the influence of SOI, this could be mistaken as an inhomogeneity.

3.3. Discussion

In this chapter many decisions were made that could be altered to expand this benchmarking study or provide an ensemble of benchmarks. For example, a different reanalysis dataset could be used, or observed precipitation values could be taken from GHCND instead of from the 20CR dataset. Neither of these possibilities were investigated here as the author believes the final choices of using 20CR data were most beneficial to the model building process owing to its completeness and greater homogeneity. However, using different data could in the future provide insight into how much the choice of underlying data impacts how realistic the benchmarks are.

Changing the choice of interpolation method when downscaling data from the 20CR to specific station locations could also be investigated in a future study. For example, non-linear interpolation could be used so that extremes could be located at positions other than the grid points, though this has the disadvantage of allowing the small possibility of unrealistic extremes as mentioned in section 3.2.2. Alternatively a spatial interpolation method could be used.

A further extension to the study would be to investigate more regions of North America, or, indeed, the world. Figures 3.4 to 3.8 show that the focus regions chosen exhibit noticeably different data characteristics and the following chapters show how well the chosen model is able to reproduce these characteristics. It would be an interesting investigation to try modelling a region where there is essentially no seasonal cycle and/ or a very predictable weather regime to see how both the models and the homogenisation algorithms coped in such circumstances.

3.4. Summary

This chapter has described the data used for this project, their sources, and the methods used to acquire them at the daily station level. It has explained why the four specific regions in North America were chosen as focus regions and has also given the reasoning for the choices of time period and level of temporal completeness. The following chapter will explain how these data were used to create clean daily benchmark time series and the models used to form these series.

4. Creation of the Benchmark Clean Data

The first two chapters of this thesis gave the motivations behind this research project and reviewed previous work in the area. Chapter three introduced the sources of data used in this study and the necessary pre-processing steps that they had to undergo to make them suitable for inclusion in the models to be used to produce realistic daily temperature data. This chapter explains the formulation of these models, including justifications for why the specific variables used were chosen. It closes with details of the predictions made from the final chosen models and the post-processing of these predictions that took place to ensure high quality benchmarks that adequately matched reality were produced.

4.1. Modelling Methods for Daily Temperature Data

As stated in section 2 of chapter two, in the past, models to produce synthetic temperature data for benchmarking studies have primarily focused on monthly or annual data. Three of the main benchmarking studies to date have been those of Venema et al. [2012], Williams et al. [2012] and Willett et al. [2014], the last of which is currently at the benchmark data creation stage. All of these studies focus predominantly on monthly data. Venema et al's approach was not feasible for this study owing to a lack of the necessary homogeneous series used to create the clean data. The method of Willett et al. [2014] would perhaps be feasible for daily temperature data, but has the capacity to be very computationally expensive when working with daily data and was not finalised when this study began. The study of variants of the pairwise homogenisation algorithm by Williams et al. [2012] allowed the creation of realistic networks on a large scale in the contiguous United States. However, the Williams et al. [2012] study also required homogeneous station information and would likely struggle to reproduce daily variabilities. This is because it downscaled gridded data with white noise and climatological offsets to create station time series, though these did match observed inter-station correlations at the monthly level. Other methods detailed in the literature review had similar drawbacks to these main studies.

For these reasons a new approach was sought. This new approach needed to have as many of the desirable criteria of previous studies, with as few of the drawbacks, as possible. It needed to be able to create realistic homogeneous benchmarks without the requirement of homogeneous input data and be able to handle incomplete or short data records. It needed to be able to cope with data at the daily resolution without exceeding computational capacity and be a directly reproducible method. It had to produce data

with realistic inter-station correlations and autocorrelations and ideally be able to make use of all the data available to it when creating both clean and inhomogeneous series.

The method adopted in this study, which is described below, sought to match as many of these criteria as possible. It used a statistical model as the basis for data creation to allow other climatic variables to impact the predictions of temperature. Using a statistical model also met the criteria of being able to cope with short or incomplete records as it allowed the prediction of temperatures at unobserved locations as long as some input information was available, which, in this case, it always was. A brief introduction to statistical models is given here to aid understanding of later model explanations. As more approaches for modelling daily temperature become available, be they statistical or climatological, they could be compared with this study to enable further quantification of algorithm performance.

4.1.1. Modelling Framework: The Generalised Additive Model

Generalised Additive Models (GAMs) are a more flexible extension to the more commonly known Generalised Linear Models (GLMs) which are in turn an extension of the linear model. The reasons for using a GAM in this study largely relate to its allowance of non-Gaussian response variables and the flexibility it allows in the incorporation of explanatory variables. The following paragraph gives a brief introduction to statistical modelling terminology and statistical models in general before overviews of the two simpler modelling approaches, and the reasons why they were not suitable for this study, are given. There follows an explanation of the GAM in more detail, which leads on to details of the selection of the specific GAM used for this work. Throughout this section Simon Wood's book on Generalised Additive Models in R, [Wood, 2006], will be the primary reference.

A statistical model is concerned with modelling a response variable $\mathbf{y}= (y_1, ..., y_n)^T$, which has some probability distribution $p(\mathbf{y}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of the model parameters. For example, $\boldsymbol{\theta}$ could be the model's mean and variance, which could in turn be modelled by explanatory variables, x_i, that are related to the response variable in some way. When using a statistical model the $\boldsymbol{\theta}$ will normally need to be estimated. The driving force behind this estimation process is the likelihood, $L(\boldsymbol{\theta}) = p(\mathbf{y}; \boldsymbol{\theta})$. When $p(\mathbf{y}; \boldsymbol{\theta})$ is viewed as the likelihood function it is treated as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{y}$ as it represents how (relatively) likely different values of $\boldsymbol{\theta}$ are for the observed data $\mathbf{y}$. If you have fixed values of $\boldsymbol{\theta}$ and are looking at $p(\mathbf{y}; \boldsymbol{\theta})$ as a function of $\mathbf{y}$ then it is a probability distribution, as already stated. The values of $\boldsymbol{\theta}$ that maximise the value of $L(\boldsymbol{\theta})$ are the maximum likelihood estimates, $\hat{\boldsymbol{\theta}}$, which can be found by setting the first derivatives of $log(L(\boldsymbol{\theta}))$ to zero. ($log(L(\boldsymbol{\theta}))$ tends to be used instead of $L(\boldsymbol{\theta})$ as it tends to be easier to work with and will give the same $\hat{\boldsymbol{\theta}}$ estimates). These estimates have desirable statistical properties including being consistent (varying less from the true $\boldsymbol{\theta}$ as sample size increases) and being asymptotically unbiased and efficient (varying less than other estimates for the same sample size). If one then takes the second derivative of $log(L(\boldsymbol{\theta}))$ then it is possible

to also form confidence intervals for the maximum likelihood estimates.

The Normal Linear Model (LM)
A normal linear model, henceforth referred to as the linear model, in statistics is a model where **y** is assumed to be normally distributed with a mean, $\boldsymbol{\mu}$, and constant variance, σ^2. For this model the mean is itself represented using a linear combination of parameters $\boldsymbol{\beta}$ multiplied by explanatory variables, **x**. The model can therefore be written in the form $\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2)$ where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and **X** is the model matrix that contains the values of the explanatory variables (or functions of them).

The assumption of linearity in the parameters means that although functions of explanatory variables can be included, they cannot incorporate the $\boldsymbol{\beta}$'s. For example, the term $\beta_1 sin(x_i)$ would be acceptable, but the term $sin(\beta_1 x_i)$ would not be. Further information on linear models including the method of fitting can be found in chapter one of Wood [2006]. The method of fitting revolves around the idea that the $\boldsymbol{\beta}$'s should be chosen to minimise the squared difference between the predictions and the observations, that is, they should be chosen to minimise the least squares function, $S = ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$. The values minimising S will be those values maximising the likelihood, $L(\boldsymbol{\theta})$.

It is evident from figures 3.4 and 3.5 that neither the normality, nor the constant variance assumption of this model hold for the daily temperature data in the focus regions of this study. It is often the case in climate studies that temperature data have their seasonal cycle removed and are standardised by dividing by their standard deviation to try and better match these assumptions. However, this is removing information that could instead be exploited by a more sophisticated model.

The Generalised Linear Model (GLM)
The generalised linear model (GLM) is more flexible than the linear model, allowing the response to be non-Gaussian. In fact, the response can have any distribution in the exponential family of distributions, where any distribution in this family can be written as

$$p\left(y;\theta,\phi\right) = \exp\left\{\frac{\left(y\theta - b\left(\theta\right)\right)}{a(\phi)} + c\left(y,\phi\right)\right\},$$

where a, b and c are functions, ϕ is a scale parameter and θ is the location, or canonical, parameter of the distribution [Wood, 2006]. The mean and variance of any distribution in this family can be written as $\mu = b'(\theta)$ and $\sigma^2 = b''(\theta)a(\phi)$. Many well known distributions can be written in this manner, including the Normal, Binomial and Gamma distributions. Thus, the linear model is a special case of the generalised linear model. Using a slightly different fitting approach the range of distributions can be extended to anywhere where the mean-variance relationship is known, for an explanation of this the reader is referred to section 2.1.10 in Wood [2006] on quasi-likelihood.

The assumption of linearity in the parameters is relaxed to some extent for the GLM as the model is now written as $\mathbf{y} \sim p(\mathbf{y};\theta,\phi)$ and a link function is defined as $g(\boldsymbol{\mu}) = \boldsymbol{\eta} = X\boldsymbol{\beta}$, which relates the mean of the chosen exponential family distribution to the explanatory

variables. Thus, the β's have to enter the link function linearly, but this link function could be non-linear as long as it is smooth (differentiable) and monotonic.

The equations it is necessary to solve to find the parameter estimates of a GLM normally require an iterative solution instead of the straightforward least squares estimation method used for the linear model. Therefore, a method known as iteratively re-weighted least squares (IRLS) is employed to get the maximum likelihood estimates of the β's. A detailed explanation of this method is given in chapter 2 of Wood [2006], an overview, heavily relying on this source is given here.

The least squares function that must be minimised is derived from the log likelihood of the model and can be written as $S = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$; the derivation of this expression can be found in section 2.1.2 of Wood [2006]. $V(\mu_i)$ is known as the variance function of a GLM and is equal to $\frac{b''(\theta_i)}{w}$ and w is a known constant. In matrix form this can be written as $S = ||\sqrt{V_{[k]}}^{-1}[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta})]||^2$ where **V** is a diagonal matrix with $V_{[k]ii} = V({\mu_i}^{[k]})$ and k denotes the iteration number. Using Taylor expansions of $\boldsymbol{\beta}$ and introducing the notation **G** which is the diagonal matrix with $G_{ii} = g'({\mu_i}^{[k]})$ this expression can be manipulated into an iterative least squares expression of the form $S = ||\sqrt{W^{[k]}}[\mathbf{z}^{[k]} - \mathbf{X}\boldsymbol{\beta}]||^2$. Therefore iteration proceeds as follows until convergence occurs:

1. Use the current iterations of μ and η to calculate $\mathbf{z}^{[k]} = g'(\boldsymbol{\mu}^{[k]})(\mathbf{y} - \boldsymbol{\mu}^{[k]}) + \boldsymbol{\eta}^{[k]}$ and the iterative weights from the diagonal matrix **W** with ${W_{ii}}^{[k]} = \frac{1}{V({\mu_i}^{[k]})g'({\mu_i}^{[k]})^2}$. If this is the first iteration then common practice is to set $\mu_i = y_i$ and $\eta_i = g(\mu_i)$.

2. Minimise S with respect to $\boldsymbol{\beta}$ to get $\boldsymbol{\beta}^{[k+1]}$ and use these new estimates of $\boldsymbol{\beta}$ to get updated values of μ and η. Increment k by 1 and repeat until convergence has occurred.

Clearly this extension to the linear model allows more features of the data to be modelled instead of removed and removing the assumption of normality also allows greater flexibility. Link functions could be investigated and functions of explanatory variables included so as to allow a desired relationship with the response to be mimicked. However, this could become very complicated very quickly, also, the aim of this study is not to perfectly explain the given data, as these contain inhomogeneities, but instead to be able to reproduce data that are like them. Thus, a still more flexible approach would be desirable with even fewer constraints on the model framework and this is what the Generalised Additive Model offers.

The Generalised Additive Model

A generalised additive model (GAM) takes a similar form to the GLM, but with even fewer restrictions in how the explanatory variables can enter the model. $\mathbf{y} \sim p(\ \mathbf{y}; \theta, \phi)$ remains the same, but now $g(\boldsymbol{\mu})$ can contain both traditional $\mathbf{X}\boldsymbol{\beta}$ terms and smooth functions of the explanatory variables, $f_j(x_{ji})$. These smooth functions are commonly fitted using a spline based approach.

Splines are formed from a set of basis functions that can be combined in such a way as to

create a smooth curve that mimics the behaviour of f(·). If left unconstrained these smooth curves can be over fitted to the data, therefore, it is advisable to add a penalty term so that the spline will be penalised if it becomes too unsmooth. This means that the estimate for the function $f(x_i)$ is chosen to minimise $\sum_{i=1}^{n}[y_i - f(x_i)] - \lambda \int_{Range(x)}[f''(x_i)]^2 dx$ where λ is a smoothing parameter and the smaller λ is the less the smooth function is penalised. Because the form of the GAM differs from the GLM, the fitting method also changes and is now Penalised-Iteratively Re-weighted Least Squares (P-IRLS). This fitting mechanism is described below, and is paraphrased from chapter four of Wood [2006].

Before the fitting process is explained it is helpful to introduce some notation. If a spline can be written as a series of basis functions, then each smooth function can be written in matrix form $f_j = \widetilde{X_j}\widetilde{\boldsymbol{\beta}_j}$ as defined on page 167 of Wood [2006]. This means that the entire GAM link function can be written as $g(\boldsymbol{\mu}) =$**X**$\boldsymbol{\beta}$, subject to some identifiability constraints to ensure that this is a one-to-one function. Here **X** contains the traditional model matrices as from the GLM in addition to the newly defined **X**$_j$'s for each smooth function, which are the $\widetilde{X_j}$'s multiplied by a matrix **Z** that is used to ensure that the aforementioned identifiability constraints are met. Further details on **Z** can be found on page 168 of Wood [2006]. Similarly, the $\boldsymbol{\beta}$ contains the parameter vector for the parametric part of the model and the parameters from the smooth function bases.

It would be possible to get the maximum likelihood estimate of $\boldsymbol{\beta}$, but this would likely result in over fitting to the data. Which is why penalised likelihood is maximised using P-IRLS. This penalised likelihood is defined as $l_p(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T$**S**$\boldsymbol{\beta}$ where **S**$= \sum_j \lambda_j$**S**$_j$ and the **S**$_j$ are matrices of known coefficients and the reader is referred to section 4.2 of Wood [2006] for a more in depth explanation of these matrices. The λ_j are smoothing parameters.

If the scale and smoothing parameters are unknown then fitting proceeds as follows, details are not given for the fitting method if these parameters are known as they never were for this thesis. First some new terminology is defined, any terminology not defined here can be assumed to be defined in the same way as for the GLM.

Define the 'influence matrix' of the GAM as **A**=**X**(**X**^{T}**WX**+**S**)$^{-1}$**X**^{T}**W** and the trace of this matrix as tr(**A**)=$\sum_{i=1}^{dim(A)} a_{ii}$. Then the scale parameter can be estimated as $\hat{\phi} = \frac{\sum_i V(\hat{\mu_i})^{-1}(y_i - \hat{\mu_i})^2}{n - tr(A)}$. Because the scale parameter has been estimated, the smoothing parameters are estimated so that they minimise the Generalised Cross Validation (GCV) score, $V_g = \frac{nD(\hat{\boldsymbol{\beta}})}{[n - tr(A)]^2}$. Here $D(\hat{\boldsymbol{\beta}})$ is the deviance of the model, defined as $D = 2[l(\hat{\boldsymbol{\beta}_{max}}) - l(\hat{\boldsymbol{\beta}})]\phi$, and '$l(\hat{\boldsymbol{\beta}_{max}})$ indicates the maximised likelihood of the saturated model: the model with one parameter per data point', from section 2.1.6 of Wood [2006]. To minimise V_g a numerical method involving its derivatives can be used inside P-IRLS as the following steps indicate. These are the steps that must be iterated until convergence occurs. What follows is directly taken from page 187 of Wood [2006], further information on the calculation of some of these steps can be found there. Two final pieces of notation to introduce are $\rho_k = log(\lambda_k)$ and the omission of hats from estimates for ease of writing.

1. "Evaluate the pseudodata, $z_i = \frac{\partial \eta_i}{\partial \mu_i}(y_i - \mu_i) + \eta_i$, and corresponding derivatives,

$\frac{\partial z_i}{\partial \rho_k} = (y_i - \mu_i)g''(\mu_i)\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \rho_k}$. Note that $\eta_i = X_i\hat{\boldsymbol{\beta}}$, is the 'linear predictor' for the i^{th} datum.

2. Evaluate the weights, $w_i = [\frac{V(\mu_i)g'(\mu_i)^2}{w_i}]^{\frac{-1}{2}}$, and derivatives, $\frac{\partial w_i}{\partial \rho_k} = -\frac{1}{2}\frac{w_i^3}{w_i}[V'(\mu_i)g'(\mu_i) + 2V(\mu_i)g''(\mu_i)]\frac{\partial \eta_i}{\partial \rho_k}$.

3. Drop any observations (for this iteration only) for which $w_i = 0$ or $\frac{\partial \mu_i}{\partial \eta_i} = 0$.

4. Find the parameters, $\hat{\boldsymbol{\beta}}$, minimising $\sum_i w_i^2(z_i - X_i\boldsymbol{\beta})^2 + \boldsymbol{\beta}^T\mathbf{H}\boldsymbol{\beta} + \sum_k e^{\rho_i}\boldsymbol{\beta}^T\mathbf{S}_i\boldsymbol{\beta}$ and the derivative vector, $\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \rho_k}$, for the next iteration" where **H** is any positive semi-definite matrix which could be used to impose bounds on the smoothing parameters, regulate an ill-conditioned problem or just be zero if neither of these are necessary, see section 4.6.1 of Wood [2006].

Here the GAM fitting method that has been described is that used for this study; where the scale parameter, ϕ, and the smoothing parameters, λ_j, are unknown and must be estimated along with the $\beta' s$. The fitting method can vary according to what is known about the model beforehand and the amount of computational cost that is deemed acceptable. For further information on these other methods, the method described here or the possible spline bases available the reader is referred to chapters three and four of Wood [2006].

4.1.2. Model Formulation: The Gamma Generalised Additive Model

The previous chapter and previous section respectively introduced the data and modelling possibilities available for this work. The focus now turns to the specific model that was chosen, its method of implementation and the reasons for these choices.

The model family

The model framework used for this study was the Generalised Additive Model. This model demanded the choice of an appropriate distribution in the exponential family and the choice of a link function. Figure 3.4 showed that the data are skewed and continuous, this suggested that the Gamma could be an appropriate distribution, but the Gamma is a positively skewed distribution and the data are negatively skewed. Therefore, a transformation had to take place. Incorporated into this transformation there had to be an addition of a constant as the Gamma cannot work with negative values.

The transformation that was applied to the data was $TMEAN60 = 60 - TMEAN$. This transformation was applied to ensure that all recorded values of $TMEAN60$ were now positive and that no temperatures that have ever occurred in the United States were impossible to get from this model. For example, the hottest maximum temperature ever verifiably recorded, which is therefore hotter than the hottest mean temperature, is 57°C [Shein et al., 2013]. After transformation this value would become 3, therefore it would still be within the range of the Gamma distribution.

The Gamma distribution can be written as $p(y) = \frac{1}{s^a\Gamma(a)}y^{a-1}e^{-\frac{y}{s}}$ where the two pa-

rameters a and s are the shape and the scale respectively. The Gamma function is $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$ where t is a dummy variable. Using GLM theory, it can be shown that the mean of the Gamma distribution is as and the variance is as^2. Hence, it can be seen that as the mean increases so does the variability.

Before the data transformation, winter (lower) temperatures exhibited the most variability in three of the four regions, and still high variability in the remaining region, the South West. After the transformation, the largest values were the winter values and were therefore rightly associated with larger variances. This is another aspect of the data that the Gamma distribution captured well.

These arguments show that the Gamma distribution was a logical choice for the modelling of these data subject to arithmetic transformations. The following section proceeds to show which variables were included in the models to explain these data and the justifications for these choices. Note the use of models, plural, here. The same variables were included in the model for each focus region, but the models themselves were fitted to data from each region separately to maximise their ability to reproduce individual region characteristics.

One more point to note before moving on to the model variables is that, as previously stated, the GAM allows a link function to be chosen. In this work, the logical choice of a link function was the identity link, $g(\mu_i) = \mu_i$ as this meant that predictions and plots from the model were on the same scale as the model inputs and were thus easily interpretable.

The variables

Temperatures are affected by many climatic and non-climatic factors. This section gives a brief overview of the variables included in this study that seek to explain temperature variability. There are undoubtedly other factors that could have been included, but model complexity had to be balanced with the computational cost of running the models. The final models were deemed fit for purpose using the criteria detailed later in this chapter.

The following explanatory variables were included in the models:

- Time: This was a simple numerical variable from 1 to 15340 (the number of days in a complete station series from 1^{st} January 1970 to 31^{st} December 2011). It was included to account for the long term trend in the data. Other studies have acknowledged the benefit of incorporating true climate variability in a model to create clean data, this has often been done by obtaining a trend from climate model data, for example Titchner et al. [2009]; Williams et al. [2012] and Willett et al. [2014], but in this study the longer term variability is drawn from the observations themselves.

- Day of the year: Again, this was a numerical variable from 1 to 365 to account for the position of an observation in the seasonal cycle. The 29^{th} of February was given the indicator 59.5 to ensure that the numbering of other days remained consistent in leap and non-leap years. It has already been stated and shown in section 3.2.1 that temperatures exhibit a seasonal cycle, but this has also been acknowledged by other studies and it is a known scientific fact that North American temperatures

have a seasonal cycle [Trenberth, 1983; Willett et al., 2014].

- Altitude (m): The altitude or elevation of a station is commonly accepted to affect its temperature distribution, with stations at higher altitudes generally exhibiting lower temperatures [Ahrens, 2000]. This variable was provided with the GHCND station data.

- Latitude (degrees): The latitude (north-south position) of a station defines how close a station is to the equator, those stations closer to the equator are typically warmer [Ahrens, 2000]. This variable was provided with the GHCND station data.

- Longitude (degrees): The longitude (east-west position) of a station can help to capture and incorporate land features into the model such as mountain ranges and deserts. These differing surface characteristics are known to affect the temperatures as was stated when discussing the different focus regions for this study in chapter three section 2.1 [Corcoran and Johnson, 2005]. This variable was provided with the GHCND station data.

- Distance to the coast (m): Proximity to a coast is expected to affect temperatures. For example, near coastal stations may be expected to be cooler than inland stations in summer, but warmer in winter because of the fact that land and water heat at different rates [Ahrens, 2000]. Distance to the coast was calculated using ArcGIS data and the latitudes and longitudes given in GHCND, acknowledgement is due to Mark Cherrie of Exeter University for the calculation of these values. Please note that 'Distance to the coast' is the distance to the nearest ocean, therefore the Great Lakes are not explicitly included in the models. Models with and without the Great Lakes were investigated and those without proved to have a better performance.

- Downward solar radiation flux (Wm^{-2}): Also referred to as 'levels of sun' from this point onwards. It is intuitive that the levels of solar radiation a station is exposed to will affect its temperature [Ahrens, 2000]. This variable was obtained from the 20CR data.

- Southern Oscillation Index: Included as an ENSO indicator. As stated in section 3.1.3 this variable was included as a representative of ocean and atmosphere interactions and larger scale temperature variability. ENSO is known to affect at least some regions of the USA, as stated in chapter three section 1.3 [Wang et al., 2013]. This variable was obtained from the Australian Bureau of Meteorology.

- U-wind (ms^{-1}): U-wind is eastward wind, that is wind that blows from the west towards the east would have a positive value and wind that blows from the east towards the west would have a negative value. This is how the wind direction was defined in the 20th Century Reanalysis and the terms should not be confused with the more commonly used phrases 'westerly', a wind from the west and 'easterly' a wind from the east. Wind direction may affect temperature depending on the conditions where the wind is blowing from, e.g., a coast, a mountain range etc [Ahrens, 2000; Corcoran and Johnson, 2005]. Wind speed would also be expected

to affect recorded temperature measurements. This variable was obtained from the 20CR data.

- V-wind (ms^{-1}): V-wind is a northward wind, that is wind that blows from the south towards the north would have a positive value and wind that blows from the north towards the south would have a negative value. Again, these are not the same as northerly and southerly winds. In the United States winds from the south would be expected to bring warmer weather in general and those from the north would be expected to bring colder weather in general owing to the effects of latitudes on temperatures stated above and in [Ahrens, 2000]. This variable was obtained from the 20CR data.

- Precipitation rate ($Kgm^{-2}s^{-1}$): Precipitation rate and temperature are often linked and therefore inclusion of a precipitation indicator is wise. For example, especially in summer, larger amounts of precipitation are likely to result in lower temperatures for most of the United States [Zhao and Khalil, 1993]. This variable was obtained from the 20CR data.

- Precipitable water content (Kgm^{-2}): This variable considers the entire atmosphere as a single layer and reports how much moisture is present that could theoretically fall as precipitation. Considering the atmosphere as a single layer is acceptable as the majority of the moisture will be concentrated relatively near the surface in the atmospheric boundary layer [NASA, 1991]. Both Trenberth et al. [2005] andRuckstuhl et al. [2007] refer to precipitable water as integrated water vapour and Ruckstuhl et al. [2007] suggests that integrated water vapour is linearly related to specific humidity. From these studies it is therefore reasonable to assume that a humidity variable such as specific humidity would behave similarly in the model to precipitable water content. Precipitable water content was obtained from the 20CR data.

- Sea Level Pressure (Pa converted to hPa): This variable can help provide information on larger scale weather systems, e.g., storm systems, that might not be captured through the inclusion of other variables [Ahrens, 2000]. This variable was obtained from the 20CR data.

- Temperature (K converted to °C): This variable was included from the 20CR data as it is expected that reanalysis temperatures will be a helpful predictor of observed temperatures, including their longer term trends [Compo et al., 2011, 2013]. The reanalysis data were not of a high enough resolution to allow reanalysis temperatures to be used to directly create the clean series without any modelling needing to take place.

Not all variables are strictly necessary for all regions. For example, distance to the coast and longitude serve a very similar purpose for the Wyoming model and distance to the coast and altitude are closely linked in the South East. However, it was decided to keep the models uniform across regions to make this work as generalisable as possible. It is hoped that a future researcher could use this model anywhere in North America and it would be appropriate because of the array of explanatory variables that have been incor-

porated. If the model were to be taken to a very different area, Africa for instance, the researcher may want to consider changing the variables included. Africa may also benefit from a less complex model as there is less data availability and fewer variables would better guard against overfitting. Overfitting was not a concern in the USA as the data are plentiful, therefore, variables that had some justifiable reason for inclusion in any region were included in all regions. That is, the selection of variables was primarily physically based and investigations into dropping certain variables were only carried out very early on in the model creation stage before it was decided that less was not more in the case of this project. If a simpler model was sought that was not necessarily the same in all regions then an approach such as stepwise selection based on the Akaike Information Criteria (AIC) could be used to determine the best model.

Model Selection

To reduce the computational cost of model fitting the data were thinned so that only every other day at a station was taken for modelling purposes. This thinning was done using the time index of the data. This index ran from 1 to 15340. To ensure that days from all points in the time series were well represented odd indexed days were taken at one station and even indexed days were taken at the next. Thus, the models included information from the whole time series, but were computationally cheaper to fit. This thinning also reduces, though does not remove, the autocorrelation in the data, which is beneficial as the GAM is designed to work with independent data. The ranges of the values for all the variables were compared before and after thinning and although a few extremes were reduced, none were substantially reduced.

Once the data were thinned and the explanatory variables were chosen it was necessary to decide how they would be included. They could be included parametrically (linearly in practice here); as smooth functions; or as smooth surfaces that take into account the effect of more than one variable at a time on temperature. One could include all variables as smooth functions, but this is more computationally costly, and of little benefit when relationships are linear or near linear. This requires assessment by eye, as, given the opportunity to be very unsmooth a relationship will take it. For this reason, scatter plots were formed between variables that were suspected to have a linear relationship with mean temperature, namely altitude and reanalysis temperatures.

Altitude was tricky, as, of course, stations exhibit a range of temperatures while remaining at the same altitude. This meant that the relationship between temperatures and altitudes was not constant. It should also be remembered that altitude is not an immediate indicator of topography, it is easy to assume that high altitude stations must be mountain top sites, but Pepin and Siedel [2005]; Pepin and Norris [2005] looked at 'high elevation sites' (over 500m) and were still able to classify them into groups which included both mountain summit and valley sites. These different topographies result in different influences on daily mean temperatures, i.e. wind is likely to have a greater effect at mountain summit sites [Pepin and Lundquist, 2008], but all sites in the studies by Pepin had high altitudes.

In spite of this knowledge that higher altitude sites aren't necessarily high topographi-

cally, the scatter plots, not shown, exhibited the traditionally expected slight negative correlation between station height and temperature overall. This was considered sufficient justification to include altitude parametrically and not as a smooth function. Including a categorical factor of topography as well as altitude would be an interesting addition to the model as the aforementioned studies do show its effects on temperatures, however it was not investigated further in this work.

A similar argument could be made in favour of including latitude linearly, as, generally speaking, stations get cooler the further away they are from the equator. However, latitude was included as a smooth function because it was deemed appropriate to include longitude as a smooth function and it was logical to include both co-ordinate variables in the same manner. Longitude was included as a smooth function because it is not necessarily the case that the temperatures across a region are linearly related to their east-west position. For example, factors such as deserts and mountain ranges will affect temperatures and the longitude smooth function can pick up these features to a certain extent.

Reanalysis temperatures, as expected, were highly positively correlated with observed temperatures and therefore allowing them to enter the model linearly was a simple decision.

The remaining variables were then included as smooth functions and their shapes suggested that this was an appropriate step. However, it is apparent when looking at climatic data that relationships between temperature and other variables can exhibit seasonal cycles, which suggested that the smooth surface option of the GAM should also be investigated. This option allows two explanatory variables to be interacted with each other, so that the effect of one on the response differs according to the value of the second.

The advantages of including a variable interacted with 'day of the year' are twofold. Firstly, it will mean that the seasonal cycle can vary between stations in a region, which would not be the case if a smooth function of 'day of the year' by itself were included. This can then lead to a more realistic relationship with temperature being modelled, which benefits the creation of the benchmark data. Secondly, there is evidence to suggest that the effects of inhomogeneities are not constant throughout the year [Trewin, 2013]. Therefore, if an interaction term is included and one of its variables is perturbed then a seasonally varying inhomogeneity can be created. This methodology will be explained in much greater detail in the following chapter.

Only one smooth surface could be included in the model otherwise the two surfaces would interact with each other and lead to the possibility of unrealistic values being predicted. The choice of which variable to include with 'day of the year' involved much careful consideration. Initially all explanatory variables were considered as potential candidates and a scatter plot was produced between 'day of the year' and the explanatory variable in question in each of the four focus regions. These plots showed which variables had a naturally occurring seasonal cycle, those which clearly did didn't need to be considered further as they were already causing seasonal variation within and between stations. Unsurpris-

ingly the 'levels of sun' variable fell into this category. Precipitable water content also had a reasonably distinct seasonal cycle in all but the South West region and therefore was also not considered further.

Latitude, longitude and distance to the coast could have different effects on mean temperatures at different times of the year, but this will be due to other factors such as precipitation or wind regimes. Therefore, none of the location variables were considered for inclusion as a smooth surface. SOI was not considered for inclusion as a smooth surface as its value is constant for a month at a time and is the same for all stations. This means that the full benefit of a smooth surface would not be realised if SOI were the variable to be included with 'day of the year'. This left the options of sea level pressure, eastward or northward wind and precipitation rate. Therefore, further investigation was carried out on the models that included one of these four smooth surfaces. Sea level pressure was soon ruled out, as, although different pressures had different effects on temperature in summer, this was not the case in winter.

The remaining three possible models were then analysed for all regions to assess which produced the best predicted data. The analysis consisted of looking at the correlation and mean squared error between the predictions and the observations; the adjusted R-squared value (which should be high) and the generalised cross validation score (which should be low) for the models and comparing models using ANOVA tables. Alongside comparing the variable of interaction, how the interaction was to be incorporated was also investigated.

Smooth function interaction terms between two continuous variables can be incorporated in one of two ways: as a thin plate regression spline smooth surface, or as a tensor product smooth surface. Thin plate regression spline smooth surfaces are constructed from two-dimensional thin plate regression spline bases which are combined in order to produce the desired surface, just as one dimensional splines are combined to form a smooth function. These surfaces are good if the two variables are on similar scales, but not as good if the scales are very different. Tensor product bases are good even if the variables are measured on very different scales as bases are built for each variable separately and then combined using a tensor product, see section 4.1.8 of Wood [2006].

The analyses revealed that the best model was obtained by including a smooth surface of 'day of the year' and 'precipitation rate' using a tensor product smooth. However, there is a drawback to this model. Some of the inhomogeneities in this study were produced by scaling certain explanatory variables by a small factor. If this were to be done with 'precipitation rate' in order to create an explicitly seasonally varying inhomogeneity, then any days with no precipitation would be unaffected by the scaling, thus the inhomogeneity addition method would not be as effective as it had the potential to be. For this reason the model that came a very close second in the analyses was used. This was the model incorporating 'day of the year' and eastward wind as a smooth surface.

The final model can therefore be written as:

$$TMEAN60_{it} \sim Gamma(a, s_{it})$$

where the mean of the Gamma distribution $\mu_{it} = as_{it}$ and

$$\mu_{it} = \beta_0 + \beta_1 Altitude_{it} + \beta_2 Tempforecast_{it} + f_1(Dyear_{it}, UW_{it}) + f_2(Time_{it}) + f_3(Lat_{it}) + f_4(Long_{it}) + f_5(Sun_{it}) + f_6(SOI_{it}) + f_7(VW_{it}) + f_8(Precip_{it}) + f_9(PWC_{it}) + f_{10}(Coast_{it}) + f_{11}(SLP_{it})$$

Here i is a station index from 1 to N, where N is the number of different stations used as inputs to the model and t is a time index, taking values from 1 to 15340. It may seem unnatural to index location and altitude variables according to time as well as station, but, as will be explained later, these variables will not necessarily be held constant for the whole station record length. For example, a station may be relocated.

After the model formulation was finalised it was necessary to decide how to include the smooth functions. Decisions when introducing smooth functions involved choosing the bases from which they were built and an upper limit for how unsmooth they were allowed to be. How unsmooth a smooth function is in the GAM can be broadly described by the function's effective degrees of freedom, or edf. The edf can be thought of as an approximation to the degree of the polynomial you would need to get the same amount of variability. For the smooth function of time it was decided to heavily restrict the variability, forcing it to be only as variable as a quadratic. The reason for doing this was because the time component is included to incorporate the long term underlying variability in the data that has not been captured by other variables, not to reproduce every variation that happens over the time period. The danger of allowing the smooth function of time too much freedom would be that it could mimic a network wide inhomogeneity, such as a change in the time of observation, thus undermining the assumption of homogeneity in the clean data.

All other smooth functions were allowed up to nine effective degrees of freedom. All bar one, the smooth function of eastward wind in Wyoming, took advantage of this maximum and thus it was necessary to investigate whether the limit should be increased. This investigation consisted of analysing the residuals from the fitted model. When smooth functions were fitted with partial residuals overlaid, the scatter about the smooth function appeared uniform, this is a sign of a well fitting model [Wood, 2006]. However, when smooth functions were fitted to the deviance residuals, with respect to each of the covariates in turn, there was clear suggestion of unmodelled variability. Deviance residuals are the square roots of each point's contribution to the deviance, where the deviance is the scaled difference in log likelihood between the saturated model and the fitted model. Modelling the deviance residuals with respect to a covariate simply means that a GAM is fitted with the residuals as the response variable and the explanatory variable in question as the only smooth function. Because, in all cases when this was done, the smooth function was shown as being significant this suggested that an investigation into increasing the allowable degrees of freedom for the smooth functions should be carried out. Therefore, all models were refitted with just over a doubling of their allowable effective degrees of freedom and the resulting models were compared to the originals.

When the allowable degrees of freedom for smooths was increased, some smooths be-

came smoother and others became more variable, some also started noticeably affecting each other. This is because climate variables will be correlated with each other. Therefore, a smooth function may behave differently depending on whether another variable is or isn't included in the model, or depending on how many degrees of freedom other variables have. For example, the range of effect sizes on temperature from the longitude smooth function became incredibly wide in Wyoming, the South East and the North East. In some cases this increase was partly compensated by the ranges of other smooths increasing too, but not always, meaning that some predictions from these models would be completely inadequate. It is also worth noting that for both the South East and the North East the models with less variability allowed gave better (lower) GCV scores, suggesting that they were preferable. Therefore, for these reasons, and in the interest of computational efficiency the original models were preferred over those with increased degrees of freedom.

In this work thin plate regression splines are used as bases for the smooth functions. Thin plate regression splines fit the data well without being too computationally expensive or reliant on user choices. They can also be used as the basis of smooth surfaces allowing a continuity between the single and interaction terms in the model.

The models described here were all fitted using the package `mgcv` in R. As already outlined, a model for each region was fitted separately to maximise the inclusion of the individual region characteristics. The `mgcv` package is well documented and Wood [2006] devotes a lot of attention to its capabilities. The models should be easily reproducible from the information given in this section. One small point to note is that, although with restricted effective degrees of freedom over-fitting was not a major concern, because GCV is known to sometimes over fit, a slight alteration to the GAM was made when applying it in R. This alteration was setting the parameter `gamma` to 1.4, as advised in Wood [2006].

Model Outputs

As shown in the model formulation, explanatory variables entered the GAM in one of three ways. As parametric terms, i.e., linearly with a coefficient indicating the sign of the effect the term has on mean temperatures; as smooth functions, allowing easily for non-linear relationships; or as smooth surfaces allowing two variables' impacts on temperature to interact with each other.

The explanatory variables that entered the models parametrically were altitude and reanalysis temperature. The coefficient for altitude indicated an increase in altitude lead to a decrease in temperatures in Wyoming, the North East and the South West as expected. In the South East the coefficient suggested increasing altitude lead to increasing temperatures, but the coefficient was smaller than in the other regions, as are the altitudes, thus this relationship does not seem unreasonable. Also, as expected, the coefficient for reanalysis temperature indicated that as reanalysis temperatures increased so did modelled temperatures.

The remaining explanatory variables entered the model as smooth functions or smooth

surfaces. Not all smooth functions are displayed in the interests of space and ease of interpretation. When interpreting those that are displayed the following explanations should be kept in mind. Firstly, smooth functions are centred to ensure that they are identifiable. That is, the sum of all the elements of the smooth must be zero to ensure that the smooth function is a one-to-one function [Wood, 2006]. Secondly, this identifiability constraint does not stop different smooth functions interacting with each other, as was stated in the previous section. Thus, the shape of a smooth function does inform the reader what impact a particular covariate value has on predictions, but only in relation to the other terms and smooth functions in the model. This means that smooth functions do not always have intuitive and easily interpretable shapes. For example, in Wyoming longitude and distance to the coast are highly correlated, which means the distance to the coast function appears to have a large influence on temperatures, which seems counter-intuitive for a landlocked state.

However, this study was concerned primarily with creating realistic data, not trying to explain the data that already existed. Hence, smooth functions were included to allow for more complex non-linear relationships to be modelled without having to rationalise their shape. Therefore, non-intuitive smooth function shapes were not cause for concern as long as the temperature series created from the models were realistic, which the following section illustrates was the case.

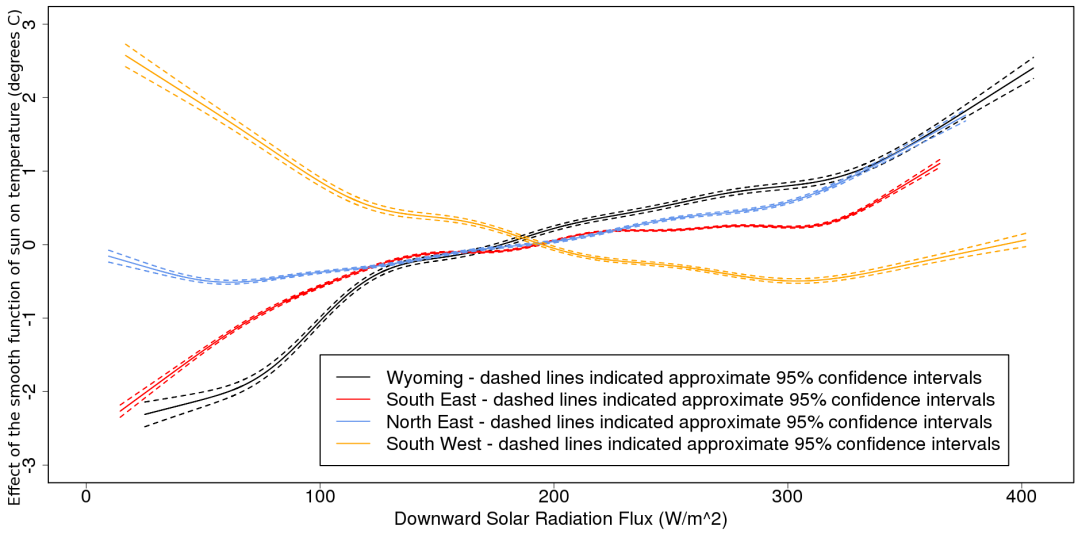

Figure 4.1. Centred smooth functions of levels of sun for Wyoming (black), the South East (red), the North East (blue) and the South West (orange).

Figures 4.1 and 4.2 show examples of two sets of smooth functions from the fitted models, for levels of sun and northward wind respectively. These serve as good illustrations for the caveats provided above. In figure 4.1 it can be seen that although in Wyoming, the South East and the North East higher levels of sun have a warming impact on temperatures, this is not the case in the South West. This is not what would be anticipated in the South West, however, in conjunction with the other model terms, predictions in this region are reasonable. For the northward wind smooth functions the interpretations do seem relatively straightforward. In all regions winds blowing from the south bring warmer temperatures, while winds from the north bring colder temperatures everywhere but the

South West, where the relationship is simply that stronger winds have a warming impact on temperatures.

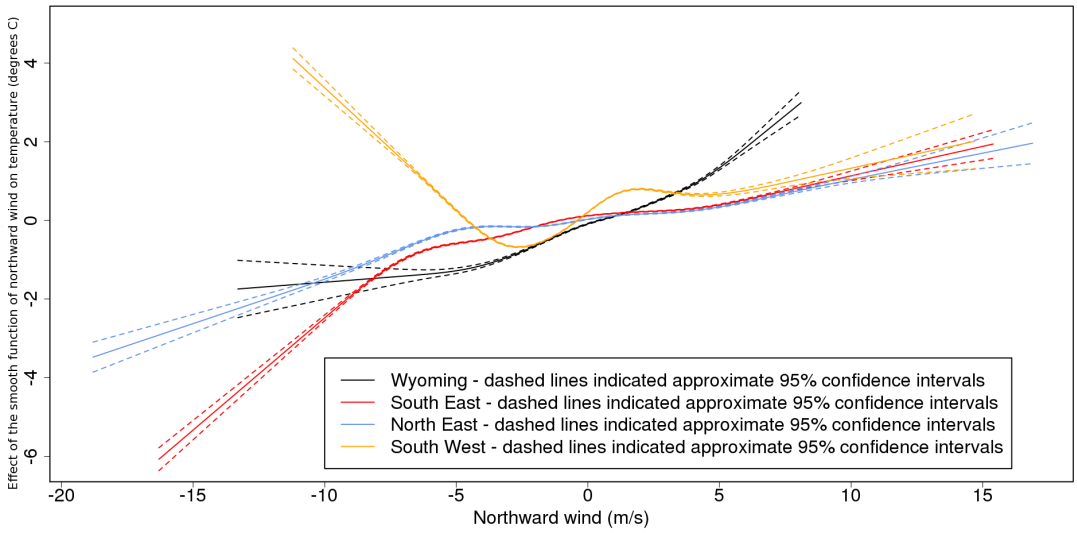

Figure 4.2. Centred smooth functions of northward for Wyoming (black), the South East (red), the North East (blue) and the South West (orange). Positive values indicate wind blowing towards the north and negative values indicate wind blowing from the north.

The two variables that were interacted in smooth surfaces for each region were day of the year and eastward wind. Figures 4.3 to 4.6 show these smooth surfaces. As is the case with the rest of the model components it should be remembered that these smooth surfaces are simply a contributing factor to the final predictions and so they should not be over interpreted.

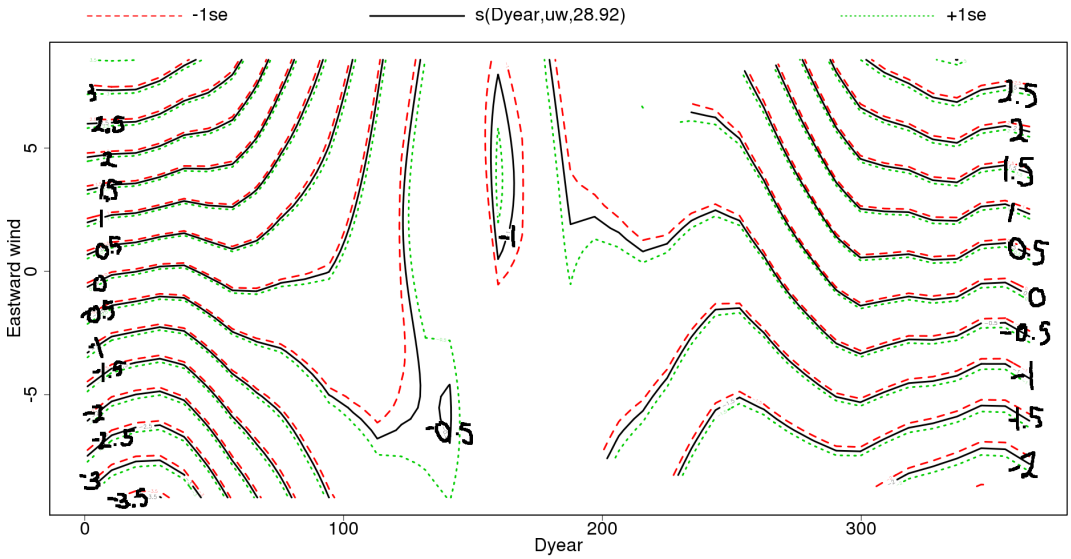

Figure 4.3. A contour plot to illustrate the smooth surface of the day of the year interacted with the eastward wind in Wyoming. The x-axis gives the days of the year from 1 to 365 and the y axis gives the range of eastward winds. The values on the black lines within the plot indicate the effect of the given combination of eastward wind and day of the year on temperature. For example, in Wyoming, winds from the east have a cooling effect on the model predictions in winter. As with the smooth functions there is uncertainty in the values of the smooth surface. Therefore, red dashed lines indicate minus one standard deviation, while green dashed lines indicate plus one standard deviation from the contours.

In Wyoming, figure 4.3, the effect of eastward wind on temperatures varies most at the

ends of the year and least in the middle of the year. At the ends of the year winds from the east cool temperatures, while winds from the west warm temperatures.

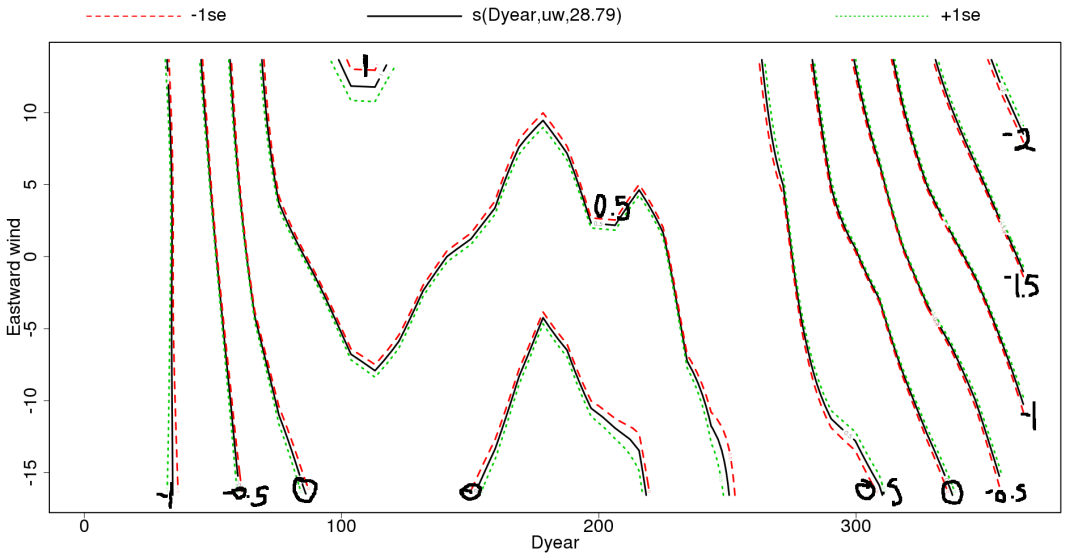

Figure 4.4. A contour plot to illustrate the smooth surface of the day of the year interacted with the eastward wind in the South East. The x-axis gives the days of the year from 1 to 365 and the y axis gives the range of eastward winds. The values on the black lines within the plot indicate the effect of the given combination of eastward wind and day of the year on temperature. It can be seen that the direction and strength of the wind don't actually change predictions that much as the black lines are near vertical in places. However, the time of the year does have an impact, with winds showing a cooling effect in winter and a warming in summer. As with the smooth functions there is uncertainty in the values of the smooth surface. Therefore, red dashed lines indicate minus one standard deviation, while green dashed lines indicate plus one standard deviation from the contours.

In the South East, the speed and direction of the wind matter less, but the time of year still affects the impact of eastward winds on temperatures. At the ends of the year winds cool the temperatures and in the middle of the year they cause a slight warming.

In the North East the relationship is a little more complex. At the ends of the seasonal cycle winds of any speed cool the temperatures, with those from the east (the coast) cooling temperatures more. In the middle of the year strong winds from the east still cool temperatures, but those from the west or weaker winds from the east warm them.

In the South West, at the ends of the seasonal cycle all winds cool temperatures, but those from the west (the coast) cool them more. In the middle of the seasonal cycle almost all winds warm temperatures, but those from the west warm them least and the strongest westerly winds cool them.

4.2. Data Simulation: The benchmark data

As already stated, the aim of this work was not to explain the data that already existed, but to create realistic synthetic data that could be used for assessing homogenisation algorithm performance. It was important that the data that were created were clean initially so that there could be confidence in the conclusions drawn from the study. To be a realistic

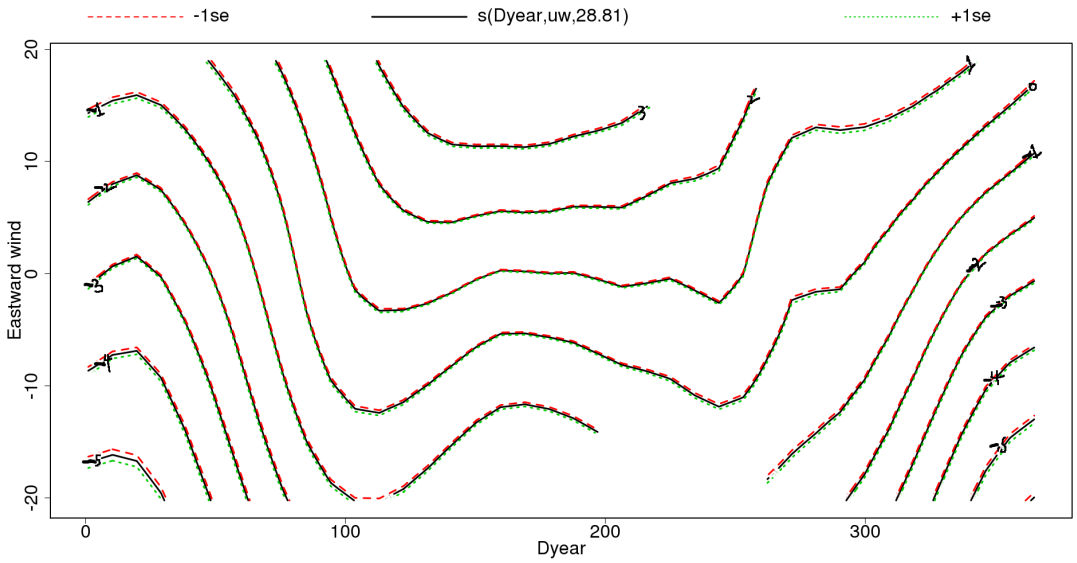

Figure 4.5. A contour plot to illustrate the smooth surface of the day of the year interacted with the eastward wind in the North East. The x-axis gives the days of the year from 1 to 365 and the y axis gives the range of eastward winds. The values on the black lines within the plot indicate the effect of the given combination of eastward wind and day of the year on temperature. For example, in the North East winds in winter always have a cooling effect on predictions, but those from the west (positive eastward wind values) have less of a cooling effect than those from the east. As with the smooth functions there is uncertainty in the values of the smooth surface. Therefore, red dashed lines indicate minus one standard deviation, while green dashed lines indicate plus one standard deviation from the contours.

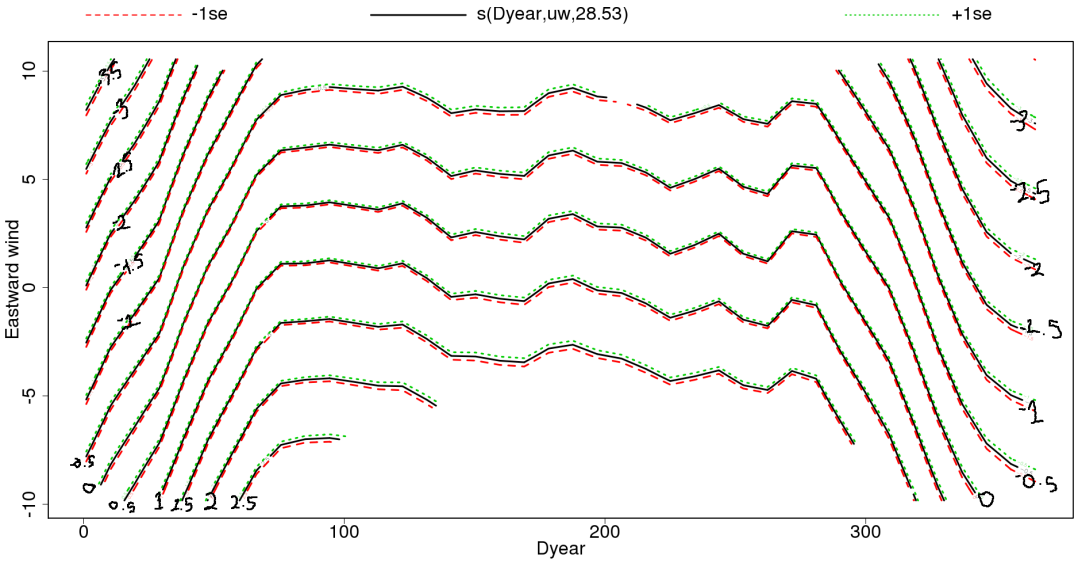

Figure 4.6. A contour plot to illustrate the smooth surface of the day of the year interacted with the eastward wind in the South West. The x-axis gives the days of the year from 1 to 365 and the y axis gives the range of eastward winds. The values on the black lines within the plot indicate the effect of the given combination of eastward wind and day of the year on temperature. For example, in the South West, at the very ends of the year (late December and early January) all winds have a cooling effect on predicted temperatures, with those from the west having the largest effect. As with the smooth functions there is uncertainty in the values of the smooth surface. Therefore, red dashed lines indicate minus one standard deviation, while green dashed lines indicate plus one standard deviation from the contours.

benchmark the synthetic data needed to match the real world well in terms of cross-correlations between stations, standard deviations and autocorrelations in difference series between stations and autocorrelations within stations [Willett et al., 2014]. All these

measures needed to be evaluated on deseasonalised series, that is, series where the seasonal mean cycle had been removed. The reason for preferring deseasonalised data for these checks is that they have a better signal to noise ratio than non-deseasonalised data. That is, the longer term variability is more easily distinguished from the day to day variability. They also have the advantage that measures of correlation should not be dominated by the presence of a seasonal cycle. In the climate literature deseasonalised series would more commonly be known as climate anomalies.

In this work deseasonalisation took place using the following step by step process:

1. Take a station series and work out the mean value for each day of the year using all available years.

2. Use these values to create a series of means the same length as the original series.

3. Subtract the series of means from the original series to create the deseasonalised series.

4. Repeat for remaining stations.

Inter-station correlations were of interest owing to the processes taken by many homogenisation algorithms. As stated in section one of the literature review, algorithms seeking to detect inhomogeneities can be broadly split into two categories; absolute methods where homogeneity tests focus on a single station at a time and relative methods where neighbouring stations are used to try and determine the homogeneity of the station in question [Costa and Soares, 2009]. If inter-station correlations are too high then stations are too similar and inhomogeneities would be easier to detect than in reality, if they are too low then inhomogeneities become harder to detect than in reality [Williams et al., 2012].

As the current study was looking at daily data another aspect that could be investigated was the extremes. Extremes are of interest when evaluating whether the variability of temperatures is changing, thus an algorithm should not smooth out extremes or create non-existent ones. How well the created benchmarks match observed extremes will be reported here and this will be done using non-deseasonalised data.

4.2.1. Predictions

The initial predictions from a GAM are the mean behaviour of the response variable, in this case temperature. These predictions will be identical, down to computer precision, each time predictions are made from the model for the same explanatory variables. A selection of these predictions can be seen in figures 4.7 to 4.10. In each of these figures plot (a) shows the predicted temperature density plot for the region as a whole, plot (b) shows an example of a 'good' station, where 'good' indicates the density of the predictions matching the observations well by eye and plot (c) shows an example of a 'bad' station,

where 'bad' indicates the density of the predictions not matching the observations well by eye.

These plots show that taking the regions as a whole the predictions (blue) are matching the observations (black) reasonably well. The general shapes of the distributions are being captured, but the peaks of the distributions are being overshot at the expense of the extremes. This over- and under-shooting is also evident in the examples of the 'bad' stations and, to a lesser extent, even in the examples of the 'good' stations.

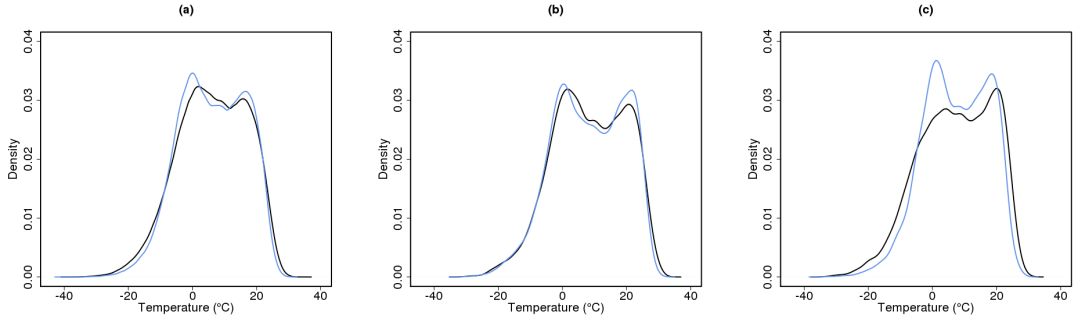

Figure 4.7. Density distributions of observed temperatures (black) and model predictions (blue) in Wyoming for (a) Wyoming as a whole, (b) a 'good' station, (c) a 'bad' station. Where 'good' and 'bad' are here determined by eye.

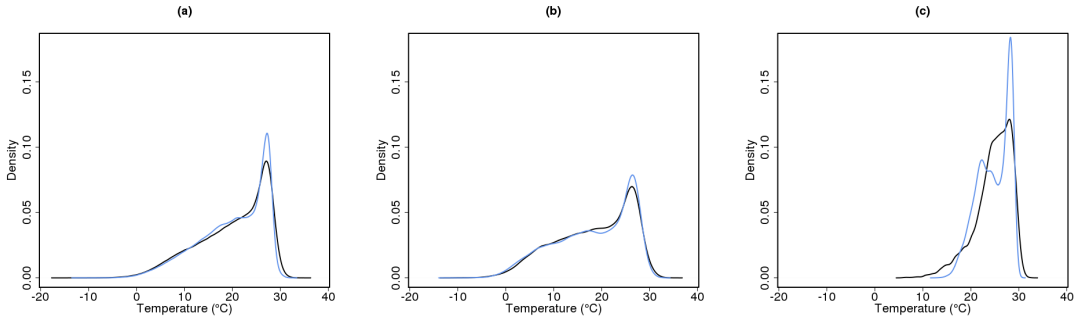

Figure 4.8. Density distributions of observed temperatures (black) and model predictions (blue) in the South East for (a) the South East as a whole, (b) a 'good' station, (c) a 'bad' station. Where 'good' and 'bad' are here determined by eye.

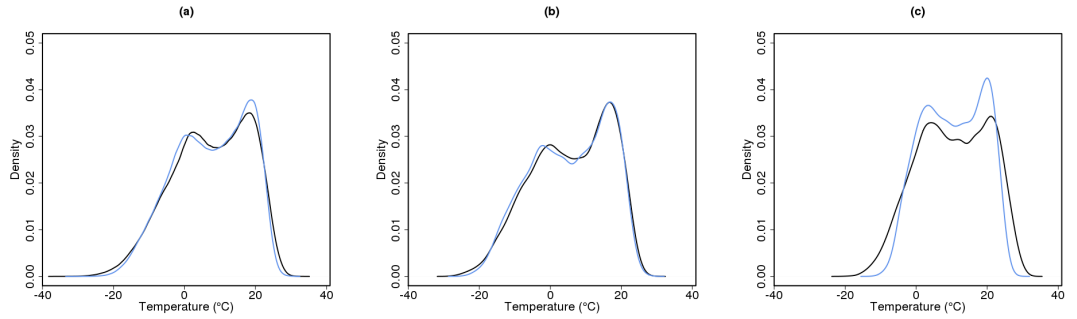

Figure 4.9. Density distributions of observed temperatures (black) and model predictions (blue) in the North East for (a) the North East as a whole, (b) a 'good' station, (c) a 'bad' station. Where 'good' and 'bad' are here determined by eye.

A further problem with simply using the mean predictions from the model is that all stations are predicted to be too similar, resulting in inter-station correlations that are too high. This is illustrated in figure 4.11 which shows the density distributions of inter-station correlations for the observed and the predicted temperatures, where all series have been

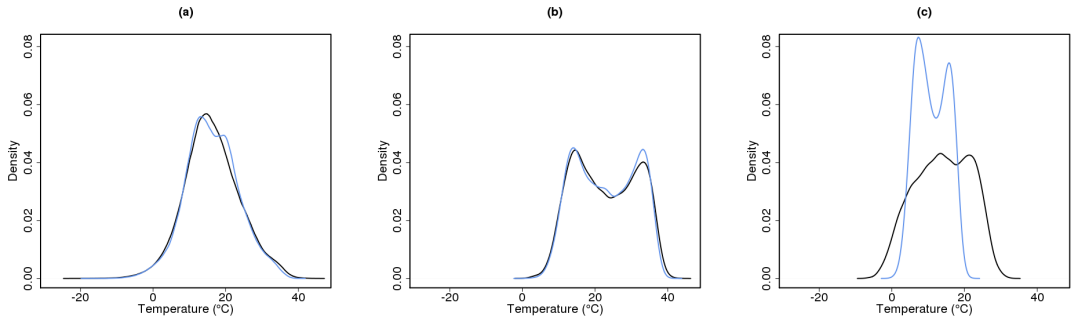

Figure 4.10. Density distributions of observed temperatures (black) and model predictions (blue) in the South West for (a) the South West as a whole, (b) a 'good' station, (c) a 'bad' station. Where 'good' and 'bad' are here determined by eye.

deseasonalised. It must be noted that these comparisons are not exactly like for like as they are being made between homogeneous (predicted) and inhomogeneous (observed) data. However, the difference in inter-station correlations is greater than could be caused by the addition of inhomogeneities.

Given that extremes are an area of investigation that daily data allow for and inter-station correlations should be realistic it was desirable to do some further processing of the predictions to create benchmark data that could reproduce these aspects better.

It should be noted at this point that the problems with using mean predictions straight from the model are also those encountered if downscaled reanalysis temperatures alone are used for the creation of the benchmark data. This is the reason that reanalysis temperatures were only used as an explanatory variable in the final model and not instead of the final model.

4.2.2. Adding Realistic Variability

To decrease the inter-station correlations and increase the temperature range noise was added on to the mean predictions. The process for calculating this noise was as follows:

1. Let $60 - \hat{\mu}_{it}$ be predictions from the GAM. It is necessary to subtract $\hat{\mu}_{it}$ from 60 to ensure predictions are on the $TMEAN$ scale and not the $TMEAN60$ scale. These $\hat{\mu}_{it}$ are the means of Gamma distributions. Therefore, to add extra variability, noise can be generated from a Gamma distribution with $\hat{\mu}_{it}$ as the mean.

2. To generate values from a Gamma distribution both the shape, a, and the scale, s, parameters must be known. Using the facts that $a \cdot s$ is the mean of a Gamma distribution, and that the mean is already known, only one other piece of information is required to work out the value of the remaining parameter. Owing to how the models were fitted, an estimate of a could be obtained from the fitted model output. s_{it} could then be estimated as $\hat{s}_{it} = \frac{\hat{\mu}_{it}}{a}$.

3. Generate a value from this Gamma distribution and denote it by $\hat{y}_{it}$, which can be thought of as a more variable prediction, $\hat{y}_{it} = \hat{\mu}_{it} + \epsilon_{it}$, where ϵ_{it} is the added

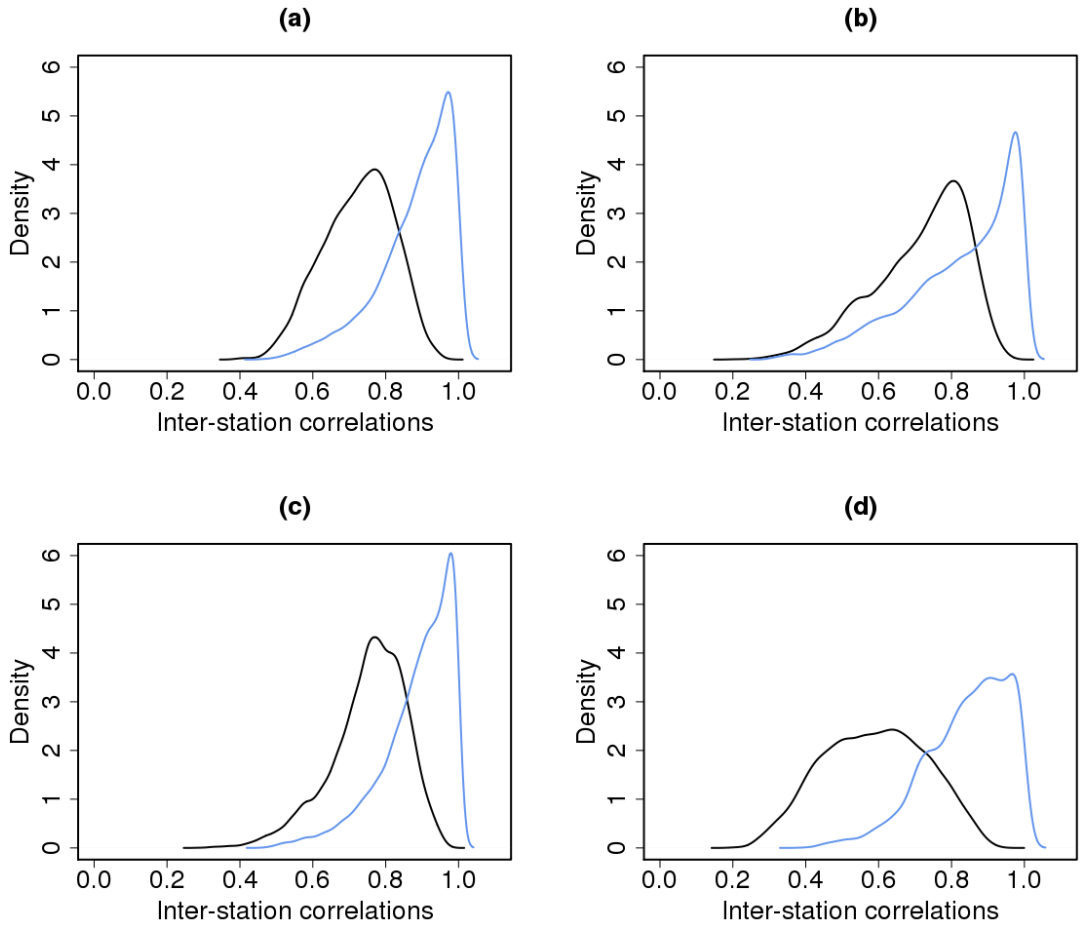

Figure 4.11. Density plots of the inter-station correlations found in observed (black) and predicted (blue) temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. The reader is cautioned that these plots do not indicate inter-station correlations greater than one, it is just an artefact of the plotting process that the blue curves extend beyond one on the x axes.

variability.

Figure 4.12 shows the density of the temperature distributions, transformed back to the true scale, if these noise added predictions are used. It illustrates that adding extra variability in the form of Gamma noise had the desired effect of increasing the range of modelled temperatures. However, it increased the range too much, cold extremes are now overshot in all regions, though warm extremes are still being missed. Also, although adding this extra variability improved the match to the peak of the distribution in Wyoming and the North East, this was not the case in the other two regions.

A more conclusive reason for not using the noise added variability with no further processing can be seen in figure 4.13. This figure shows clearly that inter-station correlations of the predictions are now much too low. This would make the benchmark harder to homogenise than reality, thus undermining the study.

To rectify this problem of decreased inter-station correlations, without reverting back to the raw predictions from the model, a post-processing of these new predictions was applied. This post-processing technique can be thought of as an extension of the previous variability adding method, as outlined below.

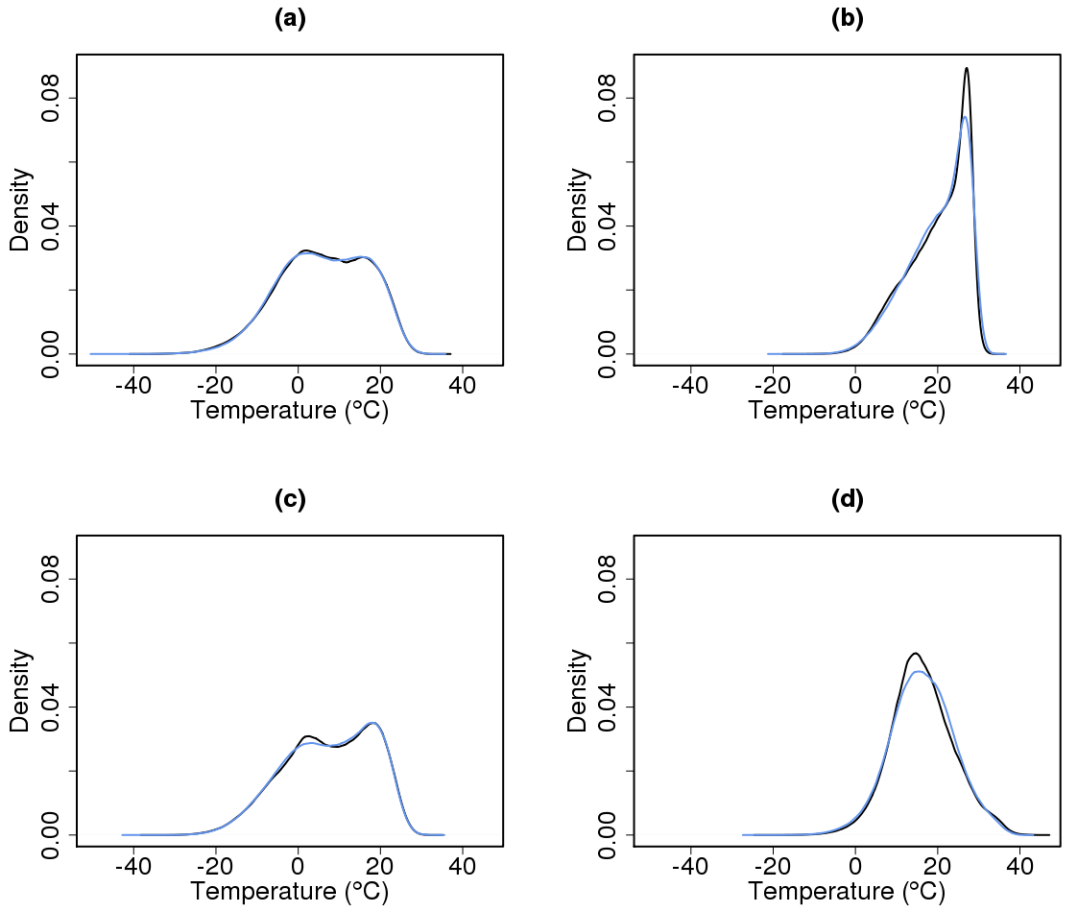

Figure 4.12. Density plots of the temperature distributions found in observations (black) and noise added predictions (blue) from temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West.

4. Extract ϵ_{it}, the extra variability added on to the prediction $\hat{\mu}_{it}$. This extra variability is not correlated between stations on any given day, though the values will have been generated from distributions with relatively similar parameters.

5. Fit a two dimensional loess surface to the given ϵ_{it}'s for each value of t separately. This requires the selection of a smoothing parameter, sp. The larger sp is the more smooth the surface will be.

A loess surface is simply a two dimensional smooth of a region obtained using the current information available (the ϵ_{it} and their geographical locations) and a local fitting method. A local fitting method means that the value of the surface at any given point is influenced by other points in the vicinity of it, usually weighted by how far apart the points are. What proportion of available points are taken into account is decided by sp. For this work sp was held constant for smooths for all days within a region, but was allowed to vary across regions. The weighting function used for the points taken into account at each part of the smooth was a tricubic weighting function, (proportional to $(1-(dist/maxdist)^3)^3$). The loess surface is a polynomial surface and in this work polynomials of degree two were used. The code used for the fitting and prediction from a loess surface was slightly adapted from existing functions available for the R software and is therefore attached as electronic appendix A to allow the reproduction of this work.

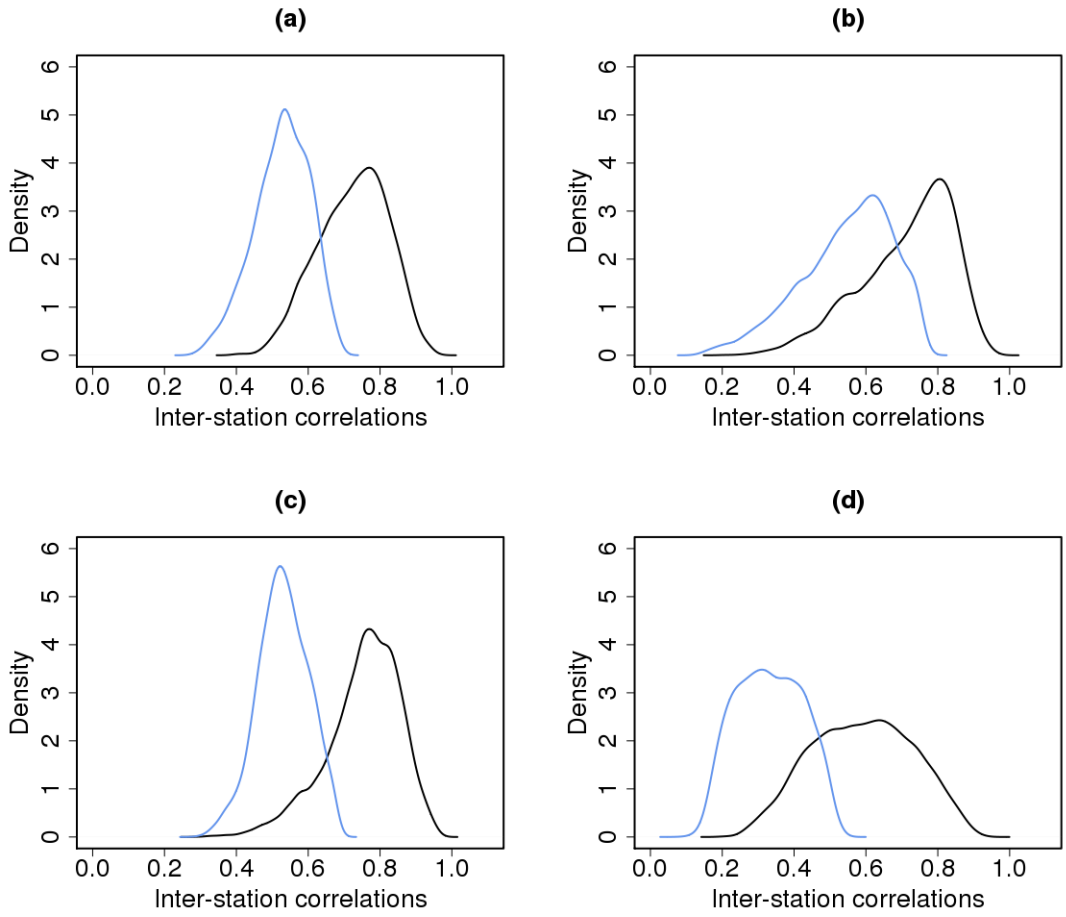

Figure 4.13. Density plots of the inter-station correlations found in observations (black) and noise added predictions (blue) from temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West.

6. Predict back from the loess surface to get a value of smoothed variability, sm_{it}, for all stations and times.

7. The modelled temperature data can now be calculated as $T_{it} = 60 - (\hat{\mu}_{it} + sm_{it})$.

8. Assess these predictions' inter-station correlations and repeat until an appropriate sp value is found.

Multiple values of sp were investigated for each of the regions. As already stated, it was decided to allow sp to vary across these regions in order to obtain the best inter-station correlations in each place instead of having a compromise that was acceptable, but not optimal everywhere. One size would not fit all because the sizes of the regions, their climatic variability and their station density are not constant. For example, smoothing according to a given number of stations could over-smooth Wyoming while under-smoothing the South West.

As will be explained in the next chapter, multiple scenarios of each region were created for this work. Scenarios two and three had an increased station network density relative to scenario one, but in all three scenarios the amount of smoothing was specified by taking the same number of stations into account. The reason for this was that in the less dense scenarios a certain number of stations would typically be more spread than

in the more dense scenarios, thus more smoothing would occur. In the more dense scenarios the underlying station mean predictions would already be more similar, which would have already boosted the inter-station correlations, thus, less smoothing of the noise was desirable. Therefore, although the same *number* of stations was taken into account in the smoothing across all scenarios in a region this amounted to a smaller *proportion* of all available stations in the more dense regions.

Table 4.1. A table to show an overview of the values chosen for sp in the different regions and scenarios and what proportion of stations this amounts to.

Region	Scenarios	Number of stations in total	Number of stations taken into account (sp)	Equivalent proportion of stations
Wyoming	1	75	25	33.3%
Wyoming	2 and 3	158	25	15.7%
Wyoming	4	75	15	20%
South East	1	153	45	29.4%
South East	2 and 3	210	45	21.4%
North East	1	146	40	37.4%
North East	2 and 3	207	40	19.3%
South West	1	151	35	23.2%
South West	2 and 3	222	35	15.8%

These numbers were decided on after careful consideration of multiple alternatives and also research into how neighbouring stations are used in the homogenisation process. Many homogenisation algorithms look for a small number of highly correlated neighbours, therefore it was deemed more important that the upper end of inter-station correlations was captured rather than the lower end. As already stated, inter-station correlations that are too high should make homogenisation easier than in reality because the stations will be more similar, and this is preferable to making the situation more complicated. Therefore, over-smoothing was preferred to under-smoothing as long as no extreme problems were encountered. Some algorithms also focus on local regions around the station being investigated, therefore getting inter-station correlations right at shorter distances was of more concern than making them perfect over longer distances. The exact number of stations that were taken into account during the smoothing process is given in table 4.1.

Figure 4.14 looks at density plots of the inter-station correlations of stations in each of the regions when the smoothed variability is added on to the predictions. More stations contribute to the inter-station correlation calculations in the more dense regions (scenarios two and three) than the less dense regions (scenarios one and reality), but these plots show that the inter-station correlations are similar across scenarios which is highly desirable. There is still a tendency for inter-station correlations to be a little too high, but, as argued in the previous paragraph, this is preferable to them being too low. These plots do show evidence of a few spuriously low inter-station correlations though. The stations contributing to these low inter-station correlations were investigated and it was found that the model was not performing as well at these locations, either in terms of the predictions themselves, or just in terms of the inter-station correlations. The culprit stations were therefore removed in the interest of not creating low quality stations in the benchmark that might bias the results.

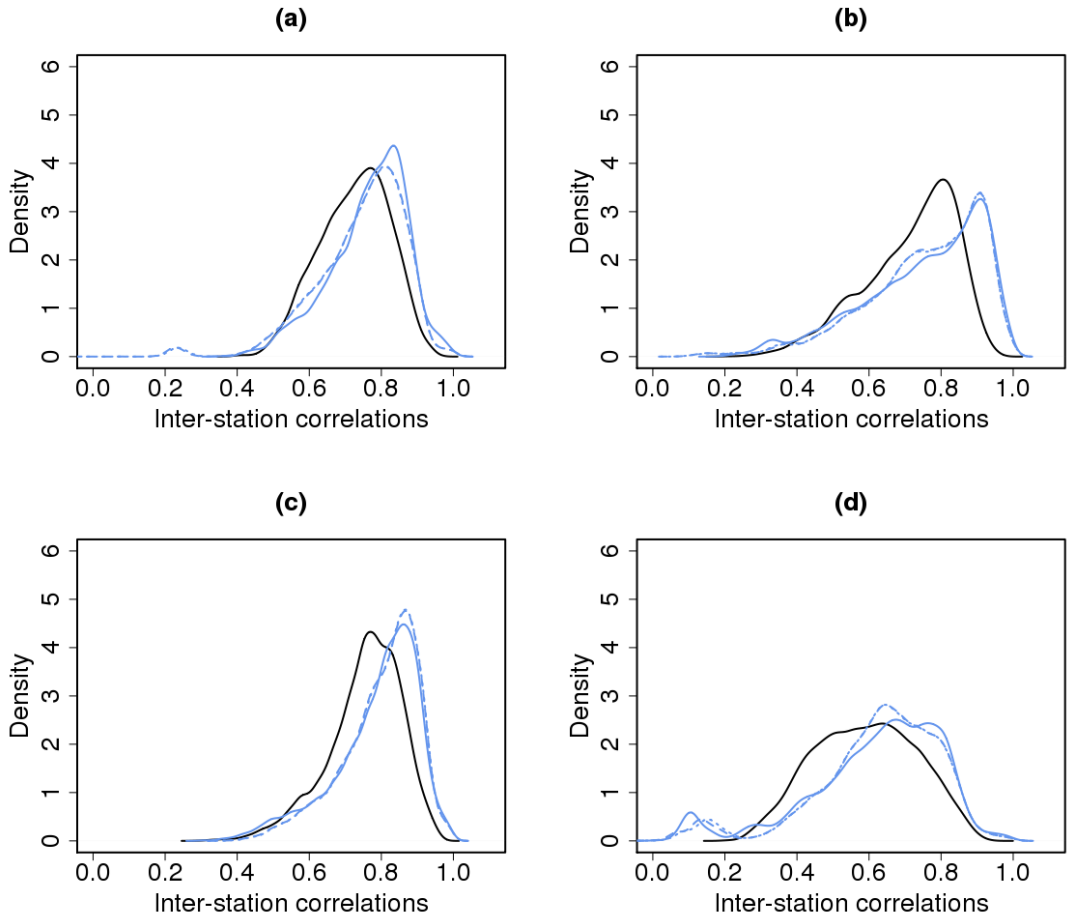

Figure 4.14. Density plots of the inter-station correlations found in observations (black) and predictions with added smoothed variability (blue) from temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Solid lines indicate scenario one, dashed lines indicate scenarios two and three.

One station was removed from the more dense scenarios for Wyoming, it was likely worse as it was reasonably isolated and the only station in a relatively mountainous region. Two stations were removed in the South East less dense version, both on the northern coast of the Gulf of Mexico. One of these also needed to be removed in the more dense version along with one that was very isolated on the western coast of the Gulf of Mexico. Two stations were removed in the less dense scenario of the North East, one on the North Atlantic coast and one on the border with Canada. Neither of these needed to be removed in the more dense scenarios, but one at the very bottom of the focus region, and two others on the North Atlantic coast did need to be. The South West was the region with the most inadequate stations, as expected from its varying climate across the region. However, there were still only six that needed removing in the less dense scenario; two were poor only in the less dense scenario and therefore could be left in in the more dense scenarios; two were poor in more and less dense scenarios and two were poor in the more dense scenario, but had to be removed from both. An additional four then needed removing in the more dense scenarios. All the stations that were removed in the South West were coastal stations, but there were also coastal stations that were perfectly acceptable, indicating that the model performance is not inadequate for all coastal stations.

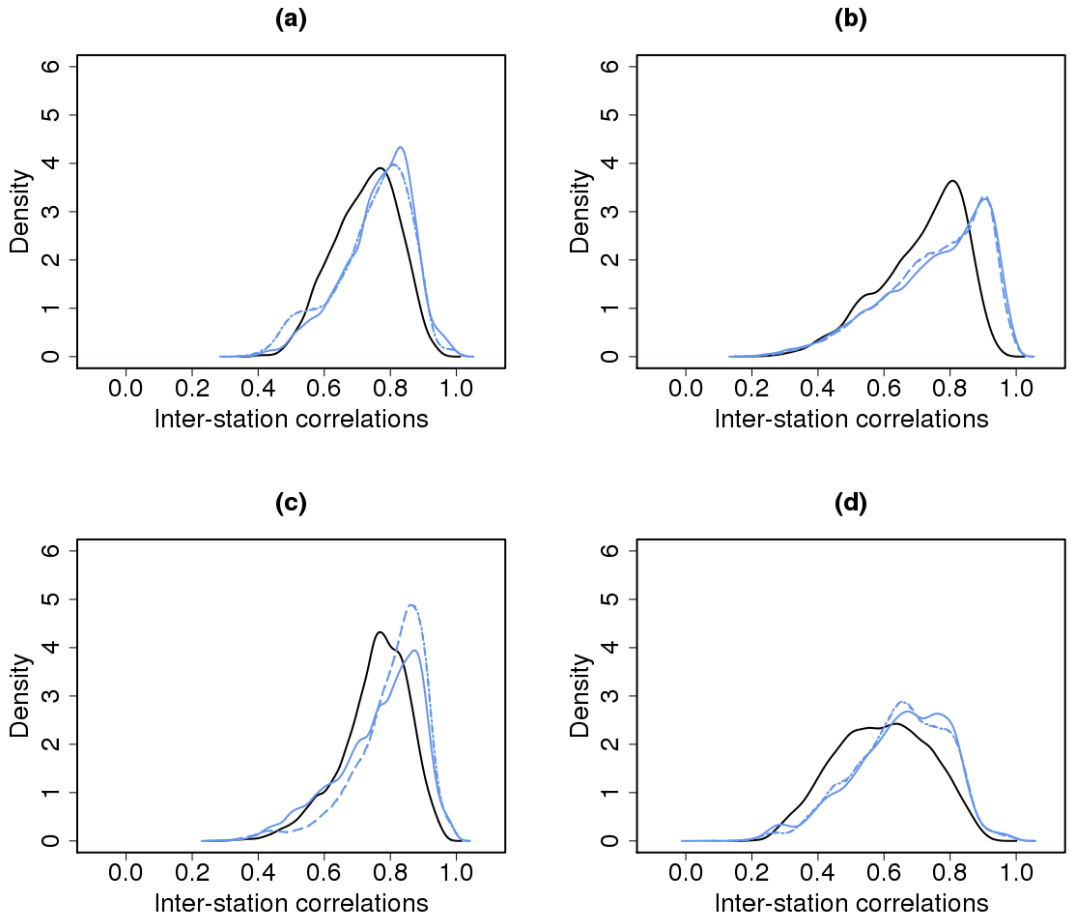

Figure 4.15. Density plots of the inter-station correlations found in observations (black) and predictions with added smoothed variability (blue) from temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Solid lines indicate scenario one, dashed lines indicate scenarios two and three. Here, stations with unrealistic inter-station correlations or predictions were removed.

Density plots of the inter-station correlations in each of the regions after the removal of the dubious stations can be seen in figure 4.15. The higher inter-station correlations have changed very little, but the majority of the unrealistically low inter-station correlations have now been removed.

As stated already, many algorithms look for stations nearby when homogenising data. Therefore, inter-station correlations between stations that were less than 75km apart were extracted from the data and plotted. This distance was chosen as it is the distance used in the neighbour based quality control checks for GHCND [Durre et al., 2010]. Figure 4.16 shows scatter plots of the inter-station correlations for the true and predicted stations within a 75km radius of each other. It is evident that predicted stations are prone to be over-correlated, this is especially evident in the South East and North East. Although this was not the desired situation it was deemed to be acceptable. It must also be remembered that the modelled inter-station correlations are for clean stations, whereas the inter-station correlations they are being compared to are unlikely to be clean.

Figure 4.17 mimics figure 4.12, but with the new smoothed noise. It shows that smoothing the noise has maintained the ability to match the shape of the overall temperature density

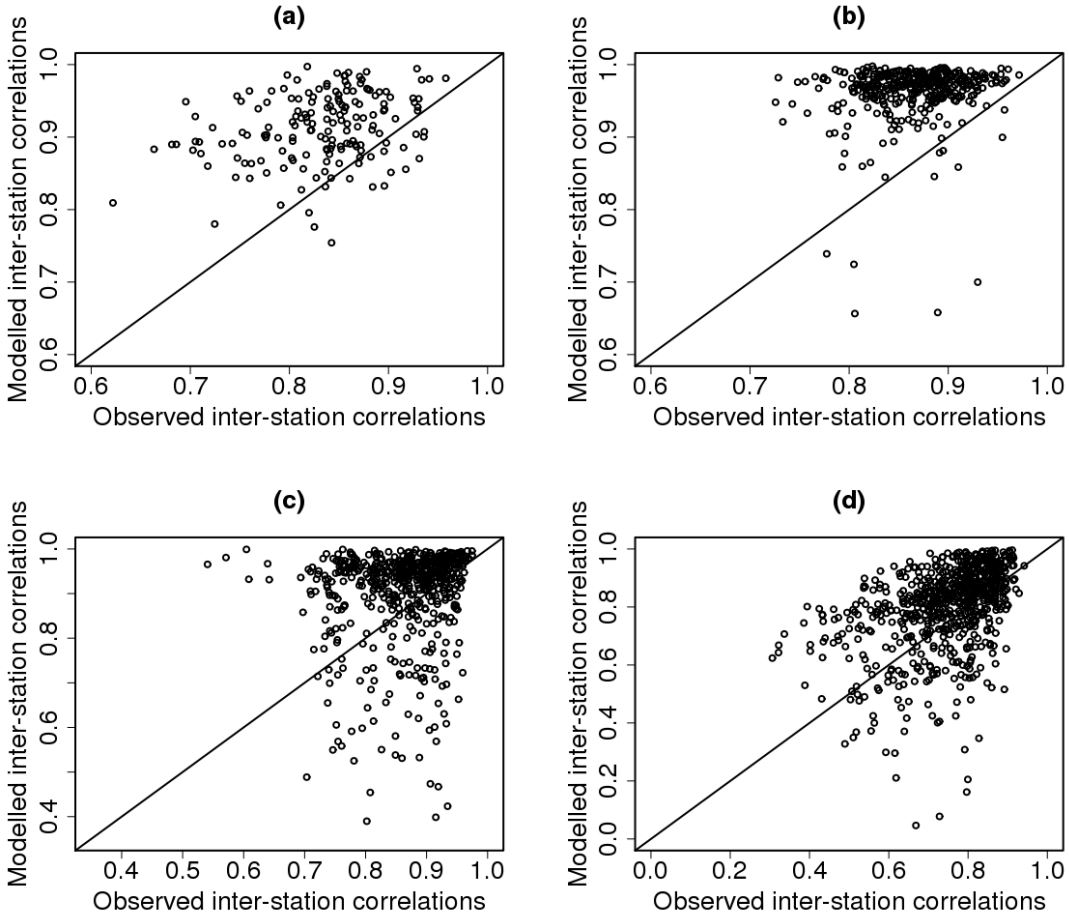

Figure 4.16. Scatter plots of the inter-station correlations found in observations and predictions with added smoothed variability for temperature stations less than 75km apart in (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Points are from scenario 1, but inter-station correlations are very similar for scenarios 2 and 3. The diagonal line is the line $y = x$, points above this line indicate over-estimated inter-station correlations, points below it indicate under-estimated inter-station correlations in the predicted data with added smoothed variability. Note that the x and y axes are consistent within plots, but not across plots. The comparison of predicted and observed inter-station correlations is marginally hampered by the fact that the predictions don't have inhomogeneities in and the observations do. However, the effects of inhomogeneities on inter-station correlations are not believed to be large enough to make this figure invalid.

distributions in each region well, without having to sacrifice the inter-station correlations. However, looking closely at this figure it is evident that some of the most extreme values of the distributions do not match between the observed and predicted data. In Wyoming the distribution is slightly too negative, with between 7 and 27 values in different scenarios being cooler than the cold extremes observed in reality, by up to five degrees, and the uppermost warm extremes being missed. In the South East the opposite is true, though to a lesser extent, with a maximum of only two cold extremes being missed in any scenario and only five warm extremes being gained. In the North East the largest cold extremes are missed, but the warm extremes are relatively well matched for all scenarios. Finally, in the South West, 24 cold extremes are missed in scenario one and one or two are too extreme in scenarios two and three. The upper tail is not well captured in the South West though with 200 warm extremes being missed in the denser scenarios and nearly 600 in scenario one. This amounts to losing the top 4 to 5 degrees of the temperature

distribution in this region. The numbers given here are just the number of times that predictions are outside (or missing) the ranges observed in reality, i.e. extremes are not being compared on like for like days or at like for like stations. When recovery of extreme values was being used as an assessment for algorithm performance, extremes were compared between clean and returned equivalent stations on the same day.

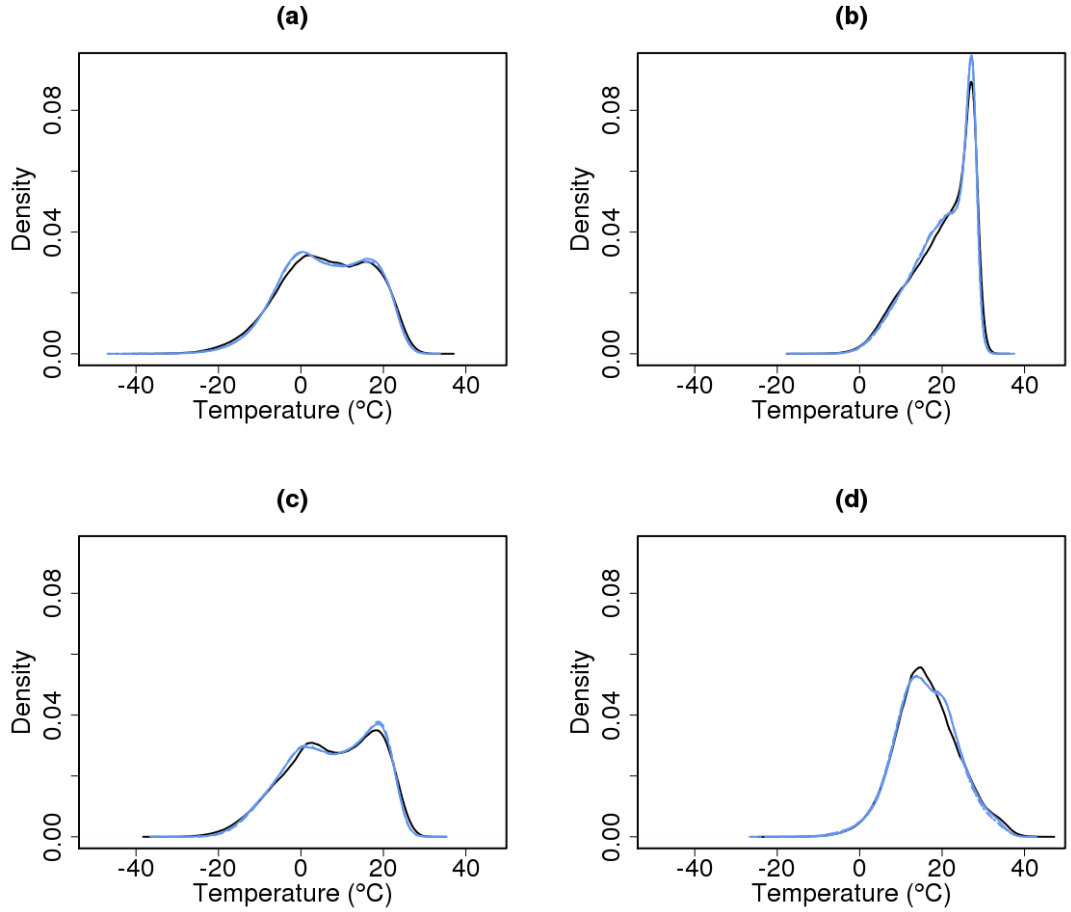

Figure 4.17. Density plots of the temperature distributions found in observations (black) and predictions with added smoothed variability (blue) from temperature station networks for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West.

The autocorrelations within deseasonalised series were also assessed as another way of examining how realistic the created data were. Even after deseasonalisation, daily data exhibit autocorrelation. For the observed data this autocorrelation tails off relatively rapidly, dropping below 0.1 by lag 15 in all regions. For the created data autocorrelations were below 0.1 by lag 15 in all regions apart from the South West. The South West exhibits more persistent auto-correlation, this is observed in reality, but not to the same extent.

Figure 4.18 illustrates the general behaviour of autocorrelation in regional average series. These series were created by taking the mean of all deseasonalised values for each day of the time series for each region separately. The general features displayed in these plots could also be seen by looking at the average difference in autocorrelations between individual observed and predicted series. The autocorrelation at lag one is consistently too low, this tendency to be a little low persists for the first few lags before becoming a tendency to be too high. There were of course a few stations that had predicted autocor-

relations which varied from the observed autocorrelations more than others, but overall the autocorrelations in the created time series were deemed a good match to reality.

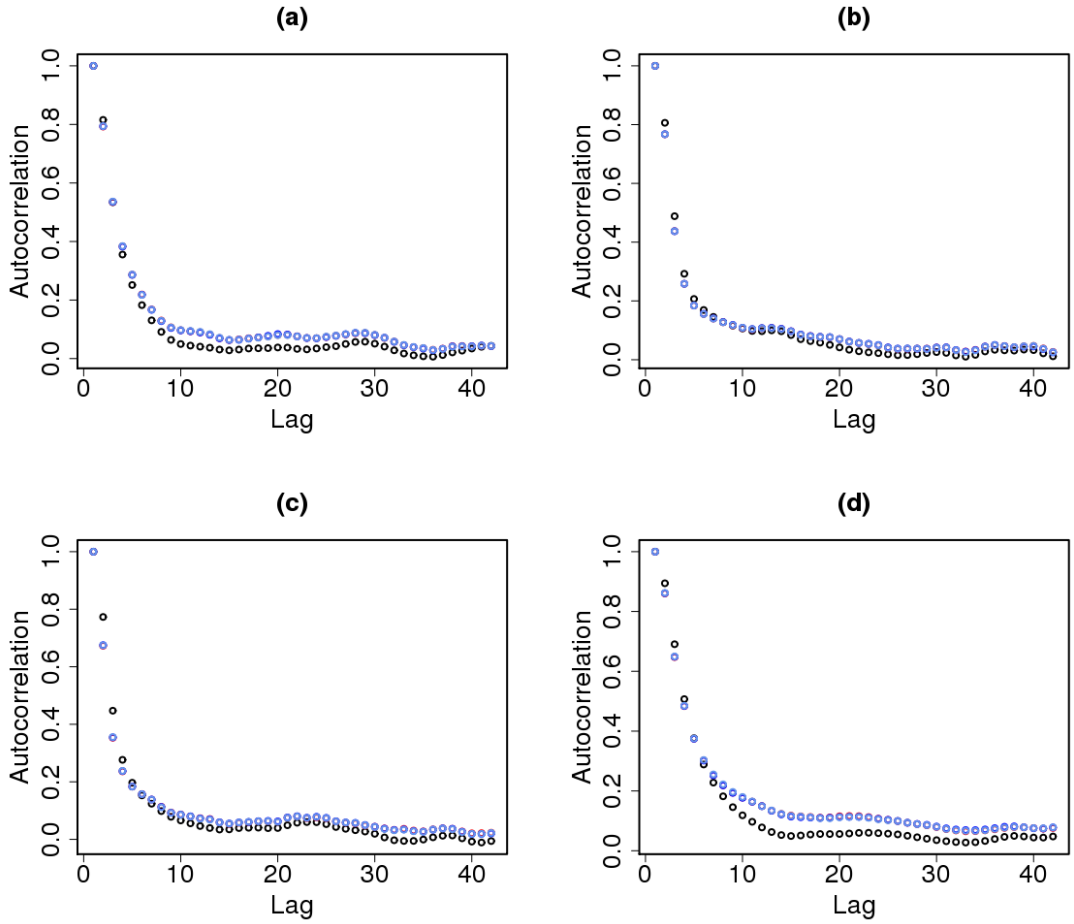

Figure 4.18. Plots to illustrate the autocorrelations up to lag 40 found in the regional average series in observations (black) and predictions with added smoothed variability (blue) for temperature networks in (a) Wyoming, (b) the South East, (c) the North East and (d) the South West for scenario 1. Plots for scenarios 2 and 3 are nearly identical to this.

It is not just autocorrelations of deseasonalised series themselves that are of interest. Also of interest are the difference series created between two highly correlated deseasonalised series. These difference series are commonly assumed to behave as white noise by homogenisation algorithms, but this is not the case in reality as can be seen in figure 4.19. The autocorrelation plots are not the same for all difference series in all regions, but they clearly cannot be assumed in general to be white noise.

A deseasonalised difference series is referring to the difference series created by differencing a deseasonalised station with its most highly correlated neighbour. The stations that are most highly correlated with each other are not necessarily the same in the observed and predicted data. For this reason when creating deseasonalised difference series for predictions the series to difference against was chosen in two different ways. In the first case the predicted station was differenced with the most highly correlated station in the predicted region. In the second case the predicted station was differenced with the predicted station that represented the station that was most highly correlated with it in the observations.

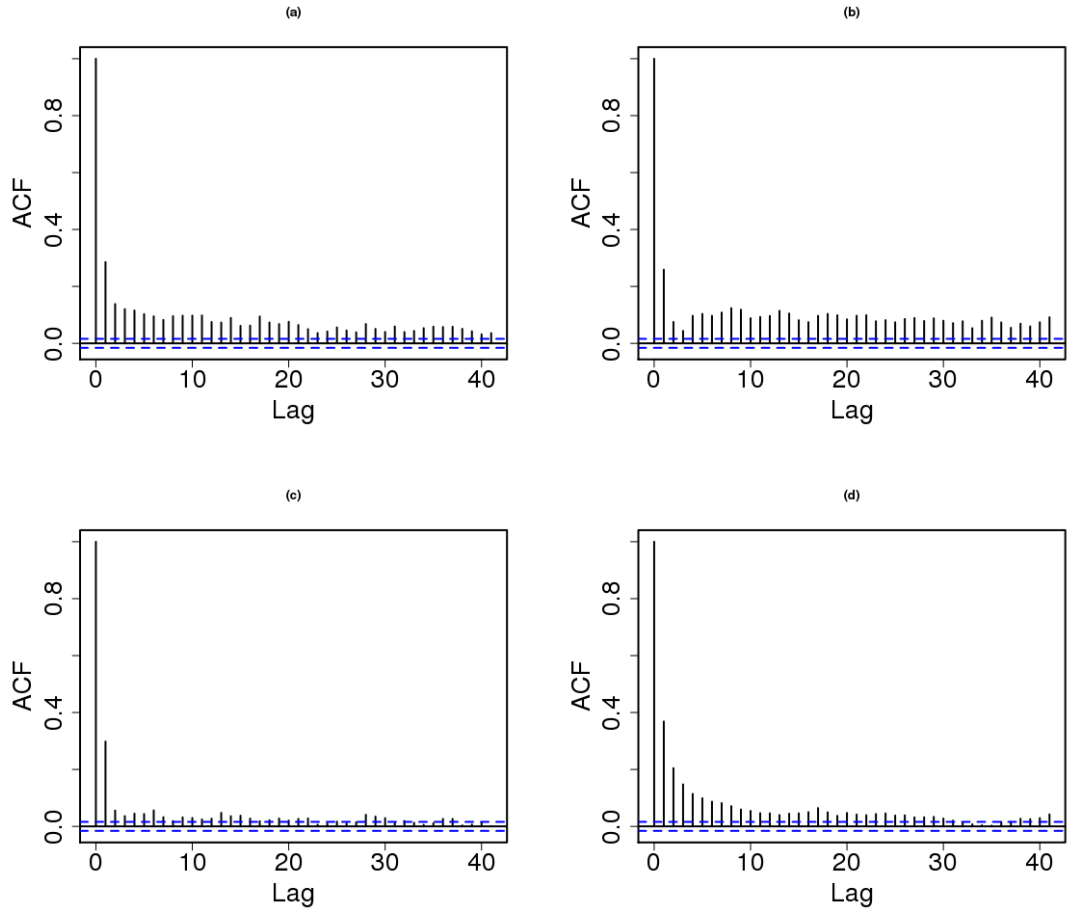

Figure 4.19. Example autocorrelation plots of the difference between a deseasonalised series and its most highly correlated neighbour in observations for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West.

Figure 4.20 looks at the average autocorrelation at each lag across the deseasonalised difference series pairs for all regions and for observations and predictions. It displays average autocorrelations at each lag for the deseasonalised difference series that have been differenced according to their most highly correlated in the observations (addition signs) and those differenced according to their most highly correlated neighbour in the predictions (multiplication signs). The averages were determined by looking at each lag in turn of all deseasonalised difference series and taking the mean autocorrelation value at each lag. It is evident that the autocorrelations in the deseasonalised difference series are considerably higher in the observations (solid circles) than in the predictions (addition and multiplication signs). This means that the created data will be easier to homogenise than the real data, as the created data match an algorithm's white noise assumption where the real data do not.

A likely reason for the lack of autocorrelation in the deseasonalised difference series of the predictions is because the created stations are not as varied as their real world counterparts. This is supported by the higher inter-station correlations found in the predicted data. Low autocorrelations when series have been differenced according to their most highly correlated neighbour in that region is also supported by the smaller standard deviations in the deseasonalised difference series for predictions compared to observations that can be seen in figure 4.21. A lower standard deviation implies less variability in

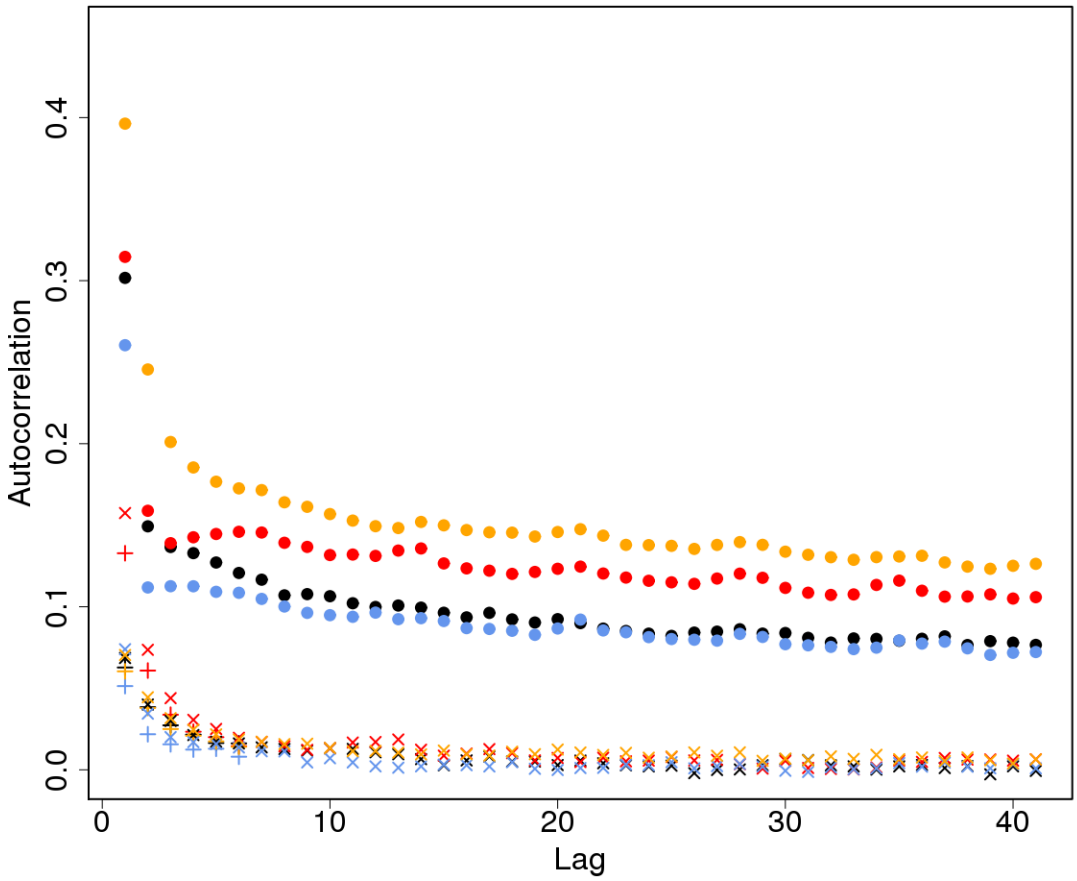

Figure 4.20. Plots to illustrate the average autocorrelation at each lag of the difference series between a deseasonalised series and its most highly correlated neighbour for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Solid circles relate to the observations, addition signs to predictions with added smoothed variability where the most highly correlated neighbour according to the observations has been used and multiplication signs where the most highly correlated neighbour has been determined using the predictions with added smoothed variability themselves. As the values represented by addition and multiplication signs became very similar after lag seven the values represented by addition signs were omitted after this lag. The reason for omitting the addition signs is that any algorithm working with the data would only know which station was most highly correlated with another station in the predictions as they wouldn't have access to the observations. The averages shown here are all taken over scenario one, but the results are similar in scenarios two and three.

the difference series, leading to the conclusion that the series being differenced in the predicted data are more similar than those being differenced in the observed data. Difference series that have been created by differencing according to the predicted station that is the equivalent of the most highly correlated neighbour in the real world do have reasonably realistic standard deviations. Therefore, in these cases the lack of variability cannot be the culprit for the lack of autocorrelation.

To build in temporal autocorrelation using the existing model formulation required the variability added on to the predictions to be smoothed in time as well as in space. This smoothing took place after step 6 of the former outlined prediction process. That is, the loess spatially smoothed noise values, sm_{it}, were smoothed in time before being added back on to the mean predictions. The temporal smoothing was implemented using a weighted moving average on the sm_{it} values. A drawback to this weighted moving average approach is that it increased the autocorrelations for the first n-1 lags for an n

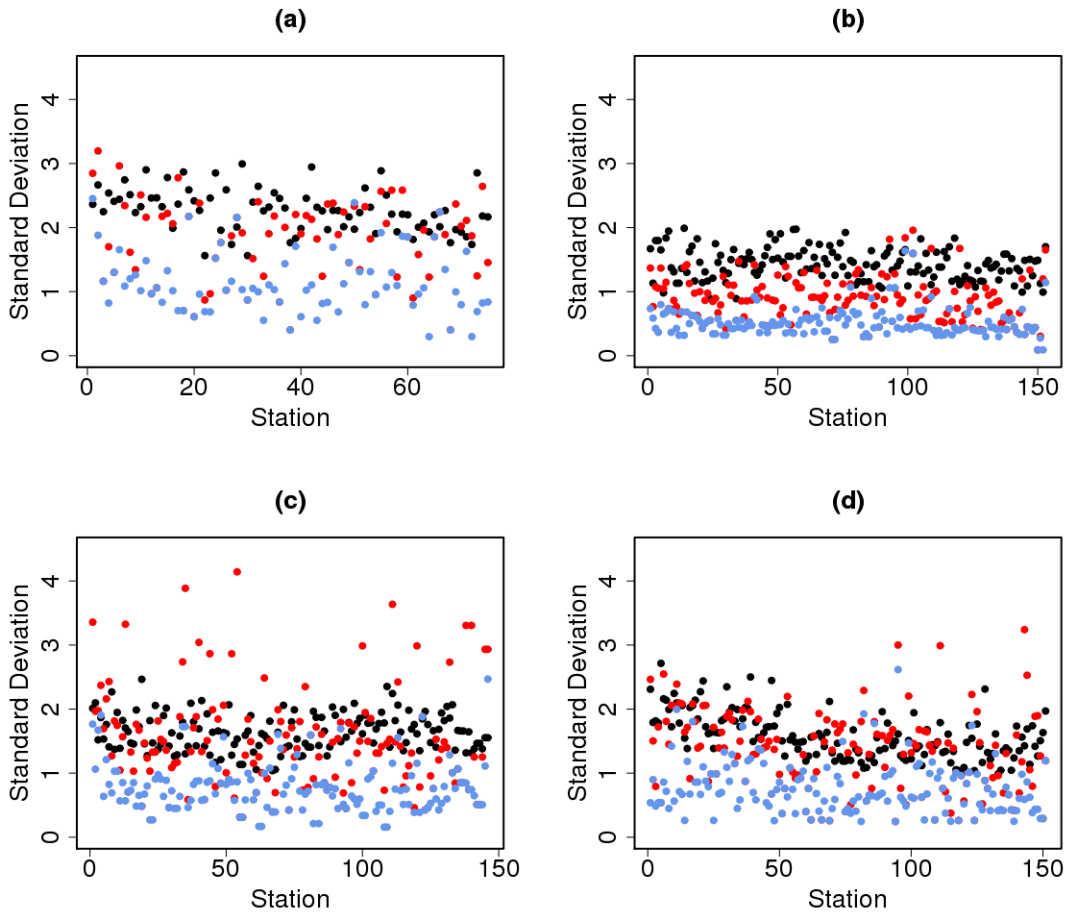

Figure 4.21. Scatter plots of standard deviations of the difference between a deseasonalised series and its most highly correlated neighbour for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Black points relate to the observations, red points to predictions with added smoothed variability where the most highly correlated neighbour according to the observations has been used and blue points where the most highly correlated neighbour has been determined using the predictions with added smoothed variability themselves. These standard deviations are for scenario one, but the results are similar in scenarios two and three.

point weighted moving average, but then the autocorrelations reverted to being similar to before. Even so, a cut off point of weighting had to be decided as, generally speaking, the greater the amount of temporal smoothing the greater the inter-station correlations and the smaller the range of inter-station correlations. Given that inter-station correlations were already relatively high this was undesirable. Therefore, to keep inter-station correlations down to some extent, the amount of spatial smoothing was reduced to only 15 stations. This level of smoothing when combined with temporal smoothing was small enough to restrict inter-station correlations, but large enough to avoid unrealistic extremes entering the data through the variability addition process.

Many lengths of weighted moving average were considered, but a nine point weighted moving average was chosen because it is at lag 8 that the median autocorrelation first drops below 0.1 in the observations on average and, as already stated, it is the first n-1 lags that are affected for an n point weighted moving average. As in section 3.2.1, this threshold of 0.1 was chosen as a relatively arbitrary cut off point, but it was deemed small enough that the autocorrelations dropped below it relatively quickly, meaning that large temporal smooths that had detrimental effects to inter-station correlations did not

have to be used. The weighting of the moving average was decided by looking at median autocorrelations in the observed series. More specifically the median value of the autocorrelation at each lag was calculated to get an idea of how the autocorrelations at lag n changed as n increased and these values were then divided by the sum of all the median autocorrelations up to lag n. The weight at each lag is then $\frac{m_i}{\sum_{i=1}^{n} m_i}$, where m_i is the median autocorrelation in the observed series at lag i and n in this implementation was 9. The division step ensures that the weights sum to one and that lags most highly autocorrelated with the observation in question in reality will maintain this high level of autocorrelation in the predictions. Median autocorrelations were used here instead of mean autocorrelations to prevent any potential outliers having undue influence.

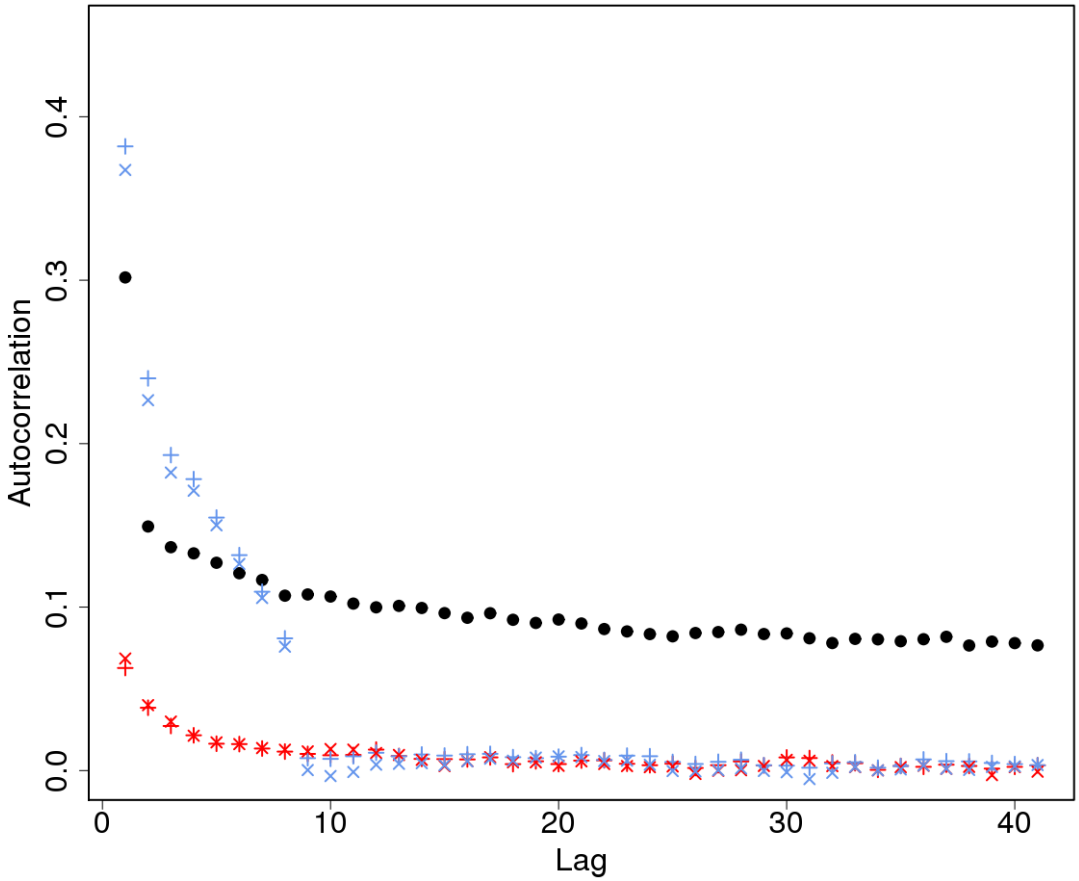

Figure 4.22. Average autocorrelation plots of the difference between a deseasonalised series and its most highly correlated neighbour for Wyoming scenario 1 (red) and 4 (blue). Average autocorrelations here are calculated by taking the mean autocorrelation at each lag from all the most highly correlated deseasonalised difference series pairs. Solid circles relate to the observations, addition signs to predictions where the most highly correlated neighbour according to the observations has been used and multiplication signs where the most highly correlated neighbour has been determined using the predictions themselves.

The average autocorrelations produced using this method can be seen in figure 4.22. Average autocorrelations here are calculated by taking the mean autocorrelation at each lag from all the most highly correlated deseasonalised difference series pairs. This illustrates that the autocorrelations are now too high on average initially. However, when plots are looked at on a station by station level, not shown, this level of spatial and temporal smoothing was deemed to be the best of those analysed. The different levels of smoothing were also assessed after inhomogeneities had been added to the data, which further

suggested a 15 station smooth with a nine point weighted moving average to be the best choice.

Standard deviations for the deseasonalised difference series in this scenario were also investigated. They can be seen plotted in figure 4.23, which shows that they are all too similar to each other. However, given that these values are of approximately the right magnitude, the data were deemed sufficient according to this criteria.

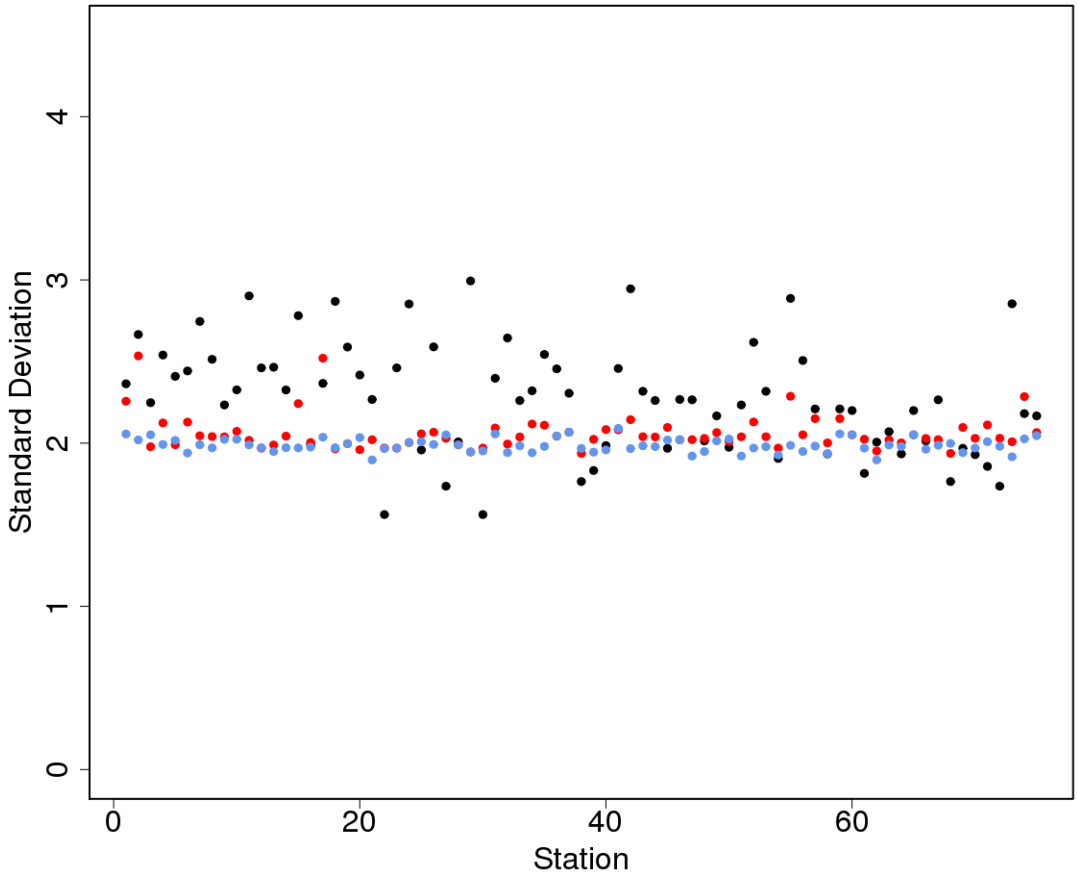

Figure 4.23. A scatter plot of standard deviations of the difference between a deseasonalised series and its most highly correlated neighbour for Wyoming. Black points relate to the observations, red points to predictions where the most highly correlated neighbour according to the observations has been used and blue points where the most highly correlated neighbour has been determined using the predictions themselves.

This scenario with temporal smoothing in addition to spatial smoothing was only implemented in Wyoming and was termed scenario 4. The reason for focusing on Wyoming was that this was the region homogenisers were asked to prioritise, and also the smallest region (just 75 stations). Remaining scenarios were left with no temporal smoothing because of the compromises smoothing temporally raised. Proceeding in this way allows comparisons to be made based on whether meeting or violating the assumption of deseasonalised difference series being white noise affects algorithm performance.

Throughout this discussion inter-station correlations have been a point of interest, figure 4.24 therefore shows the inter-station correlations for scenario 4, both for all stations and for stations less than 75km apart. It is evident that the very highest inter-station correlations are being missed, but on average the inter-station correlations are similar to those found in scenario one; this is desirable as will be explained further in chapter five. The

highest inter-station correlations tend to be those at the shorter distances, thus there is now more under-estimation of the observed inter-station correlations at shorter distances, which is not desirable as these are the stations that will most likely be used as reference series. However, it is evident that not many stations are affected by this problem and therefore the scenario was still deemed fit for purpose because of its good performance on average. There is also a reduction in the range of inter-station correlations at short distances because of the loss of the highest correlations. If all distances are looked at this loss is not as noticeable and the overall over- and under-estimation of inter-station correlations is relatively similar in this scenario and scenario one.

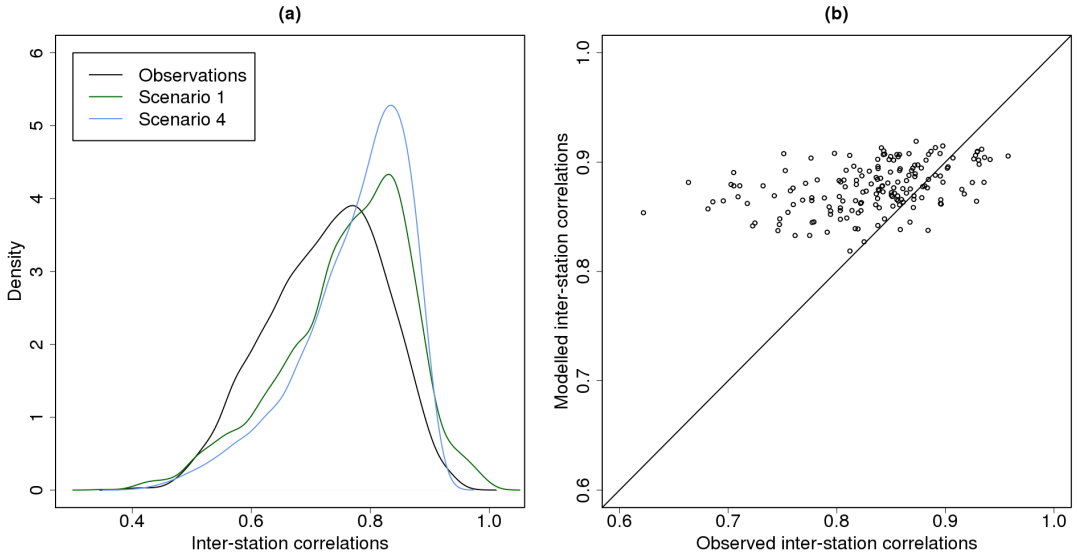

Figure 4.24. (a) A density plot to illustrate inter-station correlation distributions for stations in Wyoming, for observations (black), scenario 1 (dark green) and scenario 4 (blue). (b) A scatter plot to compare observed and predicted (scenario 4) inter-station correlations for stations within 75km of each other.

It is of course also important that the predicted data match the observed data well and this can be seen to be the case in figure 4.25. There are only 3 values that are more extreme in the predicted data than they are in the observed data. One at the upper end and two at the lower end.

The final assessment that all clean scenarios underwent was the application of the Pairwise Homogenisation Algorithm (PHA) [Menne and Williams JR., 2009]. This algorithm is automated and requires that the data are aggregated to the monthly level; it then searches for inhomogeneities using pairwise comparisons between stations. The inhomogeneities found are attributed to the station believed to be the culprit and the magnitude and uncertainty of the shift required to homogenise the data is then provided.

Table 4.2 shows the number of 'inhomogeneities' identified in the clean scenarios that have a shift magnitude greater than the shift uncertainty. It also shows how many clean stations are therefore affected. Some of these 'inhomogeneities' will be false alarms of the PHA, i.e., the PHA will have wrongly said that there is a change point where in fact no change point exists, others could be genuine shifts that the modelling process has inadvertently produced. However, it should be noted that, there are no occasions where an 'inhomogeneity' was found in the clean data at the same time and for the same station

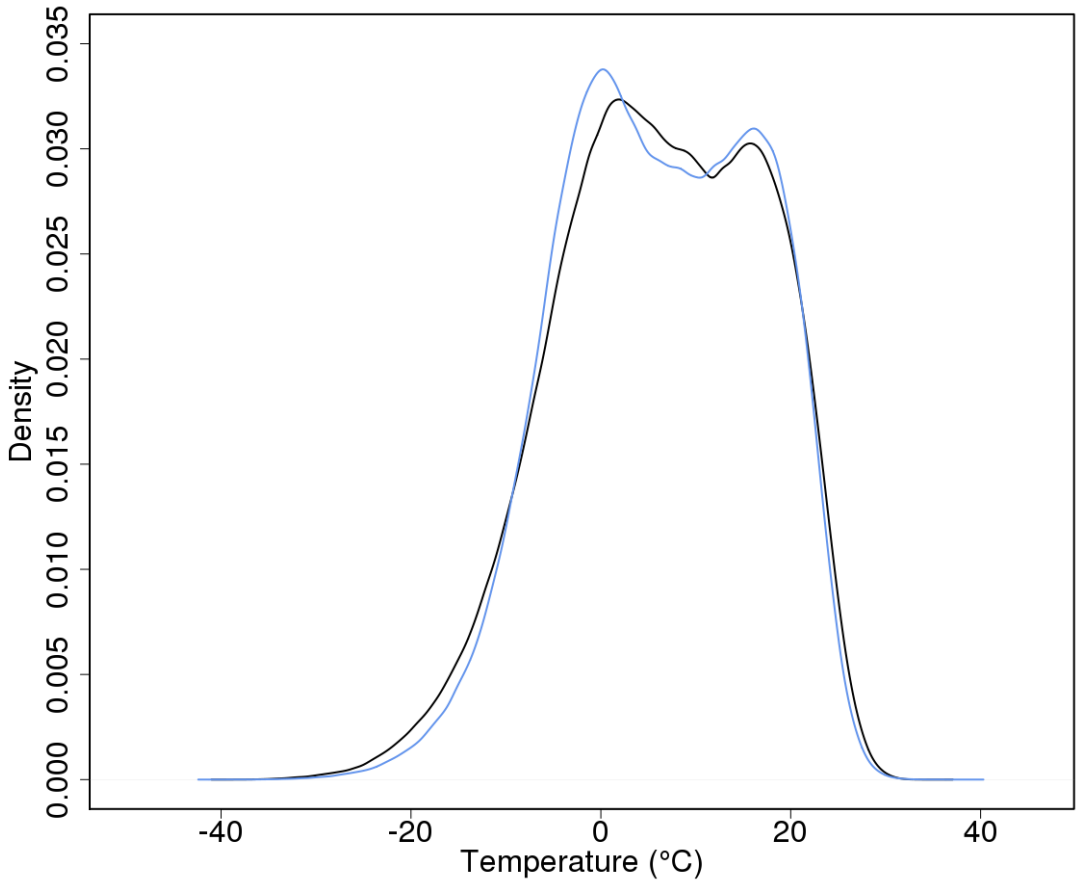

Figure 4.25. A density plot to illustrate the temperature distributions for observations (black) and scenario 4 predictions (blue) in Wyoming.

as the PHA found an inhomogeneity in the observations. Therefore, the author has high confidence that this modelling process can successfully take inhomogeneous data and create homogeneous series from them. The greatest numbers of 'inhomogeneities' are found in the South West, this is as expected as it is the most complicated of the regions.

A note was made of all these 'inhomogeneities' found in the clean series and if the same inhomogeneities were found by any of the participating algorithms then they were not counted as false alarms, this ensures that algorithms aren't wrongly penalised for what could feasibly be a modelling mistake. If the PHA was an algorithm that was being assessed in this thesis then discounting these 'inhomogeneities' would not be a fair approach as it would favour the PHA over other algorithms. There is still a possibility that this approach will favour 'PHA-like' algorithms, but 'PHA-like' algorithms could be classed as those which use reference series and, as all the algorithms in this study use reference series, the author believes that none should have an unfair advantage because of the exclusion of the PHA 'hits'.

Table 4.2. A table to show how many 'inhomogeneities' were identified in each of the clean scenarios and therefore how many stations are affected.

Region and Scenario	Number of stations in total	Number of stations with inhomogeneities with magnitude greater than uncertainty	Number of identified inhomogeneities with magnitude greater than uncertainty
Wyoming 1	75	4	4
Wyoming 2	158	5	5
Wyoming 3	158	16	16
Wyoming 4	75	3	3
South East 1	153	11	13
South East 2	210	9	9
South East 3	210	15	15
North East 1	146	11	11
North East 2	207	8	9
North East 3	207	11	11
South West 1	151	46	50
South West 2	222	25	28
South West 3	222	31	31

4.3. Discussion

The model selected to be used for the clean data creation in this thesis contains many physically justified explanatory variables. However, as stated in section 4.1.2 there are inevitably others that could be included, for example, a topography categorical variable like that used in the analyses of Pepin and Norris [2005]; Pepin and Lundquist [2008], or a land cover variable. Adding in more variables could make the models still more detailed, however, the author believes that the current models are fit for purpose. Should a simpler model be sought then models with alternative combinations of variables could be compared using stepwise regression based on the Akaike Information Criteria.

In addition to changing the explanatory variables in the model, changes could be made to the model output by changing the formulation of the smooths within the model. The investigation carried out in this thesis looked at simultaneously increasing the allowable degrees of freedom for all the smooth functions and this was found to lead to the possibility of unrealistic results. However, further investigation into altering only certain smooths would be an interesting area for extended research. This investigation was not carried out here as the model outputs in this thesis were deemed fit for purpose according to the criteria in section two of this chapter.

In the area of assessing the homogeneity of the created clean data the author is happy with the results of the PHA which shows few 'inhomogeneities'. Whether or not these 'inhomogeneities' should be discounted from the final analysis was discussed in section 4.2.2 and the conclusion was that it is acceptable in this study to discount them as the PHA was not one of the participating algorithms, and all algorithms could be argued to be 'PHA-like'. However, the author would recommend that such 'inhomogeneities' should not be discounted if absolute homogenisation algorithms (those that homogenise without

a reference series) were present in the study as such a decision could unfairly favour 'PHA-like' relative homogenisation methods. Equally, if discounting or not discounting such inhomogeneities makes little difference to the overall conclusions about algorithm performance, the step removing PHA 'inhomogeneities' could be omitted in the interest of a simpler study. This topic of the PHA and its 'inhomogeneities' is revisited in section 2.3 of chapter seven in light of the results of the current study.

4.4. Summary

This chapter has given the reasons for the inclusion of the chosen model variables. It has explained the Gamma GAM model; its formulation and workings and the justifications for its use. These justifications include its ability to use other climatic variables to model temperatures, which will be exploited in the following chapter; its ability to cope with short or incomplete records; and its ability to create new stations where none currently exist as will be illustrated in section 2.2 of chapter 5.

This chapter finishes with the production of clean data that can be used as benchmarks for the testing of homogenisation algorithms. These data are shown to have reasonable, if high, inter-station correlations that have been decided using sophisticated smoothing mechanisms; good autocorrelations in deseasonalised series; and autocorrelations that are too low, but match algorithmic assumptions, in the deseasonalised difference series. A further dataset was created, using spatial and temporal smoothing, with increased autocorrelation in deseasonalised difference series to allow the assessment of the impact of the false assumption that deseasonalised difference series are white noise. Finally, all created clean scenarios were run through the pairwise homogenisation algorithm to ensure that they were indeed clean. The results of this algorithm application suggested that, although some potential inhomogeneities were found, the modelling process had succeeded in making largely clean data. The following chapter will detail the creation of the inhomogeneities to be added on to these clean data and the different versions of the data released to the homogenisation community.

5. Building and Evaluation of the Released Data

The previous chapter has explained how realistic clean daily temperature data were created as a benchmark for the assessment of homogenisation algorithms. These data were created using a generalised additive model with observations and reanalysis data as inputs to ensure that the model drew information from reality instead of solely relying on assumptions about the real world. The focus areas for this study are four regions in North America, chosen for their station coverage and diverse climates. The same model formulation was used for each of these four regions, but the models themselves were fitted to each region separately to maximise the information that was captured from these climatologically diverse regions.

This chapter details the investigation and creation of realistic error structures and how these were added on to the benchmark clean data in order to assess algorithm performance in response to different inhomogeneities. The structure of the GAM used for data creation is employed effectively to allow for the creation of these inhomogeneities in previously unexplored ways. The four different release scenarios created and the reasons for their creation are explained in this chapter and the chapter concludes with an overview of the characteristics of these scenarios allowing the reader to make comparisons between them.

5.1. Inhomogeneities to be investigated

There are many different factors that affect the homogeneity of a temperature series. The factors most commonly identified are changes in observation practice or instrumentation, changes in station location and changes in station surroundings, see for example Peterson et al. [1998], Reeves et al. [2007] or Trewin [2010]. Other studies also identify additional causes; shelter deterioration for example Lopardo et al. [2014] or shelter changes Hubbard and Lin [2006]. As the literature commonly identified these main causes, and personal communication with those working on homogenisation of temperature data also identified them as prominent, this study focussed on representing three of these most prominent inhomogeneity causes: station relocations, shelter changes and changes in station surroundings. The reason for not focussing on observation practice changes explicitly is that, certainly in the US, this has already been the focus of numerous studies [Quayle et al., 1991; Hubbard and Lin, 2006]. Henceforth the changes focussed

on shall be referred to as shelter changes, station relocations and urbanisation, which is a specific example of a change in station surroundings. Added inhomogeneities could be considered to mimic more than simply these three issues, as some changes will affect temperature records in similar ways, for example, both urbanisation and shelter deterioration would commonly cause trend inhomogeneities. When adding inhomogeneities using the explanatory variables however, the justifications for the perturbations used are based on mimicking these three specific inhomogeneities.

This study has sought to represent true inhomogeneities well, but not restrict inhomogeneities investigated to only those known to have occurred in the US over the period of record. It has also not restricted the locations at which inhomogeneities can occur, therefore urbanisation inhomogeneities can occur at any site, regardless of real world population density. This will enable the methodology of the project to be generalised to other regions easily. It also means that real world metadata (data about the data) cannot be used to homogenise the created data. This is advantageous in this benchmarking study as it avoids any one method having an unfair advantage over others, although the use of metadata for daily homogenisation at this time is minimal. Metadata would currently be primarily used to verify changes that have been detected and not necessarily in the detection process itself. Metadata could be created for a future iteration of this project, but at the time of data creation it was considered an unnecessary complication that wouldn't be used sufficiently to justify its inclusion.

5.2. Scenarios created

It was desirable to assess homogenisation algorithm performance in different circumstances. Some of these circumstances were covered by the creation of temperature series that represented different climatic regions in North America. However, it was also desirable to test algorithm performance in response to different data characteristics (e.g., autocorrelation) and station availability. This was done by the creation of four scenarios.

5.2.1. Scenario 1

This scenario can be thought of as the current best guess for the world. The stations provided are those that are at least 75% complete in the observed data over the period of 1970-2011, minus the few in each region that the model did not reproduce adequately. All the inhomogeneities listed in the previous section are allowed to be present in this scenario, thus an algorithm's ability to detect and correct for both trend (urbanisation) and step (relocation or shelter) changes can be investigated. Figure 5.1 shows the locations of the stations in this scenario in black for all four regions provided. There are 75 stations in Wyoming, 153 in the South East, 146 in the North East and 151 in the South West.

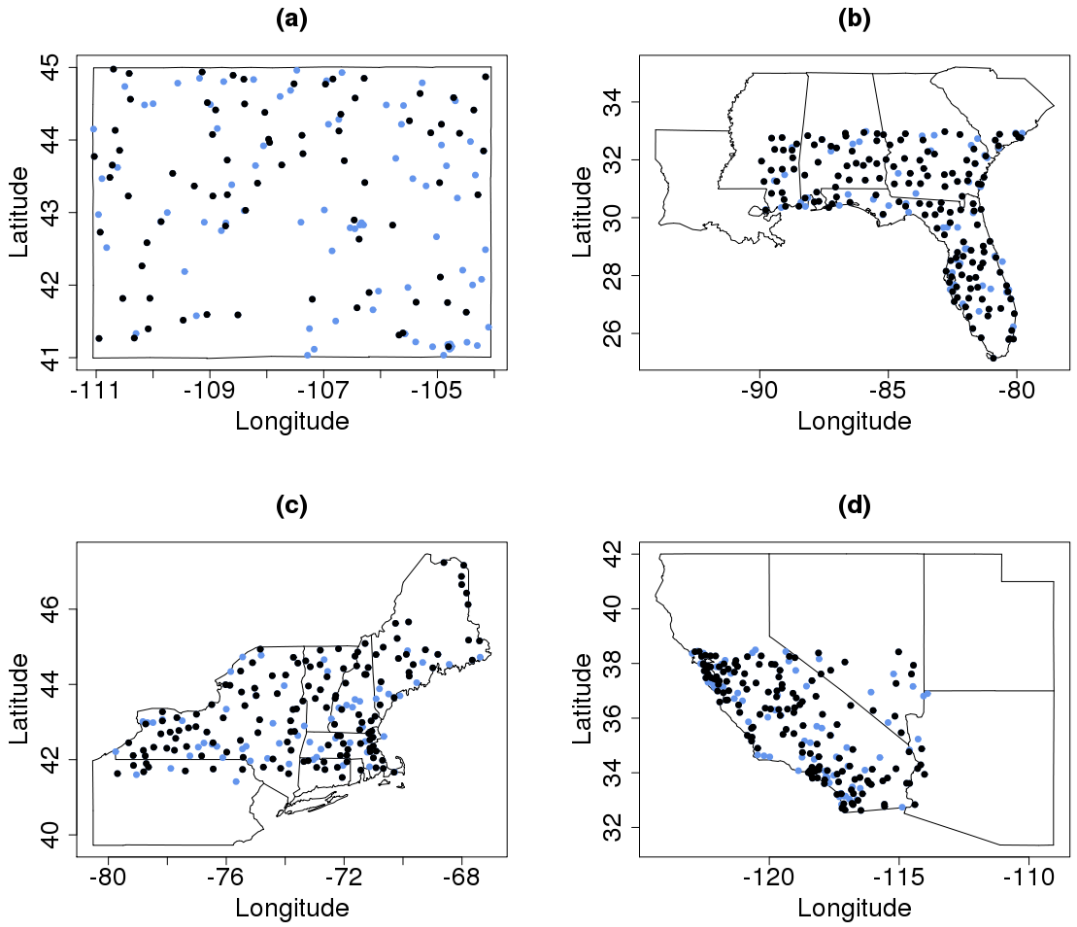

Figure 5.1. Location of the temperature stations provided in scenario 1 (black) and the additional stations provided in scenarios 2 and 3 (blue) for (a) Wyoming, (b) the South East, (c) the North East and (d) the South West.

5.2.2. Scenario 2

It was possible to work out the density of stations present in each region per unit area using Python code created by Peter Killick from the Met Office. One unit area on a Python map is approximately 9500km^2. For scenario one the average number of stations per unit area differed across regions, being 2.82 in Wyoming, 4.38 in the South East, 4.22 in the North East and 4.09 in the South West. These differing station densities mean that differences in algorithm performance may arise because of a greater or lesser availability of suitable reference stations and not just because of differences in climatic regions. Therefore, in scenario two, the station density was made uniform across each of the regions, with an average of 6 stations per unit area. This amounted to having 158 stations in Wyoming, 210 in the South East, 207 in the North East and 222 in the South West. It should be noted that these numbers are slightly less than was originally designed owing to the removal of some stations where the model performed inadequately, as stated in chapter four section 2.2. Increasing the station density to 6 stations per unit area was deemed an appropriate increase to investigate the change in algorithm performance, without the processing becoming too computationally expensive. It was decided to increase the station density and not decrease it, as it is far easier for algorithm developers to decrease the station density themselves to test performance. Decreasing

the station density can be done simply by omitting some of the given stations, whereas the creation of new stations required interpolation to new data points and new model predictions. Creating new stations at new locations allowed the capabilities of the GAM modelling approach to be exploited.

Once the necessary increase in stations was decided the locations for these stations had to be chosen. One way to achieve this would be to randomly generate latitudes and longitudes for the necessary number of stations and then proceed to acquire the GAM's input variables for these points. This was the methodology employed when creating the station relocation options as will be explained in section three of this chapter. An alternative method was to use the existing station locations stored in the GHCND database that were unable to be used in the initial modelling stage of this project owing to insufficient record completeness. These extra GHCND stations have the advantage that they are in plausible locations and are known to have been used for gathering observations at some point in the past. In order to spread the sample of new stations out over the region in question the candidates for selection were first ordered by longitude and were then subsampled at regular intervals to achieve the desired number of additional stations. These stations were then plotted, as can be seen in blue in figure 5.1, to ensure that a reasonable coverage had indeed been achieved. Missing data were added to these stations in the following manner. Stations that also existed in scenario one had the same missing data as their equivalent scenario one stations. Stations that were newly created had missing data from stations chosen randomly, without replacement, from another modelled region. This ensured that their level of missing data was realistic, but also that no two stations within a region could have the same missing data.

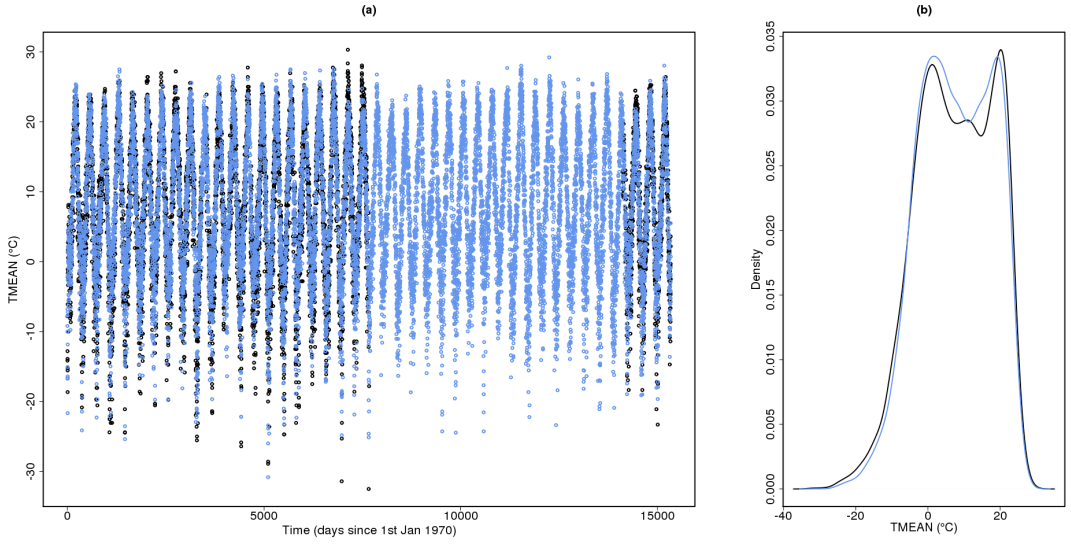

Figure 5.2. A time series (a) and density plot (b) to illustrate the capabilities of the GAM at predicting station temperature data for stations that were not included in the model building stage. Black points and lines are from the observations and blue points and lines are from the predictions.

Figure 5.2 shows an observed (black) and predicted (blue) station. This station is station 113 in Wyoming scenario two; the observed station could not be included in the model building process as it was not 75% complete, but it did have sufficient data to be able to form both a time series and density plot that the predicted data could be compared

to. It can be seen that the predictions match the observations well, thereby providing an example to testify that the GAM is able to predict mean temperatures for stations who's data were not included in the model building process.

5.2.3. Scenario 3

This is the simplest of all the scenarios created and the one that algorithms were anticipated to have the best performance on. The station density is equal to that in scenario two, making it better than the station density of stations that are at least 75% complete in reality, and there are no urbanisation inhomogeneities. Urbanisation inhomogeneities were expected to be more difficult to locate than shelter changes or station relocations as urbanisation inhomogeneities are trend inhomogeneities. Trend inhomogeneities can be difficult to detect owing to the possibility of them starting at different points with respect to different neighbouring series Menne and Williams JR. [2008]. Having said this Hausfather et al. [2013] showed urban effects being correctly identified and adjusted for suggesting that the removal of trend inhomogeneities is not beyond the capabilities of all algorithms. Hausfather et al. [2013] found that urbanisation inhomogeneities could be best removed with greater station densities suggesting that urbanisation inhomogeneities should be easier to find and remove in scenario two than scenario one.

A comparison between scenario three and scenario two allows the assessment of algorithm performance with and without the presence of artificial trends. This scenario is similar to the investigation carried out by Williams et al. [2012] where homogenisation algorithm ability was only assessed in the presence of varying step inhomogeneities, but not trend inhomogeneities. However, the focus of Williams et al. [2012] was on the algorithm's ability to recover the true trend at the regional scale, whereas this study focusses on a wider range of validation measures, as will be explained in chapter six.

5.2.4. Scenario 4

Whilst the other scenarios incorporate all four focus regions this scenario was created only for Wyoming. This was because Wyoming was the area participants were asked to prioritise owing to the model having good performance here and it being the smallest of all the regions meaning that the application of a manual algorithm would be possible. As stated in chapter four section 2.2, it was scenario four where autocorrelations in the created data were focused on to try and ensure a better match to observed autocorrelations than was found in the other scenarios. To be able to most easily assess the impact that these autocorrelations had, the type and location of inhomogeneities added to this scenario were exactly the same as in Wyoming scenario one. The size of perturbations or constant offsets made were the same as in Wyoming scenario one too, however final sizes of the inhomogeneities were not necessarily identical owing to the underlying clean data not being identical.

5.3. Inhomogeneity creation and addition

The following step by step process was implemented in order to create and then add inhomogeneities into the clean series:

1. Take a clean scenario from a single region.
2. For each station in that scenario allocate points of an inhomogeneity using a Poisson process.
3. Decide on the type (station relocation, shelter change or urbanisation) of each inhomogeneity by generating a random value from a uniform distribution on zero to one. (The types of inhomogeneity allowed are governed by the scenario chosen, as explained in the previous section.) Allocate inhomogeneities as evenly as possible, but ensure that an urbanisation inhomogeneity does not happen more than once in any given series.
4. Decide on the method of addition for an inhomogeneity by generating another random value from a uniform distribution on zero to one so that 30% of inhomogeneities are constant offsets and 70% come from explanatory variable perturbations.
5. Make any necessary changes to the predictor variables in the GAM to implement explanatory variable changes and then use this model to predict new inhomogeneous data.
6. Add the original smoothed noise and any constant offset inhomogeneities to the predicted inhomogeneous data so that the noise structure is the same in clean and inhomogeneous scenarios.

The following sections give more detail for these steps.

5.3.1. Inhomogeneity locations

Points where an inhomogeneity arises in a time series can be thought of as rare events, therefore, they can be modelled as a Poisson process. In a Poisson process the time between events is exponentially distributed with a mean time between events of θ and events are independent of each other [von Storch and Zwiers, 2001; Venema et al., 2012]. Inhomogeneities occur on average every 15-20 years in a US monthly dataset analysed by Menne et al. [2009] and Venema et al. [2012] found a similar frequency of inhomogeneities in Europe. However, those analyses will not have captured all inhomogeneities as some will have been too small and the detection methods will not have been perfect. Therefore, the frequency of inhomogeneities found in those studies can likely be considered a conservative estimate of the true number of inhomogeneities present. For this reason, in this study, inhomogeneities were inserted into the series on average every thirteen years. This value amounts to a mean of three inhomogeneities per 42 year series where change points were not allocated to the final two years of the period.

Although a lack of change points in this period is a simplification of reality it was deemed appropriate as a first step in benchmarking the performance of daily homogenisation algorithms. This simplification is common to other studies [Venema et al., 2012], this is because many methods use the final years of a record as the reference period against which to search for inhomogeneities and because methods also traditionally struggle to find inhomogeneities close to the end points of a series. Homogenisers were asked to use the most recent homogeneous sub period for reference in this study, but they were not told the length of this last HSP.

Note here that the final two years were free of change points, but they were not necessarily free of the effects of an inhomogeneity. The times of inhomogeneities represented the time of a shelter change or station relocation; or the midpoint of an urbanisation inhomogeneity. Thus, no inhomogeneities started or ended in the final two years of the record, but an urbanisation inhomogeneity could persist throughout it. Because the time allocated for an urbanisation inhomogeneity represented its midpoint the actual number of change points per urbanisation inhomogeneity is two. This will further increase the frequency of change points beyond the aforementioned average of one per thirteen years. However, as will be explained later, some of these inhomogeneities will have no noticeable effect on the series and therefore the frequency of noticeable inhomogeneities was deemed to be acceptable.

As participants were asked to homogenise relative to the most recent time period, inhomogeneities were propagated backwards in time. That is, all the points in time from each change point to the beginning of the series were affected by the added inhomogeneity. The effects of inhomogeneities combined, therefore, the earliest period of a record had most inhomogeneities acting on it. However, there was nothing to stop multiple inhomogeneities cancelling each other out, which is why the size of an inhomogeneity was determined using its relative effect size and not the cumulative size of all inhomogeneities acting at a certain point. Further detail about inhomogeneity size classification is given in section four of this chapter.

5.3.2. Inhomogeneity creation

In chapter four, three prominent benchmarking studies were identified, those of Williams et al. [2012], Venema et al. [2012] and Willett et al. [2014]. All three of these studies were blind, meaning that the truth about the data was not revealed until it was known that contributions were finalised. In the first of these studies the benchmarking investigation was assessing variants of the pairwise homogenisation algorithm; specifically it looked at the algorithms' ability to recover true climate trends in the presence of various inhomogeneity structures. The structures introduced included large change points, clustered change points and many small change points randomly inserted into the temperature series. Some of these change points were supported by metadata, others were not. No trend inhomogeneities were added, and the size of the inhomogeneities added was kept constant throughout the period over which they acted. These sizes were drawn from nor-

mal distributions with varying standard deviations (a maximum of 1) and various means, depending on whether the inhomogeneities were associated with a sign bias, which was allowed to be positive or negative. The study of Venema et al. [2012] investigated trends and seasonally varying inhomogeneities. They used a Poisson process when locating non-clustered inhomogeneities and a normal distribution to decide their size with a standard deviation of 0.8°C for the mean size and a standard deviation of 0.4°C about the mean to decide the size of the seasonal cycle. Spatially clustered change points were added into 30% of the time series in any network assigned this change point type; this change point occurred at the same time in each series, but its magnitude was allowed to vary slightly in the different stations affected. A uniform distribution between 30 and 60 years was used to decide the length in the case of trend inhomogeneities and these had to be fully contained within the focus time period. The size of the trend was selected from a normal distribution with a mean of 0.8°C. The study of Willett et al. [2014] has not yet reached the stage of creating error structures. However, it advocates the production of multiple scenarios ranging in difficulty, the inclusion of both step and trend inhomogeneities and also the creation of error structures using information from other climatic variables.

In the present study inhomogeneities were added in two different ways; by perturbing explanatory variables in the GAM or by adding constant offsets to the clean series. These added inhomogeneities can mimic both step and trend inhomogeneities and those inhomogeneities that are added using information from other climatic variables will vary seasonally, thus allowing the exploration of algorithm performance when non-constant offsets are present. No inhomogeneities are supported by metadata in this study for the reasons given in 5.1 and also because it is known that metadata is not always complete or even present and therefore methods need to be found that can correct for inhomogeneities without prior knowledge of them [Peterson et al., 1998]. The processes to create the inhomogeneities are explained in more detail below.

Using constant offsets to create inhomogeneities

Using constant offsets has been the most common method employed for inhomogeneity addition in past studies. These offsets may have pre-specified sizes, as in DeGaetano [2006] or Reeves et al. [2007]; or they may have sizes sampled from a particular distribution, as in Menne and Williams JR. [2009]. The Normal(0,1) distribution is a common distribution to use in cases of step-changes as, as already stated, Menne and Williams JR. [2005] showed that, certainly in the US, inhomogeneity magnitudes tend to be distributed Normally once they have been standardised; although there is evidence of some positive skew in their figure evidencing this. One inhomogeneity that is prevalent in US records and known to exhibit positive skew is the inhomogeneity caused by changing the time that temperature observations were made [Menne et al., 2009]. This is not an inhomogeneity that is focused on here as it is already relatively well understood and has received more focus in the past than some of the inhomogeneities chosen as focuses for this thesis.

It is known that detection skill for the smallest inhomogeneities, for example from the cen-

tre of a N(0,1) distribution, is extremely low; this artefact is known as the missing middle. In the current study inhomogeneities with sizes in this missing middle were generated to keep a realistic distribution of sizes and section 2.3 of chapter seven includes an analysis of the performance of algorithms in this area; this analysis does bear in mind that these inhomogeneities are known to be difficult to detect.

As constant offsets are the conventional method of inhomogeneity addition this means that more studies have already analysed change points added in this manner, therefore, they were not the primary focus of this study. Constant offset inhomogeneities were also not the primary focus of this study as it is known that, in reality, most inhomogeneities are not constant. Therefore, only 30% of inhomogeneities were added as constant offsets. That is, if the random number generated from the uniform distribution at step four of the inhomogeneity creation and addition process was less than or equal to 0.3 the inhomogeneity was added using a constant offset, otherwise it was added by perturbing the explanatory variables.

For the constant offset method of inhomogeneity addition shelter changes or station relocations could essentially be created in the same way as each other, as both are implemented as sudden changes and not trend changes. This combined type of inhomogeneity was implemented if the random number generated at step three of the inhomogeneity addition process was less than or equal to 0.67. The size of the inhomogeneity was chosen by sampling from the set $\{1.5, 1.25, 1, .75, .5, .25, -.25, -.5, -.75, -1, -1.25, -1.5\}$ °C, this value was then added on to the clean series at the location decided in step 2. All values are equally possible, meaning that positive and negative steps are equally likely here. Allowing both positive and negative steps is justifiable as various studies have shown that neither station relocations nor shelter changes bias all temperature series in the same manner [Hubbard and Lin, 2006; Xu et al., 2013]. Having a range of inhomogeneity sizes allows the assessment of algorithm performance in response to different sized perturbations; from those in the missing middle to those that would be expected to be more easily detectable. The sizes given here changed slightly due to rounding and stacking of inhomogeneity structures in the released data, but the process of evaluating sizes will be explained in more detail in section 4 of this chapter.

Figure 5.3a shows an example difference series between the clean and released data for a station from scenario one in Wyoming where a constant offset shelter change/ station relocation inhomogeneity has been added at the time point 14067 as indicated by the vertical red line. Figure 5.3b shows the inhomogeneity in the released series itself.

Urbanisation inhomogeneities were implemented if the value generated at step three of the inhomogeneity addition process was greater than 0.67. Urbanisation inhomogeneity lengths were drawn at random from a normal distribution with a mean of 15 years and a standard deviation of three years. Although this is a relatively short time period it is long enough to investigate the impacts of non-climatic trends on temperature series without dominating the whole series. If the length selected was less than or equal to 15 years then the trend could reach .1, .15 or .2°C over the period which it acted. That is, a constant gradient slope was added over the period of the urbanisation so that at its end

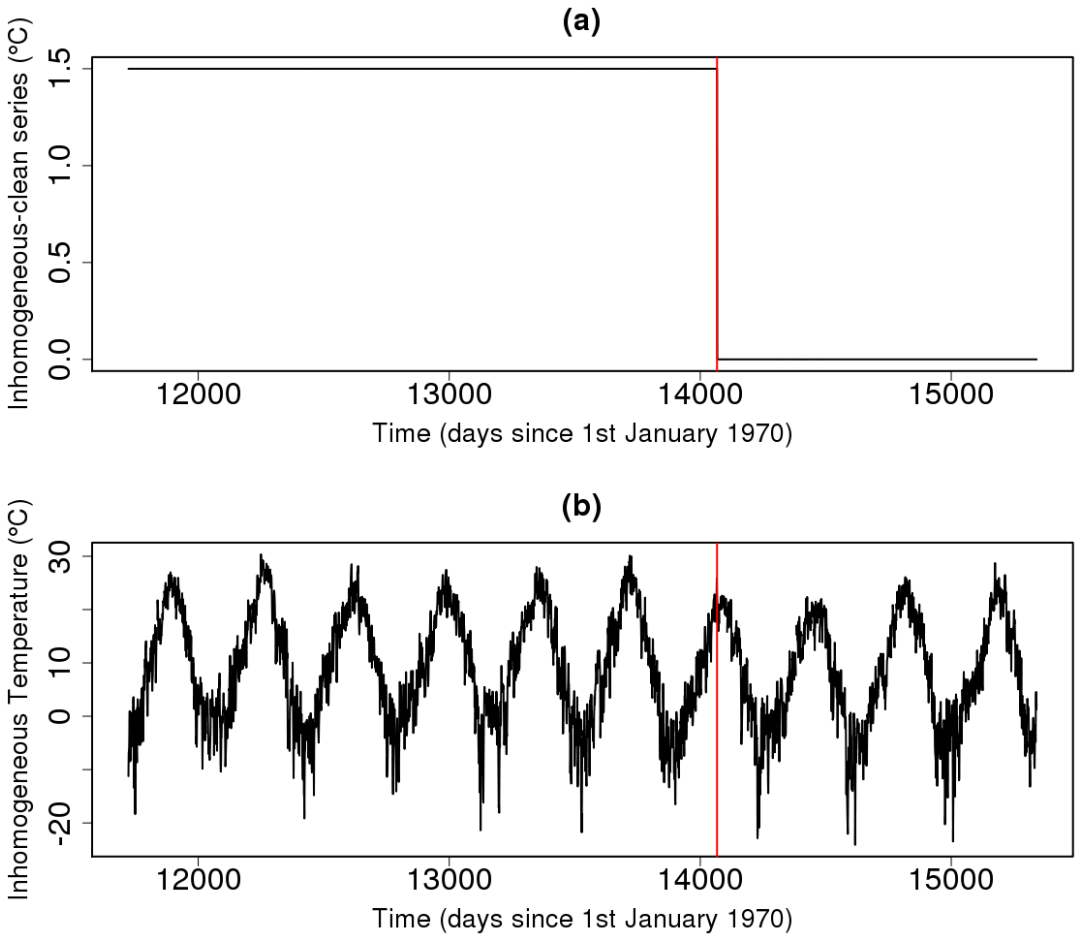

Figure 5.3. (a) Example difference (released minus clean) series, from station 58 in Wyoming scenario one to illustrate the effect of a constant offset shelter change or station relocation inhomogeneity. (b) The equivalent part of the released inhomogeneous series.

the series was .1, .15 or .2°C higher than it was before. If the length of the inhomogeneity was greater than 15 years then the range of values from which the trend was drawn was $\{.15, .2, .25, .5, 1, 1.5\}$ °C.

Figure 5.4a shows an example series where a constant offset urbanisation inhomogeneity has been added. Because the data have been rounded to one decimal place to mimic GHCND data this does not look like a trend inhomogeneity; this is an issue with the real world data too. Here, the constant switching between two different values of the difference series shows that there is a gradual change taking place, this is further evidenced in 5.4b, which is the unrounded difference series. The appearance of multiple values at a single point in time in 5.4a is just an artefact of having many time points close together, there are no real multiple observations for a single time point. Another artefact to highlight from this figure, that is true of all constant gradient urbanisation inhomogeneities is that they acted by reducing temperatures prior to their end point by ever increasing amounts; this means that they are positive trends, but are implemented by reducing past temperatures instead of increasing future ones relative to the clean baseline, because they are applied in reverse.

Perturbing the GAM's explanatory variables to create inhomogeneities

It is known that inhomogeneity effect sizes may be dependent on other climatic variables

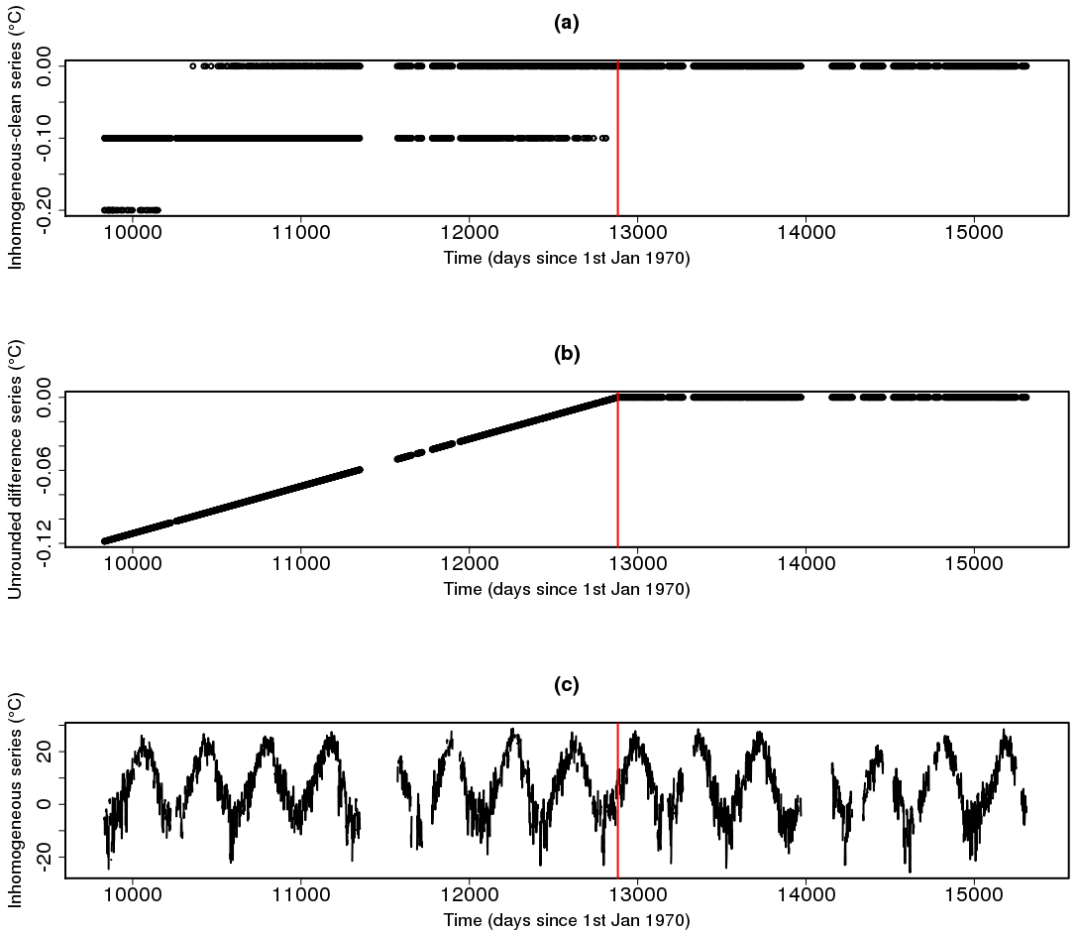

Figure 5.4. (a) Example difference (released minus clean) series, from station 73 in Wyoming scenario one to illustrate the effect of a constant gradient urbanisation inhomogeneity. (b) The difference series if rounding of values to one decimal place had not occurred. (c) The equivalent part of the released inhomogeneous series.

and that they can vary seasonally [Auchmann and Bronnimann, 2012]. In the current work the temperature series were themselves created using a GAM that drew information from these other climatic variables, which allowed the addition of inhomogeneities using a perturbation of these inputs. Creating inhomogeneities in this manner made them seasonally varying and varying according to background meteorological features. This is a new approach to the inhomogeneity addition problem, which is believed to be more realistic, and therefore 70% of inhomogeneities were added in this manner so that a range could be explored.

If step three of the inhomogeneity addition process returned a value less than 0.34 then the inhomogeneity to be mimicked was a shelter change. Numerous shelters have been used to house thermometers over the period of recorded temperature series, see Parker [1994] for an excellent overview of these. The quality of these shelters in protecting the thermometer from radiation whilst allowing free air circulation has varied considerably. For example, Trewin [2010] notes that many pre-Stevenson screens, including the widely used Glaisher stand, were over exposed to radiation and Parker [1994] comments that early UK screens may have restricted the ventilation inside a screen with other equipment when earlier designs of Stevenson screen were in use. Therefore, to mimic shelter change inhomogeneities the explanatory variables of solar radiation, eastward wind and

northward wind were all perturbed.

All three of these explanatory variables were perturbed in the same direction so that a decrease in levels of sun (to mimic better radiation shielding), was accompanied by a decrease in wind (to mimic reduced ventilation) and vice versa; that is, less exposure to one element is coincident with less exposure to another. The literature shows that altering both variables in such a manner may in fact reduce inhomogeneity effects in reality. This is because an increase in radiation can bias the thermometer readings because too much heat is trapped inside the screen, but this effect is reduced with increasing ventilation [Harrison, 2010; Parker, 1994], which is why some screens now artificially ventilate. These inhomogeneities were seasonally varying because levels of sun are naturally seasonally varying and eastward wind has been encouraged to be so by including it as a smooth surface with day of the year in the GAM. Northward wind is not seasonally varying in the model formulation for the reasons given in section 4.1.2, namely that entering more than one smooth surface in to the GAM causes the two surfaces to adversely affect each other, but it does still naturally vary on a day to day basis.

The amount by which the explanatory variables were perturbed was selected randomly from the set $\{0.85,0.9,0.95,1.05,1.1,1.15\}$ and the variables in question were then multiplied by this factor. The inhomogeneity effect sizes produced using this method were reasonable, as is discussed in section four of this chapter, therefore, no other perturbation values were considered.

Figure 5.5 shows an example of a shelter change inhomogeneity created in this manner. In this case the solar radiation, eastward and northward wind were all higher in the past than the present, implemented by the multiplication of these elements by 1.15 before day 11722, which is the date of the change. An explanation of how these perturbations are imposed can be given by considering the model used. The mean of the GAM can be written as

$$\mu_{it} = \beta_0 + \beta_1 Altitude_{it} + \beta_2 Tempforecast_{it} + f_1(Dyear_{it}, UW_{it}) + f_2(Time_{it}) + f_3(Lat_{it}) + f_4(Long_{it}) + f_5(Sun_{it}) + f_6(SOI_{it}) + f_7(VW_{it}) + f_8(Precip_{it}) + f_9(PWC_{it}) + f_{10}(Coast_{it}) + f_{11}(SLP_{it}).$$

Given the same input variables the predictions from this model would always be the same, to measurement precision. However, the predictions for all $t < 11722$ are now made from the model with mean

$$\mu_{it} = \beta_0 + \beta_1 Altitude_{it} + \beta_2 Tempforecast_{it} + f_1(Dyear_{it}, 1.15 * UW_{it}) + f_2(Time_{it}) + f_3(Lat_{it}) + f_4(Long_{it}) + f_5(1.15 * Sun_{it}) + f_6(SOI_{it}) + f_7(1.15 * VW_{it}) + f_8(Precip_{it}) + f_9(PWC_{it}) + f_{10}(Coast_{it}) + f_{11}(SLP_{it})$$

instead, thus changing the values and imposing the inhomogeneity. This inhomogeneity has caused positive shifts in the past in the temperatures with a seasonal cycle, which further investigation reveals creates larger differences in summer than in winter. The artefact of rounding to GHCND precision has again made the inhomogeneity more staccato than it would be if a higher measurement precision was recorded.

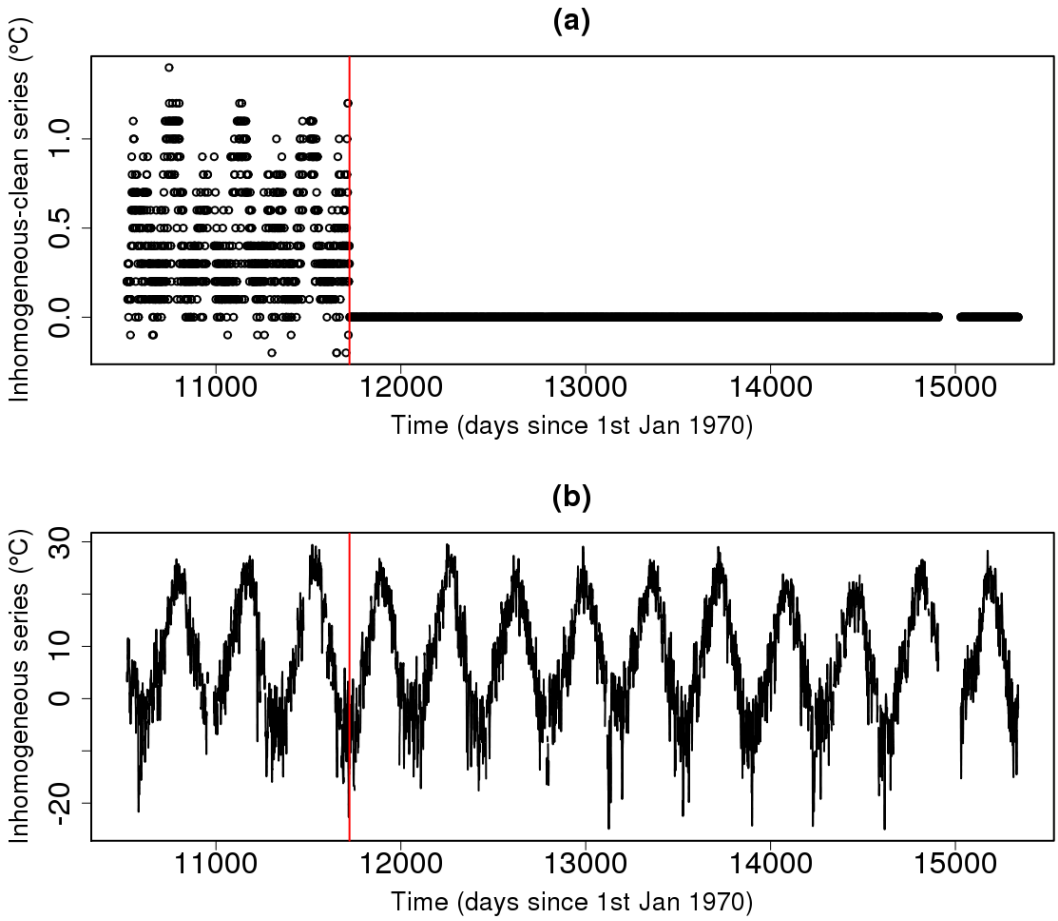

Figure 5.5. (a) Example difference (released minus clean) series, from station 10 in Wyoming scenario one to illustrate the effect of a shelter change inhomogeneity caused by the perturbation of explanatory variables. (b) The equivalent part of the released inhomogeneous series.

If step three generates a number that is greater than or equal to 0.34, but less than or equal to 0.67 then this signifies a station relocation. With the method of explanatory variable perturbation, creating such an inhomogeneity is a relatively straightforward step, all the explanatory variables just need to be taken from a new station location. To simplify the process further each station was assigned a single station that it could be relocated to. The latitude and longitude of the station to have the inhomogeneity were perturbed by an amount that created a station displacement greater than 500m, but less than 5km. All explanatory variables were then acquired for this new location. Station relocations are typically not over very large distances, which is why the maximum displacement allowed was 5km. For the majority of explanatory variables getting the values at the new location meant interpolating the smooth surfaces produced in section 2.2 of chapter three to new locations. Elevations were not interpolated, instead they were obtained from the US geological survey National Elevation Dataset [Gsech et al., 2002; Gsech, 2007]. These elevations were sometimes noticeably different from the elevations of the original station, however, this was not cause for concern, as some station moves in complex terrain will result in large elevation changes and large elevation changes can provide more scope for a noticeable temperature change than just geographical location changes [Trewin, 2010; Xu et al., 2013].

Figure 5.6 shows an example station relocation inhomogeneity that was produced by the

perturbation of explanatory variables. In this case the station moved 2.37km and had an elevation change of 30.8m. The inhomogeneity does have some seasonal variation (largely masked by the rounding process), but it can be seen that its effect is generally one of cooling in the past, which fits with the fact that the older station location had the higher elevation.

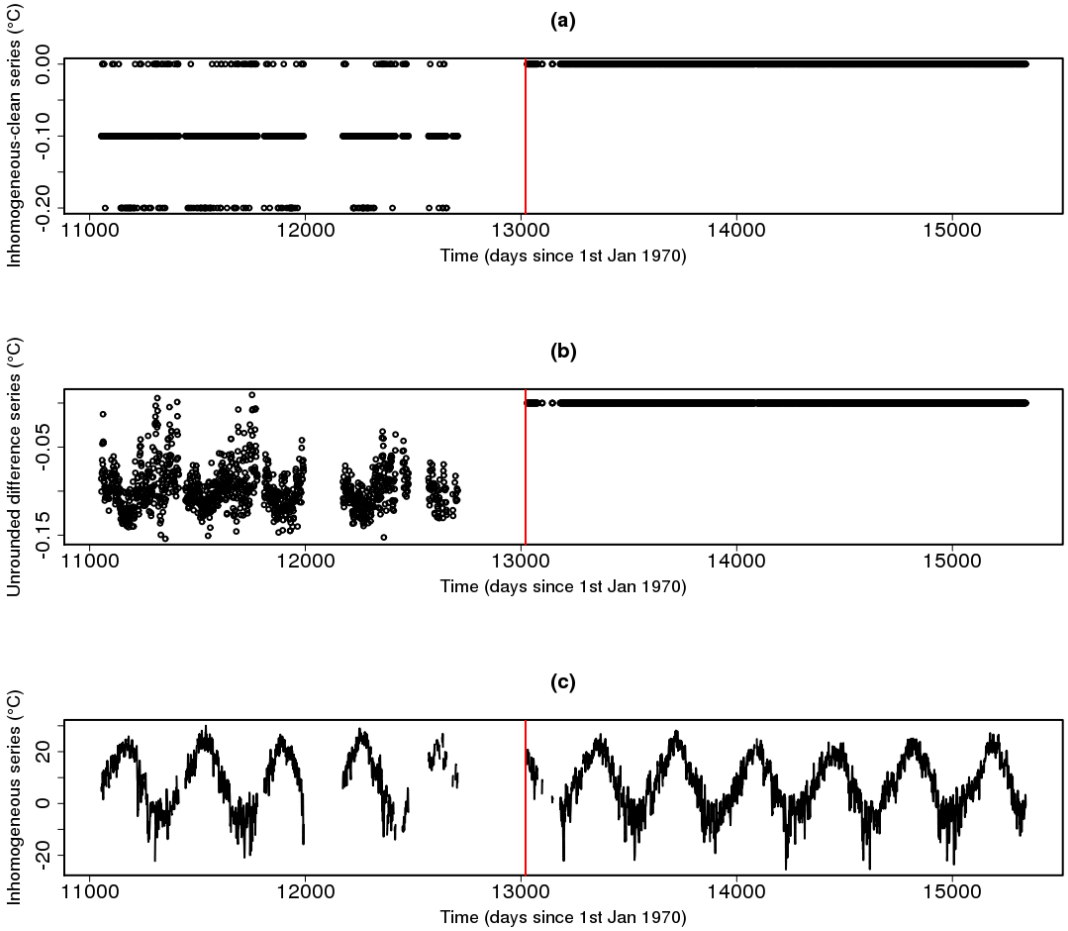

Figure 5.6. (a) Example difference (released minus clean) series, from station 20 in Wyoming scenario one to illustrate the effect of a station relocation inhomogeneity caused by the perturbation of explanatory variables. (b) The difference series if rounding values to one decimal place had not occurred. (c) The equivalent part of the released inhomogeneous series.

If more than one station relocation occurred in a single series and all were to be mimicked by explanatory variable perturbations, then the station alternated between the original location and the single relocation option. If no other inhomogeneities occurred between these points then this would result in a reversal of the original inhomogeneity effect, thus creating a platform inhomogeneity [Domonkos, 2008b]. If more computer power was available then allowing multiple relocation options would be a straightforward extension to the existing code. However, this extension was deemed unnecessary for this project, especially as platform inhomogeneities are a true artefact of temperature series, though a reversal of a relocation would not normally be the cause [Domonkos, 2008b], though it can be, as was shown in figure 2.1!

The final option for the addition of an inhomogeneity was an urbanisation caused by the perturbation of explanatory variables. Increasing urbanisation can lead urban areas to exhibit the urban heat island effect; this effect manifests itself with increased tempera-

tures. It is caused by the presence of buildings leading to a greater retention of heat during the day and a reduced loss of heat at night. The effect is amplified with clearer skies and lower wind speeds [Parker, 2010]. For these reasons, the explanatory variable urbanisation effect was created by gradually increasing levels of sun and gradually decreasing levels of eastward and northward wind from the start point to the end point of the inhomogeneity. The maximum increase or decrease possible was 2.5% if the urbanisation lasted less than 15 years and 2.5%, 5% or 7.5% if the urbanisation was longer than 15 years. The advantage of using explanatory variables here is that the urban heat island's strength in response to cloudy or clear skies and high or low wind speeds will naturally be mimicked. Some experimentation was also done into varying precipitable water content as a proxy for humidity, but this was found to confound the effects of the other explanatory variable perturbations and therefore this line of investigation was not carried any further.

Figure 5.7 illustrates an example urbanisation inhomogeneity. This inhomogeneity is still in progress at the end of the series, which can be seen to be warmer in general than when the inhomogeneity began. This figure also illustrates, though not very noticeably, a slight error that arose in the inhomogeneity addition process that was not discovered until after the data release. This error is that all urbanisation inhomogeneities added by perturbing explanatory variables start with a step change, which will therefore make them easier to find than would otherwise be the case. Although this situation is not ideal, it is not unrealistic. When analysing HCN difference series Menne and Williams JR. [2009] found 40% exhibited a step change accompanied by a trend change, although some of these trend changes may have in fact been small step changes.

5.3.3. Inserting the inhomogeneities

The previous subsection has expanded upon the different methods of inhomogeneity addition. Here these methods will be explained in relation to the final predictions from the GAM used for the released data.

It has already been stated that inhomogeneities propagate backwards in time, so the inhomogeneities that come last chronologically will be implemented first, affecting all time points before their date of addition. Other inhomogeneities then continue to be added in reverse order until the final inhomogeneity (that closest to the beginning of the series) has been added. Explanatory variable perturbation inhomogeneities are added by altering the values of the climatological variables input to the model itself, as was explained in section 5.3.2. Therefore, all explanatory variable inhomogeneities need to have been added before the predictions from the model are made to ensure the correct data are being perturbed. Because the constant offset inhomogeneities do not affect the explanatory variables they are added at the appropriate times after the predictions have been made. In terms of their effect on the model predictions, constant offset inhomogeneities essentially add a constant value to the β_0 term of μ_{it} within appropriate ranges of t as determined by their location values given by the random number generation of step two

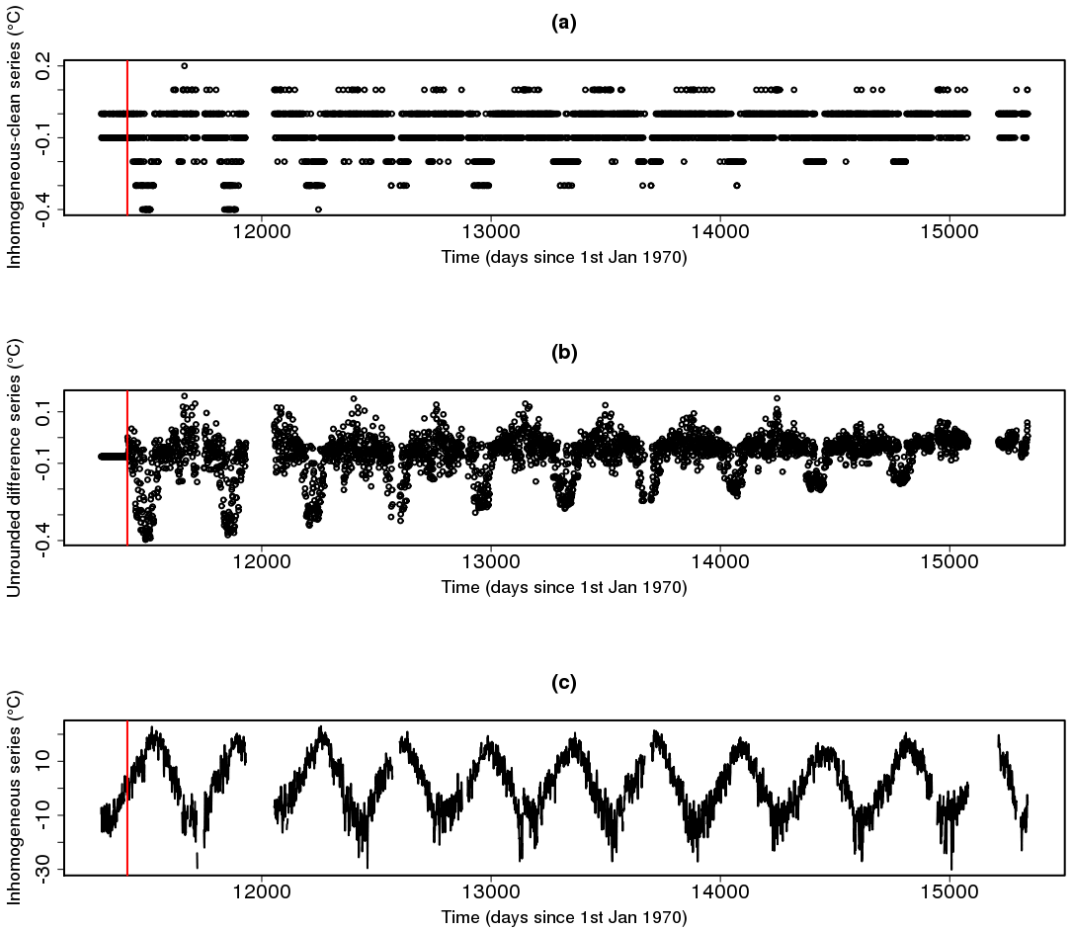

Figure 5.7. (a) Example difference (released minus clean) series, from station 49 in Wyoming scenario one to illustrate the effect of an urbanisation inhomogeneity caused by the perturbation of explanatory variables. (b) The equivalent difference series if rounding to one decimal place had not occurred. (c) The equivalent part of the released inhomogeneous series.

of the inhomogeneity addition process.

The same smoothed noise that was added to the clean predictions is added to the corrupted predictions to ensure that the only changed variability in the data is from intentionally added inhomogeneities and not from the data generation process.

5.4. Evaluation of the scenarios

Once the scenarios were created an investigation was carried out into their properties and, where possible, these properties were compared to the real world so that their fidelity to reality could be assessed. This evaluation process benefited after the release of the data from comments of homogenisers, especially Dr Peter Domonkos.

An initial evaluation involved ensuring that no completely unrealistic (greater than observed extremes) values had occurred. This was done by comparing the ranges of values in each corrupted scenario with the range of temperatures found in the observations. Although extremes did not match perfectly, as was the case in the clean scenarios, es-

pecially in the South West, none were found to be unrealistic and therefore this aspect of the data was deemed appropriate. Figure 5.8 shows the temperature density distributions in each of the scenarios and regions. Black dashed lines represent the observed values, blue dashed lines represent the clean scenarios and red dashed lines represent the corrupted scenarios. The blue lines can hardly be seen in these figures though they are sometimes visible between the red dashes, this illustrates that the distributions of temperature on a regional scale are very similar with and without inhomogeneities. However, it can be seen that the peak of the distribution in figure 5.8b, which represents the South East, matches the observations better after corruption than before.

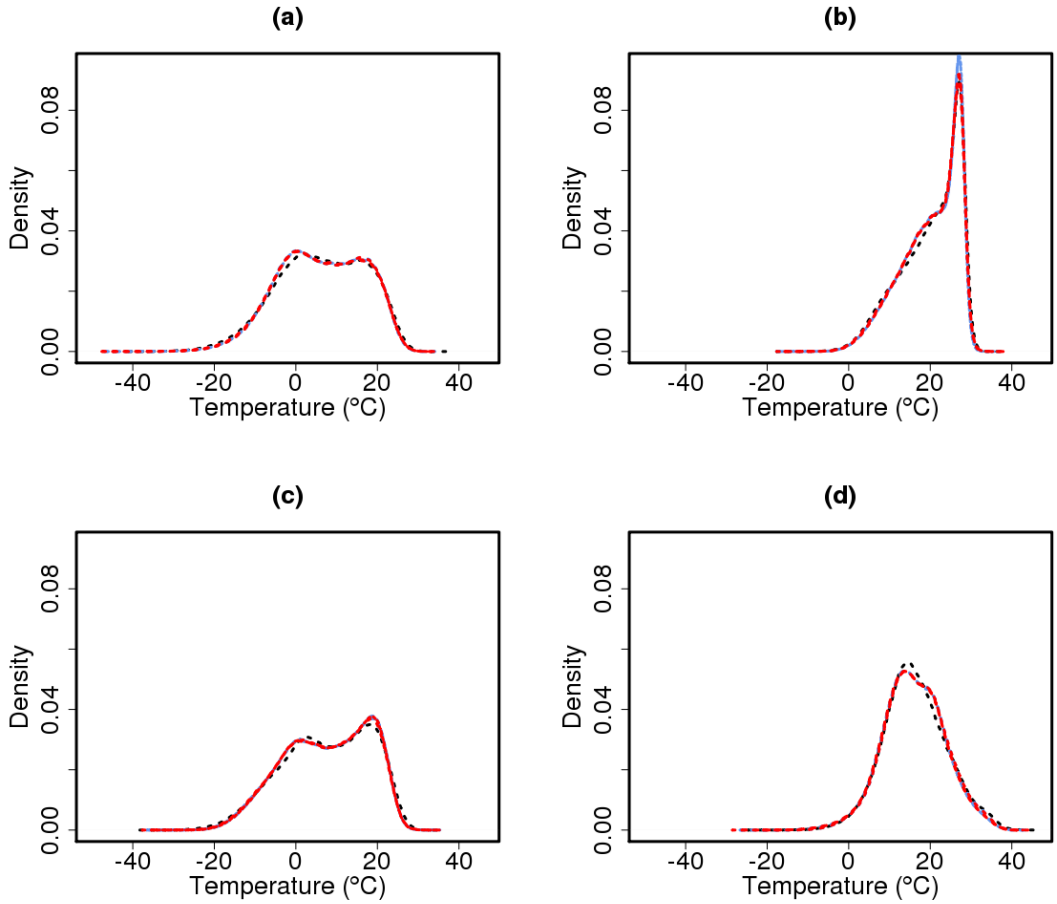

Figure 5.8. Density distributions of temperatures in a) Wyoming, b) the South East, c) the North East and d) the South West. Black dashed lines represent observations, blue dashed lines represent clean scenarios and red dashed lines represent released scenarios.

5.4.1. Inter-station correlations, autocorrelations and standard deviations exhibited in the released data

Section two of chapter four assessed the inter-station correlations and autocorrelations of the created clean data, but these comparisons were against the real world data, which are very unlikely to be free of inhomogeneities. The corrupted data created in this chapter can fairly have their qualities compared to those of reality, as they are designed to match real world inhomogeneity structures.

Figure 5.9 looks at the inter-station correlations in the corrupted data in relation to the clean and observed data. It can be seen that, as expected, corruption lowered the inter-station correlations in all regions, but only by small amounts meaning the inter-station correlations are still higher on average in the created data than in the observations. The inter-station correlations can be seen to not be identical across scenarios within a region, but are broadly similar, with the larger scenarios showing slightly higher inter-station correlations in Wyoming and the North East. The same findings about inter-station correlations were seen when only stations within 75km of each other were investigated, as was done in figure 4.16, this can be seen in figure 5.10. The red points representing corrupted data inter-station correlations against observed inter-station correlations in figure 5.10 can be seen to be lower than their black counterparts which were for clean predictions. However, there is still clearly a bias for inter-station correlations to be higher in the predictions than the observations even when inhomogeneities are present in both data sets. As already stated, this means that the released data are expected to be easier to homogenise than the observed data would be.

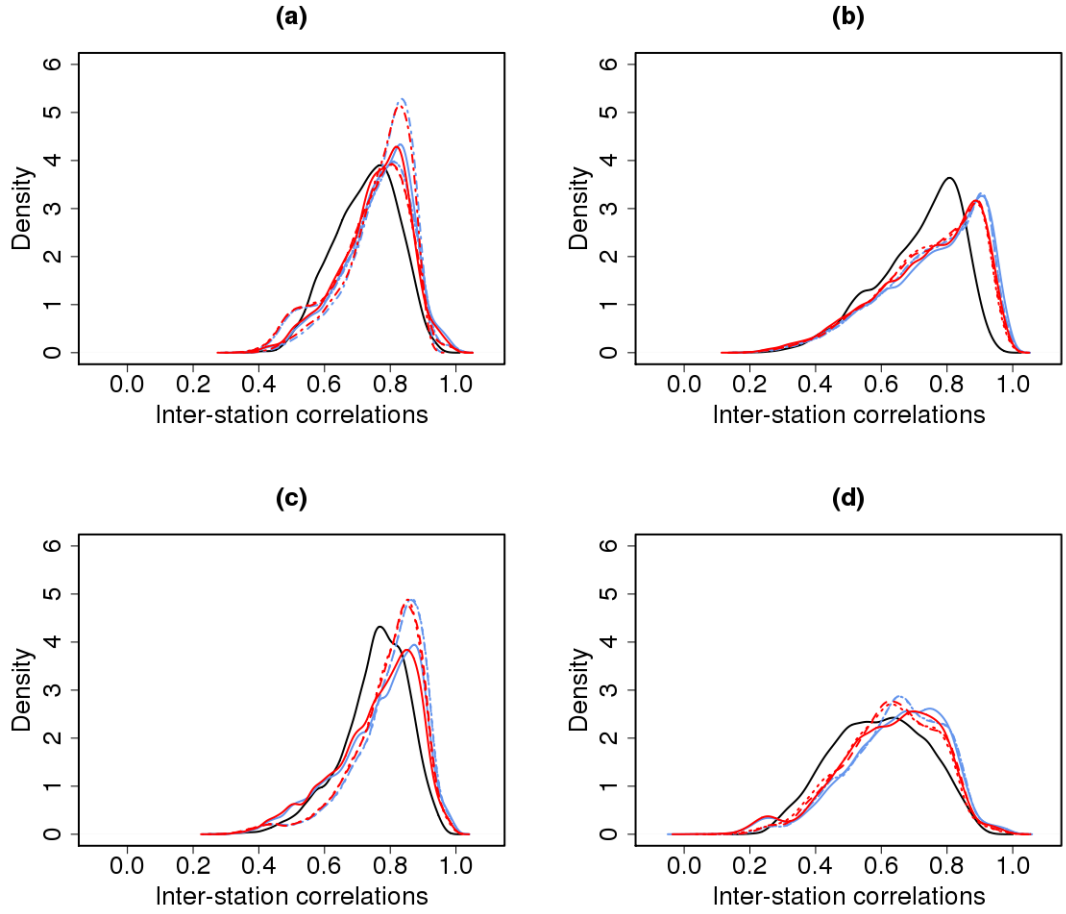

Figure 5.9. Density distributions of the inter-station correlations found in a) Wyoming, b) the South East, c) the North East and d) the South West. Black lines represent observations, blue lines represent clean scenarios and red lines represent released scenarios. Solid lines represent scenario one, dashed lines represent scenario two, dotted lines represent scenario three and the dot-dashed lines in plot (a) are for scenario four.

If regional averages of the corrupted stations are created, and the autocorrelations plotted for these regional averages, then there appears to be no difference in autocorrelation between the clean and corrupted data. Small scale experimentation showed that this

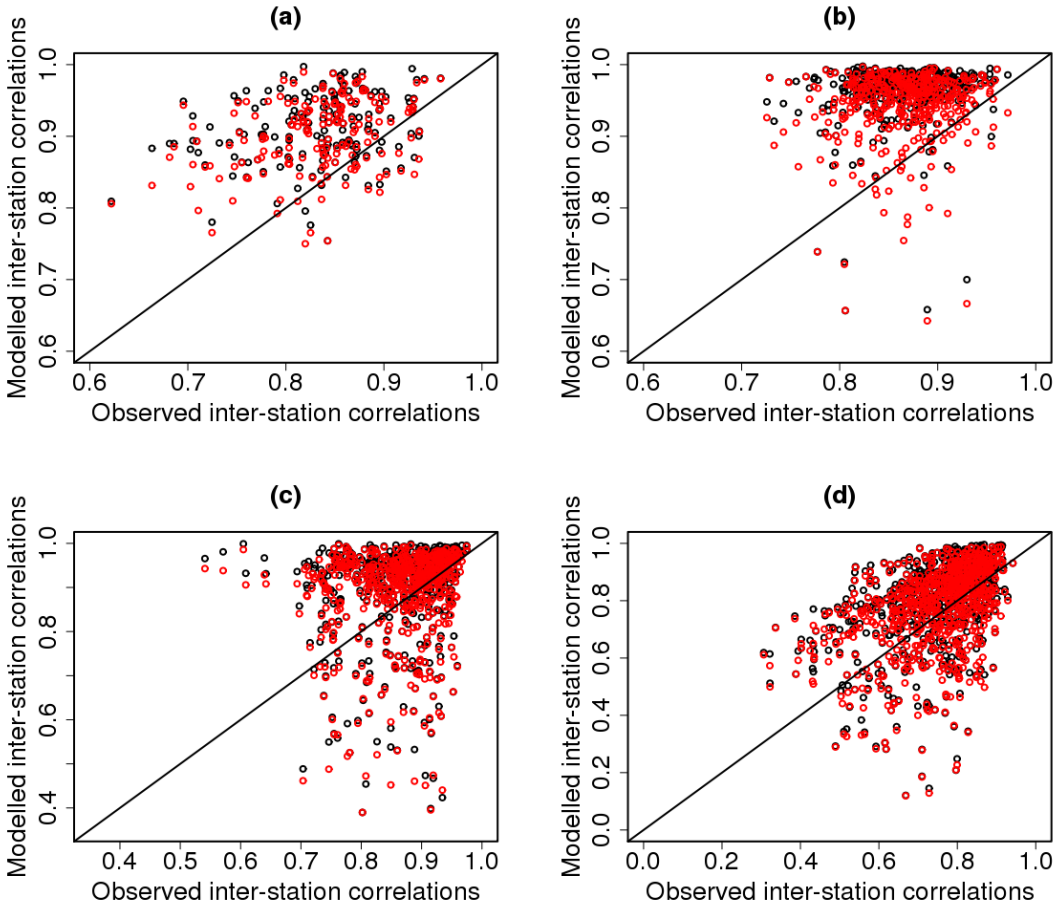

Figure 5.10. Scatter plots of the observed versus predicted inter-station correlations found in a) Wyoming, b) the South East, c) the North East and d) the South West between stations that are less than 75km apart. Black dots are inter-station correlations before inhomogeneities were added and red dots are inter-station correlations after inhomogeneities have been added. It is evident that the addition of inhomogeneities does decrease inter-station correlations, but not by very large amounts. The inter-station correlations displayed here are only for scenario one, but similar findings were obtained when other scenarios were also investigated.

lack of difference was largely the case on a station by station level too, but figure 5.11 illustrates it is not always the case. Looking on a lag-by-lag basis the autocorrelations can be seen to be fractionally higher in the corrupted data; this is more evident at later lags. Figure 5.11 shows the autocorrelations for the observations (black), the clean data (blue) and the released data (red) for station one in scenario one in each of the four regions. Figure 5.11d, for the South West appears to show no change in autocorrelations and this is not surprising given that this station sees just one step change in its record. Wyoming and the North East both have step and trend changes in station one, with differing effects on the autocorrelations, while the South East sees only step changes.

Also of interest are the autocorrelations in the deseasonalised difference series between most highly correlated neighbours. Figure 4.19 showed that the common assumption that these series are white noise is invalid and figure 4.20 showed that the clean data did not reproduce these autocorrelations well. Figure 5.12 shows that, where the clean series had a tendency to under-estimate autocorrelations in deseasonalised difference series, the corrupted data over-estimate these autocorrelations on average. The autocorrelations displayed here are averages of the autocorrelations in each deseasonalised difference

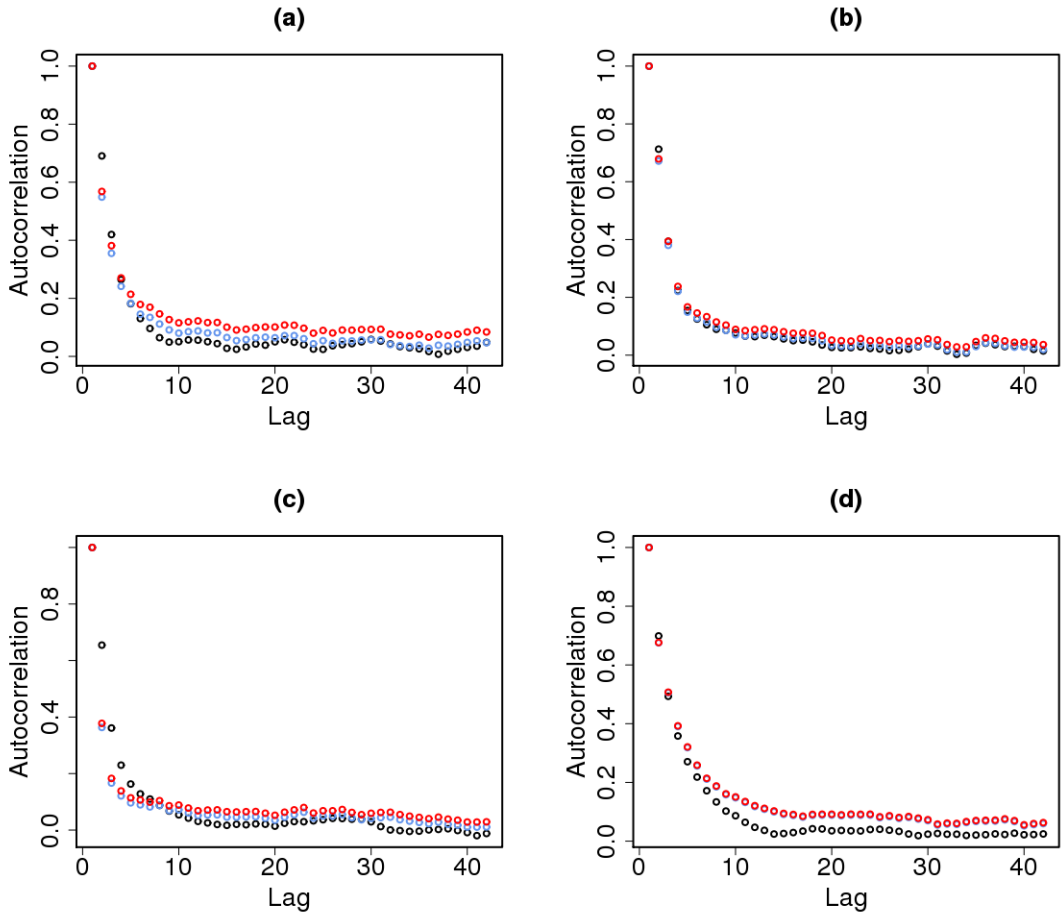

Figure 5.11. Autocorrelation plots for station one of scenario one in a) Wyoming, b) the South East, c) the North East and d) the South West. Black points represent observed data, blue points represent clean data and red points represent released data.

series at each lag on a scenario by scenario basis. Not all stations will behave in the same manner. Interestingly the scenario that doesn't allow trend inhomogeneities (scenario three, plotted in orange) displays more persistent autocorrelations in the deseasonalised difference series. This finding was a surprise as series with trends might be expected to exhibit greater autocorrelation than those without. However, a possible explanation as to why there are lower autocorrelations exhibited in trend scenarios is that the effects of trends somewhat cancel out and they are also more gradual whereas a step change might be expected to force the difference series to have a more persistent sign thus increasing the autocorrelation. The autocorrelations in scenario four in figure 5.12a can be seen to also be increased from the clean state, but not as far as reality after the point at which the weighted moving average stops acting (lag eight). Therefore, scenario four can be considered as closest to reality whilst still being a little too easy because of being closer to the algorithmic assumption of white noise difference series (because of the lower autocorrelations) than the observations are.

Finally, just as in the clean scenarios, the standard deviations in the deseasonalised difference series were examined. To show all the standard deviations for all scenarios would result in an illegible plot, therefore, figure 5.13 shows only the standard deviations for the real data, the clean version of scenario one and the corrupted version of scenario

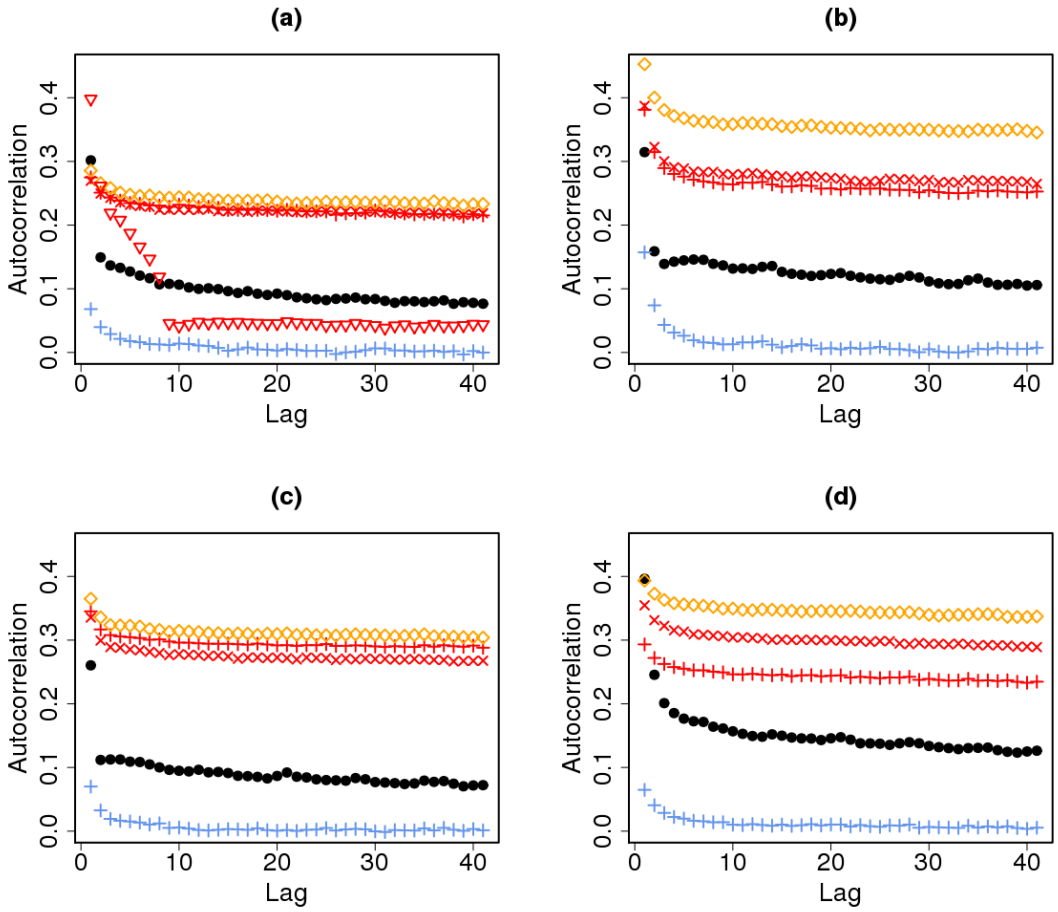

Figure 5.12. Averages of the autocorrelations in each deseasonalised difference series at each lag for a) Wyoming, b) the South East, c) the North East and d) the South West. Black points represent observations, blue addition signs represent clean data, red points represent released data in trend scenarios (addition signs are scenario 1; multiplication signs are scenario 2 and triangles are scenario 4) and orange points represent released data in scenario 3.

one (and scenario four in Wyoming). The series have been differenced with respect to their most highly correlated neighbour in their scenario, therefore the series differenced against are different in reality, scenario one and scenario four. The clean series are differenced with respect to the most highly correlated neighbour in the released data and not in the clean data, in order to make these comparisons direct, but in more cases than not the highest correlated neighbour is the same in both situations. It can be seen that, as expected, corruption increases the variability of the series, with red points being higher than blue points. Scenario four is still the best match to reality in terms of standard deviations, though the standard deviations are still relatively constant. There are no completely consistent patterns in the ordering of the standard deviations of the scenarios, but very generally speaking scenario two is prone to having lower standard deviations than both scenarios one and three. In all cases though the pattern in standard deviations across scenarios is not uniform and therefore broad conclusions about the relationship between algorithm performance and scenario standard deviation will not be reported on a scenario by scenario level.

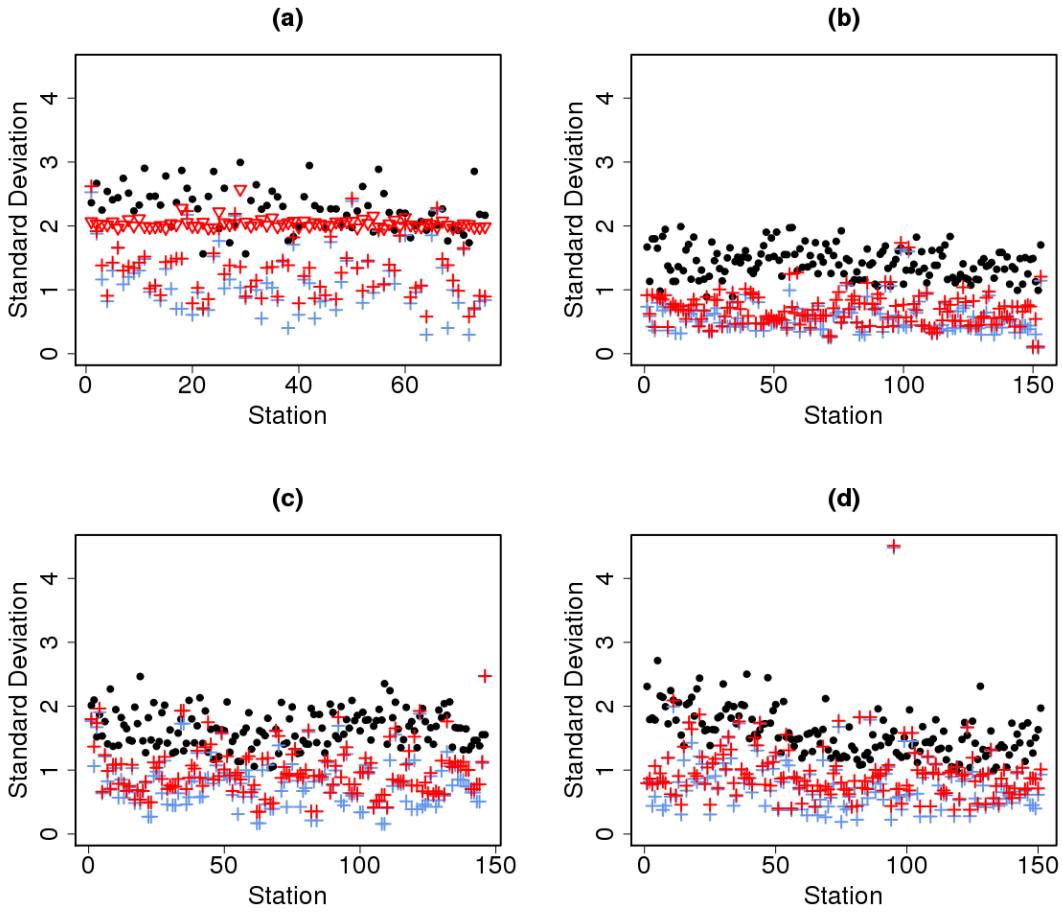

Figure 5.13. Standard deviations in deseasonalised difference series for a) Wyoming, b) the South East, c) the North East and d) the South West. Black points represent observations, blue addition signs represent clean data, red addition signs represent scenario one released data (and red triangles are scenario 4).

5.4.2. Inhomogeneity size and frequency

Inhomogeneity sizes can be defined in two different ways, which will here be defined as an inhomogeneity's *cumulative size* and its *relative size*. Its cumulative size is its mean size over the period which it acts compared to the clean series, that is, if d is the difference series between clean and released data, then the cumulative size of an inhomogeneity would be the mean size of d for the period of the inhomogeneity. This means that the cumulative size of an inhomogeneity takes into account all the other inhomogeneities that may be acting on the same time period. Thus, if two consecutive inhomogeneities have opposite effects and return the data to homogeneity, the cumulative size of the second inhomogeneity will be zero.

The relative size of an inhomogeneity is its size relative to the next homogeneous sub period. This is calculated by working out the mean value of d for two consecutive sub-periods and then working out their difference to get the relative inhomogeneity size.

The outcome of these definitions is that an inhomogeneity can have a cumulative size of zero, without having a relative size of zero. When this happens there is a platform inhomogeneity, as defined in 5.3.3, because the effect of one inhomogeneity has corrected

for the effect of another. In reality this might happen owing to a change which causes an inhomogeneity that is then discovered and rectified [Domonkos, 2011]. Table 5.1 shows the number of platform inhomogeneities in each region as well as other regional characteristics. The proportion of platform inhomogeneities is higher in scenario three than in the other scenarios. The cause of this is likely due to the lack of trend inhomogeneities in scenario three. This means in scenario three a step change of one sign and magnitude has an increased chance of being followed by another step change of opposite sign and the same magnitude, thus creating a platform inhomogeneity; whereas in the other scenarios a trend inhomogeneity could start between them causing enough of a disturbance to stop a platform forming.

Personal communication with, and investigation by, Peter Domonkos revealed that the proportion of platform inhomogeneities is too low in the created data and should therefore be increased, or considered more directly in a future study. It is emphasised in Domonkos [2013] that the benefit of including these inhomogeneities is not solely to test whether they can be detected, but because their presence will also impact on the detection of more substantial inhomogeneities. It should be noted here that in other studies, for example Venema et al. [2012], platform inhomogeneities refer simply to change points with opposite signs, but not necessarily the same magnitudes, the reported proportion of platform inhomogeneities is therefore higher in such studies than in the present work.

If an inhomogeneity has a relative size of zero, regardless of its cumulative size, then this means that the change point separating two homogeneous sub-periods is indistinguishable and is therefore termed *unidentifiable*. Homogenisers were not expected to be able to find unidentifiable inhomogeneities and therefore the location of these inhomogeneities was noted in the data creation, but they were not searched for when assessing algorithm performance. Table 5.1 shows the number of unidentifiable inhomogeneities in each region. The proportions of added inhomogeneities that were unidentifiable is greater in the South East and the South West than in Wyoming or the North East which have relatively similar proportions of unidentifiable inhomogeneities in equivalent scenarios.

In all regions there were never any unidentifiable inhomogeneities caused by constant offset shelter changes or station relocations. This is because, as stated in 5.3.2, the smallest value these inhomogeneities could take was 0.25°C (though their relative size could become a little smaller when interacting with an inhomogeneity with a seasonal cycle). In Wyoming, the North East and the South West scenarios that allowed urbanisation, urbanisation was the most common cause of an unidentifiable inhomogeneity. Given that urbanisation is a slow trend with a smaller magnitude in general than other types of inhomogeneity this was not surprising. Urbanisation was also just about the most common cause of unidentifiable inhomogeneities in scenario one for the South East. In other South East scenarios and in scenario three for the North East and Wyoming station relocations were the dominant cause of unidentifiable inhomogeneities. In the South East this is not surprising as it is the least topographically and climatologically diverse of the regions. In Wyoming and the North East although the region varies more it is still believable that a small change of all variables (a station relocation) will have a lesser impact

than a larger change of three variables (a shelter change). In the South West which is a more diverse region station relocations were the least common cause of unidentifiable inhomogeneities.

It was the relative size that was used to group inhomogeneity magnitudes into categories. These categories were small inhomogeneities, which were less than or equal to 0.2°C in magnitude, medium inhomogeneities which were between 0.2 and 1°C in magnitude and large inhomogeneities, which were greater than 1°C in magnitude. These groupings were decided so that the small inhomogeneities could be thought of as those that traditionally fall into the category of the 'missing middle', these are the inhomogeneities that participants are unlikely to find owing to them only being a little larger than measurement precision. The medium inhomogeneities still incorporate some that are likely to be missed, as other studies have considered inhomogeneities smaller than 0.5 as unidentifiable [Stepanek, 2004; Menne and Williams JR., 2009], but they are still sizeable enough to bias conclusions drawn from analyses run on them. The large inhomogeneities are those that should definitely be corrected for by the algorithms. When assessing algorithm performance the proportion of inhomogeneities of different sizes that were detected was noted as well as the similarity of the returned to the clean series. Therefore, algorithms could be judged accordingly if they had a good detection skill in general, but were unable to pick up the smallest inhomogeneities. Further details of the assessment criteria for the algorithms are given in the following chapter.

Table 5.1. A table to summarise the inhomogeneity characteristics in each of the created scenarios. Where average inhomogeneity sizes are given these refer to relative inhomogeneities and are calculated only from the identifiable inhomogeneities. Size means that sign has been taken into account, whereas magnitude means the absolute value has been taken. Thus the mean inhomogeneity magnitude is larger than the mean inhomogeneity size because there can be counteracting positive and negative effects for inhomogeneity size, but not magnitude. Also, numbers of inhomogeneities 'found' by the PHA here refer only to inhomogeneities with a suggested adjustment larger than the suggested uncertainty. The numbers in brackets in this column are the number of inhomogeneities found that are within a month of a true inserted inhomogeneity, whereas the numbers not in brackets are just the total numbers 'found'.

Scenario	No. of unidentifiable inhomo-geneities	No. of platform inho-mogeneities	Mean IH magnitude	Mean IH size	No. of small IHs	No. of medium IHs	No. of large IHs	Total no. of identifiable IHs	No. of IHs found by PHA in released scenario	No. of IHs found by PHA in clean scenario	Average no. of IHs per series)
WYW1	43	9	0.49°C	-0.035°C	104	116	37	257	81 (42)	4	3.33
WYW2	85	18	0.43°C	0.028°C	214	206	69	489	176 (92)	5	2.93
WYW3	29	28	0.49°C	0.070°C	165	219	61	445	181 (96)	16	2.82
WYW4	41	11	0.50°C	-0.024°C	101	119	39	259	88 (34)	3	3.36
SEW1	168	7	0.40°C	0.042°C	244	74	59	377	188 (91)	13	2.37
SEW2	265	12	0.36°C	0.004°C	300	98	68	466	209 (122)	9	2.11
SEW3	231	33	0.44°C	-0.031°C	252	136	55	443	220 (132)	15	2.11
NEW1	95	21	0.39°C	0.024°C	287	115	61	463	188 (100)	11	3.08
NEW2	146	12	0.38°C	0.049°C	355	176	65	596	269 (129)	9	2.82
NEW3	46	34	0.47°C	0.009°C	282	183	76	541	253 (162)	11	2.61
SWW1	156	12	0.42°C	-0.028°C	207	117	51	375	166 (67)	50	2.41
SWW2	223	18	0.44°C	0.031°C	300	162	82	544	229 (126)	28	2.39
SWW3	153	40	0.56°C	-0.009°C	248	178	92	518	251 (155)	31	2.33

Table 5.1 gives the mean inhomogeneity magnitude (the absolute value of the inhomogeneity which must therefore be positive) and the mean inhomogeneity size (the signed value of the inhomogeneity) as well as the number of inhomogeneities in each magnitude category. Note that the mean inhomogeneity magnitude is similar across all scenarios and regions and that the mean inhomogeneity size is generally close to 0°C suggesting that there is no persistent bias in the sign of the added inhomogeneities. Figure 5.14 shows the distribution of inhomogeneity sizes in all the regions combined. Plot (a) displays the inhomogeneity sizes before standardising and plot (b) shows the sizes after this process has taken place. Plot (b) also incorporates an overlaid N(0,1) density line and shows that although the inhomogeneities are not Normally distributed, the distribution they do have is approximately symmetrical like a Normal distribution, just with a higher peak and more larger values at the expense of values between the centre and the tails. The reason for comparing to a normal distribution at all is that, as already stated, Menne and Williams JR. [2005] found this to be a good distribution for representing standardised inhomogeneity sizes in the US. Overall, the plot shows that there are a good range of inhomogeneities and therefore the inhomogeneity creation and addition process was deemed a success.

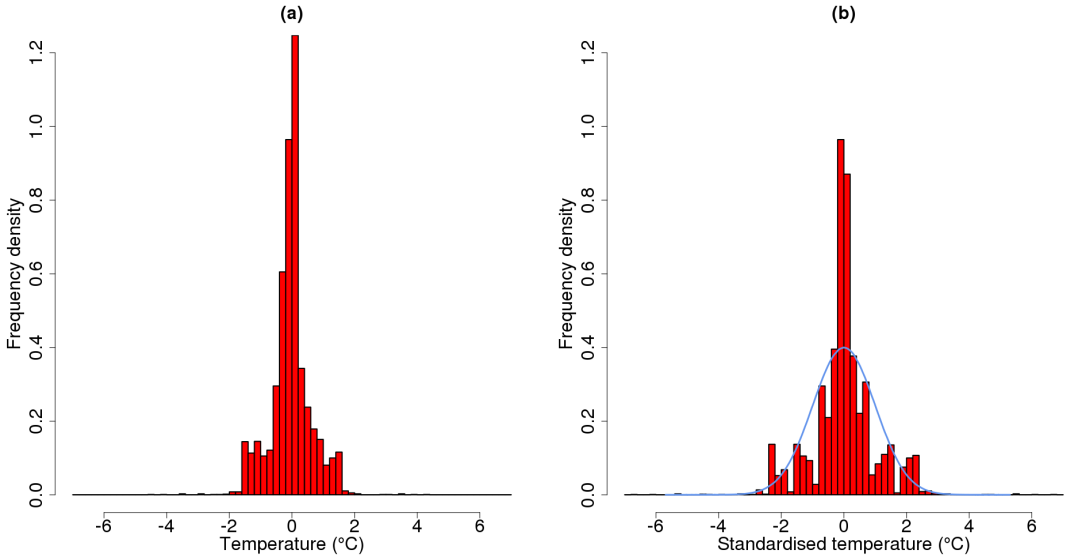

Figure 5.14. A histogram showing the distribution of inhomogeneity sizes before (a) and after (b) they have been standardised by dividing by the standard deviation of all sizes. The blue line overlaid in figure b shows the density of a N(0,1). The pattern in bar heights in figure (b) is believed to be an artefact of adding constant offset inhomogeneities of discrete sizes.

Looking at a breakdown of the sizes by different types of inhomogeneity, the following findings were obtained. As designed, approximately 30% of the inhomogeneities were added by constant offsets and 70% of the inhomogeneities were added by explanatory variable perturbation in each of the regions. However, owing to more of the explanatory variable perturbations creating unidentifiable inhomogeneities than the constant offset perturbations these percentages do change slightly. This change in distribution of inhomogeneity types is most noticeable in the South East where approximately 40% of the final identifiable inhomogeneities are from constant offsets. The majority of all constant offset inhomogeneities are medium in all regions, with this majority ranging from 38% in the North East scenario one to 64% in the South East scenario three. This higher fre-

quency of medium inhomogeneities in all scenarios is logical as the shelter and station change constant offset values were largely in this range. There were some small constant offset inhomogeneities, largely arising from the urbanisation inhomogeneities, though a few from other sources. The remaining 30% or so of constant offset inhomogeneities were classified as large and were solely from station and shelter changes.

Looking at the inhomogeneities added by perturbing explanatory variables there was a much greater tendency towards small inhomogeneities. Small inhomogeneities were the dominant category that inhomogeneities fell into in all regions but Wyoming; around 40% were classed as small in Wyoming and between 70 and 90% were classed as small in other regions. In all regions there were relatively few large inhomogeneities, with none at all in the South East, which also only saw only a small proportion of medium inhomogeneities (between 4 and 12%). This propensity for small inhomogeneities in the South East agrees with the earlier reasoning that explanatory variable changes would be expected to make less difference here owing to it being the most uniform climatological and topological region. In the remaining three regions explanatory variable shelter changes never caused large inhomogeneities, and station relocations caused them in less than 10% of cases in all regions except from scenarios two and three in the South West. These larger inhomogeneities caused by station relocations in the South West would again fit with the greater variability of this region. In the North East and Wyoming there were a small number of large inhomogeneities caused by explanatory variable urbanisations, this was due to these inhomogeneities inadvertently being forced to start with a step change.

When comparing scenarios, the similarities between scenarios one and four in Wyoming should be emphasised. These two scenarios were designed to be similar so that the assessment could evaluate the impact of different underlying data characteristics, namely autocorrelation, as its primary focus. Therefore, as already stated, they had the same inhomogeneity structures added. This means that the distribution of inhomogeneity sizes is very similar, though not identical, with scenario four having a slightly greater tendency towards larger inhomogeneities. In turn this leads to unidentifiable inhomogeneities that are not always the same in the two scenarios, in fact they differ in approximately 10% of cases. These differences lead to marginally different average homogeneous sub period lengths, but by less than a year. Overall the general agreement of characteristics in these two scenarios is clear to see in tables 5.1 and 5.2 and therefore comparing algorithm performance across them was deemed legitimate.

Table 5.1 gives the number of inhomogeneities found by the PHA in the released data. This is the same algorithm that was run on the clean data. As with the clean data, the number of inhomogeneities reported in this table is the number of inhomogeneities found with a shift size greater than their shift uncertainty. It can be seen that PHA always identifies less than half of the inserted inhomogeneities to within a month of their location and also that it 'identifies' many inhomogeneities that were not inserted into the released data. The algorithm performs best in scenario three of each region as was anticipated from this being the simplest scenario. The large number of false alarms from this algorithm were

a little concerning, but were not taken into account when assessing the performance of other algorithms. The reason for not taking them into account was primarily that the PHA is not a perfect algorithm and the aim of this study was to compare the performance of algorithms against the clean benchmark data, not against the performance of the PHA. It should also be noted that the window used here for classifying an inserted change point as a hit and not a false alarm was only one month, which is fair, but still quite strict.

Table 5.2 provides some further information on the characteristics of each region in each of the scenarios. Lengths of homogeneous sub-periods are defined in two ways, lengths in real time and lengths in condensed time. The length in real time refers how many days there were between one inhomogeneity and the next regardless of whether there were data available for these days. The condensed time is the number of days in a homogeneous sub period that the data were present for. There were a few inhomogeneities that were completely unidentifiable because there was no data during the period over which they acted, in these instances homogenisers were not expected to find them and they were removed from the records.

Table 5.2. A table to summarise additional characteristics in each of the created scenarios. HSP stands for homogeneous sub period. RT stands for real time and CT stands for condensed time (i.e. the length of an event after missing data has been removed). The averages given are the means. There is no minimum length between change points by nature of the addition process. It is not expected that all close change points will be found, but, as will be explained in the following chapter, a single correction can be classed as a hit for multiple change points using the windowing approach to validation adopted in this study. References to extremes overshot and missed here are comparing extremes between the observations and the released data, extremes here are not being documented on like for like days, it is simply a record of how many values predicted were more extreme than the most extreme value observed in reality and how many values observed in reality were more extreme than the most extreme value predicted on a scenario by scenario basis.

Scenario	Average HSP length		Longest HSP		Shortest HSP		No. of HSPs less than 1 year		No. of HSPs less than 6 months		No. of values ovsershooting observed extremes	No. of observed extremes missed	No. of non-homogeneous most recent HSPs (due to trend inhomogeneities)	No. of homogeneous stations
	RT	CT	RT	CT	RT	CT	RT	CT	RT	CT				
WYW1	3540	3258	15340	15311	7	7	28	29	20	21	5	13	7 out of 10	2 out of 75
WYW2	3903	3645	15340	15164	40	22	41	44	15	17	27	6	16 out of 17	14 out of 158
WYW3	4108	3753	15340	15318	10	3	42	49	21	22	35	4	NA	11 out of 158
WYW4	3518	3238	15340	15311	7	7	31	32	21	22	6	0	7 out of 10	2 out of 75
SEW1	4548	4232	15340	15337	26	25	36	40	17	22	13	1	14 out of 22	18 out of 153
SEW2	4918	4600	15340	15339	5	5	36	40	17	19	3	4	21 out of 36	31 out of 210
SEW3	4933	4614	15340	15338	32	31	33	37	14	15	3	2	NA	33 out of 210
NEW1	3751	3551	15340	15335	10	10	36	39	18	18	8	2	12 out of 22	4 out of 146
NEW2	4014	3782	15340	15339	2	2	47	52	27	30	0	9	6 out of 18	11 out of 207
NEW3	4245	3999	15340	15289	8	8	51	54	24	25	1	28	NA	15 out of 207
SWW1	4489	4171	15340	15340	2	2	33	35	18	22	0	348	10 out of 17	20 out of 151
SWW2	4523	4203	15340	15333	0	0	28	33	25	31	4	250	13 out of 26	25 out of 222
SWW3	4602	4277	15340	15338	13	13	31	34	10	12	2	87	NA	31 out of 222

A statistic not shown in table 5.2, but calculated during the evaluation process is the median HSP length. In all regions the median HSP length was less than the mean HSP length. This means there is positive skew in the distribution of the HSP lengths as can be seen in figure 5.15 which is for Wyoming scenario one, but is relatively similar to equivalent figures for other scenarios and regions. This positive skew means that shorter HSPs are more common than longer HSPs. This can be explained because to have an average of n HSPs per series, for every series with one long HSP, there must be another series with $n-1$ much shorter HSPs. Unless $n=2$ this means that there will be more shorter HSPs than longer HSPs, as exhibited here. Another way of explaining this is to remember that the inhomogeneity locations were generated using a Poisson distribution and so the length of the homogeneous sub-periods have an exponential distribution, which is positively skewed.

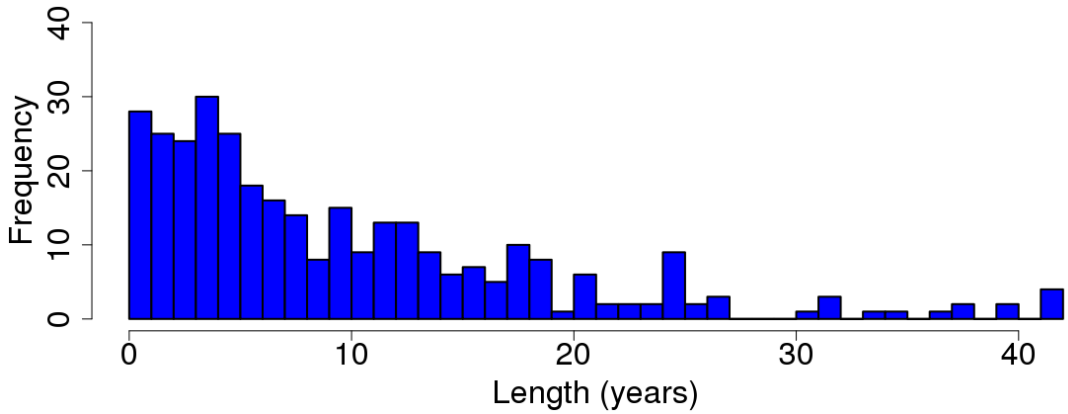

Figure 5.15. A histogram showing the distribution of homogeneous sub-period lengths in real time for Wyoming scenario one.

5.5. Discussion

As stated in section 5.1, no metadata were provided for any of the scenarios created in this thesis. The primary reason for this is that such data are often not available in the real world and therefore homogenisation algorithms cannot rely on them. However, the author believes that it would be worth creating a scenario with metadata in a future iteration of this project to see what effect providing such information does have on algorithm performance. Even if algorithms only use metadata as a checking mechanism, if such data cause a noticeable improvement in performance, then it would be further evidence of the need to digitise old station records so that algorithms can benefit from metadata in real life.

In addition to a metadata scenario another new scenario could investigate including more platform inhomogeneities, to better match real world distributions of these artefacts. Also, if more computing power were available stations could have multiple relocation options instead of only one.

A future iteration of this study could also investigate a greater range of inhomogeneity

sizes in its scenarios. Although the current size distribution is reasonable, constant offset inhomogeneities were here chosen from a discrete set of values. If they were chosen instead from a Normal distribution, as in Venema et al. [2012], then their distribution would likely be more similar to that of the explanatory variable changes, which would make comparisons between the two inhomogeneity types simpler.

Thinking about urbanisation inhomogeneity sizes, a change to a continuous distribution could also be beneficial. The possible trends in this thesis are similar in magnitudes to those used by Menne and Williams JR. [2009]. However, they are larger than the values cited in DeGaetano [2006] or Ren and Zhou [2014] for urbanisation and they therefore should perhaps be reconsidered in a future iteration of this study. It should also be noted that the size a trend could take had a discontinuity at the 15 year mark; the author recommends that should this work be carried out again a trend per year should be specified, as in DeGaetano [2006] instead of the overall size of the trend.

An interesting extension to the present work using the existing output of the PHA would be to analyse the PHA's detection ability when the criteria is simply to get a change point in the right year instead of within a month of the true change point. The reason for such an investigation would be that Venema et al. [2012] found the PHA to have a very low false alarm rate when assessing it at the yearly resolution whereas the present study looking at it at the monthly resolution found it to have a high false alarm rate.

5.6. Summary

This chapter has introduced the different scenarios that were released to the homogenisation community to allow for the testing of temperature homogenisation algorithms on daily data. These scenarios explored different inhomogeneity and station characteristics by considering different regions, different inhomogeneity types and different underlying data autocorrelations.

An overview of the method used to create inhomogeneous data has been given as well as a justification of why the specific inhomogeneities modelled were investigated. The inhomogeneities themselves were created in two different ways; the traditional constant offset approach and the new explanatory variable approach that utilises the GAM's ability to create inhomogeneities by using information from other climatic variables.

The chapter concludes with an overview of the scenarios created and compares them to the clean data that were created in the previous chapter. This shows that corrupting data does alter their characteristics, highlighting once more the importance of having clean data if reliable conclusions are to be drawn from analyses carried out on them. The frequency and size of inhomogeneities is well matched to previous studies' findings in general, with perhaps a slight tendency to over-estimate the number of small and large inhomogeneities, at the expense of those of medium size. This is acceptable however, as small inhomogeneities can bias series, but pose a bigger challenge to homogenisation

algorithms. Therefore, assessing performance of the algorithms in this study in this area is beneficial as not all algorithms are yet in mainstream use. Details of these algorithms are given in chapter seven after chapter six has introduced the methodology for assessing their performance.

6. Framework for Evaluating the Returned Data

Previous chapters have explained how the data used for this benchmarking study were created to mimic four regions in North America and subsequently corrupted to form four different data scenarios that were released to the homogenisation community. These scenarios incorporated different station densities, different combinations of step and trend inhomogeneities and different underlying data characteristics, namely different autocorrelations.

This chapter explains the validation framework implemented when assessing algorithm performance on these scenarios. The validation was split into two components; detection ability, which assesses an algorithm's ability to find change points, and adjustment ability, which assesses the similarity of the returned series to the clean series. Each of these will be discussed in turn. The validation concepts explained in this chapter will be implemented in the next chapter.

6.1. Validation framework

There are a plethora of studies that have assessed the performance of homogenisation algorithms, many have compared multiple algorithms, such as Venema et al. [2012] and Reeves et al. [2007], whilst others have assessed variants of the same algorithm Titchner et al. [2009] and Williams et al. [2012]. Within studies evaluating the performance of homogenisation algorithms different aspects of an algorithm's performance have been the focus. The predominant focus until recent years was on an algorithm's ability to correctly identify the location of a change point, where an algorithm was rewarded for identifying a change point within the window of a true change [Menne and Williams JR., 2005]. The correct identification of change points is commonly assessed using the hit rate (proportion of change points found out of total number present). Complements to this statistic are the false alarm rate (how many change points are wrongly inserted out of the number of points where no change is present in reality), the false alarm ratio (the proportion of the detected change points that are false) and the type one error rate (the percentage of times a homogeneous series is incorrectly classified as inhomogeneous) [DeGaetano, 2006; Venema et al., 2012]. Many other measures exist, and many other names for the same measures exist, as was outlined in section three of chapter two. In this study the hit rate and the false alarm rate were the focus measures, alongside the

bias and critical success index, as will be further explained in section two of this chapter.

Recently, as stated in chapter 2 section 2.3, attention has started to focus on the quality of data returned by an algorithm and not just an algorithm's ability to detect change points. Both detection and adjustment ability need to be examined, as good performance in one does not guarantee good performance in the other [Venema et al., 2012]. The quality of returned data can be assessed by measuring the difference between the clean and returned series; this was done using the centred root mean squared error (CRMSE) by Venema et al. [2012] and the root mean square prediction error (RMP) by Reeves et al. [2007].

Not restricting assessment to just detection ability allows for inspection of an algorithm's capacity to return realistic climate trends in homogenised data, an aspect of growing interest in recent decades. Venema et al. [2012] used the RMSE to compare trends in clean and returned series, but the approach outlined in Willett et al. [2014] using the percentage recovery was employed in this thesis. This is a new approach that can be used across multiple aspects of the validation.

Not looking purely at detection ability also has the advantage that an algorithm's ability to correct for trend inhomogeneities can be better assessed. The change points that happen at the start and end points of trend inhomogeneities are difficult to detect and trends are often misclassified as step changes [Menne and Williams JR., 2009], but this does not necessarily mean that their effects haven't been corrected for. Looking at more than just detection ability means that algorithms can be credited for lessening a trend inhomogeneity's impacts even if they do not find its end points.

Although there is still benefit in validation studies that assess only one aspect of algorithm performance, there is a growing need for benchmark datasets that test an algorithm's performance against multiple, different featured, datasets that are more similar to the real world. Such tests benefit from being blind as this avoids the danger of algorithms being tuned to perform well only in certain situations [Willett et al., 2014]. For the present study homogenisers were given the final location of the stations, but very little other information about the datasets used. They knew roughly what aspects would be present in the created datasets; trend inhomogeneities, step inhomogeneities and inhomogeneities created using the model or using constant offsets, but they did not know the quantities or distribution of these aspects. They were also given a brief outline of how the algorithms would be assessed, this highlighted that both detection and adjustment ability would be evaluated, even so, not all algorithms returned sufficient information for both these qualities to be assessed. The main emails sent to homogenisers explaining the study can be found in Appendix A, the website where homogenisers downloaded the data from is `http://www.metoffice.gov.uk/hadobs/benchmarks/`.

6.2. Breakpoint Detection ability

Although arguably less important than its adjustment ability, an algorithm's detection ability is still important. Whether missing inhomogeneities is better or worse than adding in spurious inhomogeneities is a much debated and contentious subject, but the fact remains that both are undesirable algorithm tendencies [Domonkos, 2011; Willett et al., 2014]. For this reason it is important to assess not just an algorithm's ability to find true change points, but also the rate at which it 'detects' false change points.

Large inhomogeneities are easier to detect than smaller inhomogeneities given their greater signal to noise ratio. The effect of leaving these larger artefacts in can be more detrimental to the climatic series. However, missing multiple small discontinuities is also undesirable as this can still lead to large errors in the analyses. Within this validation, inhomogeneity size was taken into account and it was the relative size that was being referred to. The relative size is the size of an inhomogeneity relative to the next homogeneous sub period, as defined in section 4 of chapter 5. Also, as the previous chapters showed, the created benchmarks are not perfect and this too was taken into account when evaluating the algorithm performance. This means that inhomogeneities that may be unintentionally present in the clean series did not contribute to either an algorithm's hit rate or its false alarm rate.

6.2.1. Breakpoint Detection ability concepts

Various statistics exist to summarise an algorithm's performance in terms of breakpoint detection, many of which were listed in section two of chapter two. These statistics are largely formed from the combination of four quantities; hits, false alarms, misses and correct rejections, which are commonly given the letters a, b, c and d respectively, with their sum being n [Hogan and Mason, 2012]. These quantities can be defined as follows and can be seen in table 6.1:

- Hit - a change point allocated to a day/window where there was a change point present.
- False alarm - A change point allocated to a day/window where there was no change point present.
- Miss - No change point allocated to a day/window where there was a change point present.
- Correct rejection - No change point allocated to a day/window where there was no change point present.

In annual data, where 'day' is replaced by 'year' in the definitions above, the quantities a, b, c and d are of a similar enough order to each other to ensure that they can be easily combined to form the desired summary statistics. At the monthly or daily level the fre-

Table 6.1. A table to illustrate the events that give rise to a, b, c and d. Adapted from Hogan and Mason [2012].

	Change point present		
Change point allocated	Yes	No	Total
Yes	a (Hits)	b (False alarms)	$a + b$
No	c (Misses)	d (Correct rejections)	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = n$

quency of correct rejections should vastly outweigh the frequency of all other quantities. This presents problems when calculating validation measures as they can easily become swamped by d, blurring the distinction between different algorithms. For this reason a novel approach to defining these quantities was applied in this study. This approach involved defining windows of time in the series. These windows could either be a change point window, which could be a hit or a miss; or a homogeneous sub period (HSP) window, which could be a false alarm or a correct rejection.

The idea of windowing change points is not a new one, it is common practice in homogenisation studies, see for example Easterling and Peterson [1995], DeGaetano [2006] or Menne and Williams JR. [2009]. The reason for this approach is that, although exact detection would be the ideal in homogenisation, it is unlikely to occur in practice, especially when the series are at the daily resolution. Having a series at the daily resolution was a challenge in this study as most contributed algorithms homogenised data at the monthly level and then interpolated these changes through time. These algorithms then gave the date of change as the first or last day of the culprit month, substantially reducing the chance of specifying the true date of the change.

Two change point window lengths were considered, a 60 day window, allowing 30 days either side of the true change point, and a 180 day window, allowing 90 days either side of the true change point. The reason for using two window lengths is twofold. Firstly, it allows the assessment of whether an algorithm is consistently detecting change points, but consistently being beyond a month out; in this case the 60 day window would record a miss, but the 180 day window would record a hit. Secondly, many algorithms still have an operational minimum time between detections, even if theoretically none exists. For the contributions assessed in this work this operational limit was never more than six months. Therefore, in having a 180 day window, a change point that has been placed by an algorithm between two true change points, because of the short time separating them, can be recorded as a hit for both change points. In the event that an algorithm detects two change points within a single change point window only one hit is recorded, this is to ensure that it is impossible for an algorithm to be credited with more hits than there are change points. The other hit is ignored as it would be incorrect to count it as a false alarm, miss or correct rejection.

An HSP window is a newer concept than a change point window, it treats all days between two change point windows (or between a change point window and the end of the series) as a single period. Thus, most of the time, a series will have m change point windows and $m + 1$ HSP windows, though there will be a few exceptions when two change points

are so close together that there is effectively no HSP between them. Each HSP has the capacity to be either a correct rejection for the algorithm or a false alarm, this solved the problem of correct rejections swamping everything else. If there are multiple false alarms within an HSP window the algorithm is still only given one false alarm, just as it is given only one hit for multiple hits within the same change point window.

Only penalising an algorithm once for any number of false alarms in a single HSP window has both advantages and disadvantages. One advantage is that it is simple to explain and implement. A second advantage is that it means hits and false alarms within their respective windows are treated in the same manner. A third advantage is that n will be constant across all methods in a region, thus making comparisons of different algorithms very easy. A disadvantage of this method is that it could feasibly result in a very error-prone algorithm that inserts multiple false alarms into every HSP window only being penalised by the same amount as an algorithm which inserts just one inhomogeneity into each HSP window. This point is raised to caution the reader that an algorithm's performance should be considered as a whole; looking at both similarity to clean series and detection capability. An algorithm that is prone to adding too many false alarms may not be penalised harshly enough in detection ability measures, but the impact of them will likely show in adjustment ability measures.

Given that at the station level the defined a, b, c and d will be very small the detection ability measures were all assessed on the region as the whole. That is, a was the total number of hits in a region, b was the total number of false alarms, c was the total number of misses and d was the total number of correct rejections; where all were defined using the windowing approach.

6.2.2. Breakpoint Detection ability measures

As stated in section one of this chapter, two commonly used measures when assessing algorithm performance are the hit rate, also known as the probability of detection, and the false alarm rate, which is the same as the probability of false detection.

The hit rate is defined as $H = \frac{a}{a+c}$, that is, the total number of hits out of the total possible number of hits. For a perfect algorithm the hit rate would be one and the worst hit rate score is zero, no values outside this range are possible.

The false alarm rate is defined as $F = \frac{b}{b+d}$, which here is the number of false alarms divided by the total number of HSPs. Again, the possible range is between zero and one, with zero being perfection and one signifying that all homogeneous sub periods have been corrupted by the homogenisation process. Both the hit rate and false alarm rate will be reported for each algorithm and also shown graphically. The advantage of a graphical presentation of such measures is that it provides a good visual illustration of the differences in the performance of different algorithms and different scenarios.

Given that defining d in the daily data situation is tricky, it is beneficial to have some

measures that do not include this term. The measures chosen here were the frequency bias and the critical success index (CSI) [Hogan and Mason, 2012]. Frequency bias is defined as $B = \frac{a+b}{a+c}$ and gives information about whether an algorithm is prone to over- or under-estimating the number of change points in a series, though it does not take into account whether these change points are in the right place! CSI combines a, b and c as $CSI = \frac{a}{a+b+c}$, it gives more information than the hit rate as, by the incorporation of b, it also takes into account how the algorithm performs in the absence of change points.

Given that bias and CSI do not incorporate d, the windowing method could be somewhat relaxed. HSP windows no longer needed to be constrained to count as only one false alarm or only one correct rejection. Instead each false alarm could be counted, which will distinguish between an algorithm very prone to over-inserting change points and one that has only a slight tendency to do so. This change to the definition of HSP windows makes defining change point windows slightly more of a challenge. To be equivalent to the approach for HSP windows, each hit in a change point window should be counted, but this would allow the possibility, however small, of there being more hits allocated than there were change points present. For this reason change point windows were still constrained to be a single hit or miss.

In all these measures, consistency of calculation across the same measure was key. As long as for each region, scenario and method the measure was calculated in the same way then direct comparisons could be made allowing a fair review of the strengths and weaknesses of the various algorithms.

6.3. Methods for assessing similarity of clean to returned series - Adjustment ability

It was not expected that all inhomogeneities would be detected and, in the unlikely event that they were, it was not expected that they would be perfectly corrected for. This is the reason that it was important to assess the similarity between the clean and returned series; to enable a fuller picture to be formed of how well an algorithm had homogenised the released data.

It is important to highlight at this stage that homogenisation algorithm performance was the primary focus of this study and not interpolation algorithm performance. For this reason, all returned data that had been interpolated over missing periods were made missing to the same level as the released data. Similarity measures were then calculated. This ensured fair comparisons across homogenisation algorithms regardless of whether they also incorporated interpolation algorithms.

6.3.1. Adjustment ability concepts

When comparing clean and released and clean and returned series it was possible to work on a station by station basis as well as on a region wide basis, whereas for detection ability the numbers were too small for individual stations to be analysed, and the most sensible option was to look at measures regionally. However, to report individual statistics for each station pre- and post-homogenisation would be time consuming and largely uninformative. For this reason adjustment ability statistics were made available to homogenisation algorithm developers for all stations for their benefit, but only summary plots and measures are included in this thesis.

To show the progression of a particular statistic of interest over time required the regional aggregation of the data into what will here be defined as a *compiled series*. Compiled series' were created both at the monthly and daily levels to balance conveyance of information in the level of detail with interpretability in the quantity of information. That is, a daily plot is more detailed, but a monthly plot can be more easily interpreted and is unlikely to hide any major findings, though both were produced to be certain of this. In chapter seven only example monthly plots are included, but the author checked that the conclusions stated were also valid for the daily plots.

The methodology for creating a compiled series is best explained with an example. The code relating to this example can be found in electronic appendix B. Suppose that a series for the RMSE in a region as a whole is desired. For Wyoming scenario one there are 75 stations. Thus there are 75 values (some of which may be missing) for each day from the 1st January 1970 to the 31st December 2011. In an iterative loop each day can be selected in turn and the RMSE for that day between clean and released data and clean and returned data can be calculated. This takes all 75 stations into account, but returns just a single value for each time point. These values can then be formed into a series. The principle is the same if the time period of focus is longer such as months from January 1970 to December 2011, or if the desired series only covers a single year, but draws data from all years. In this latter case the iterative loop would go from the 1st January to the 31st December and take all information from these days regardless of year and station. So all January 1sts from all stations would be used to get a single value of the RMSE for January the 1st. It doesn't matter that later years would be expected to have a lower RMSE as this would be a consistent pattern for all days of the year and therefore would affect all days in the same way. (The reader is reminded at this point that RMSE is lower in later years because inhomogeneities were propagated backwards in time).

Another concept, that was mentioned in section 6.1, is that of *percentage recovery*. This can be defined for a statistic as:

$$PR = [\frac{(statistic_{released} - statistic_{returned})}{(statistic_{released} - statistic_{clean})}] * 100.$$

This conveys how much change there has been in the value of a statistic between the released and the returned data. A value of zero indicates that no change in the statistic has taken place, that is, the returned data is no better or worse than the released data.

Values greater than zero mean that the statistic has been moved in the right direction (i.e. towards perfection) and the closer this value is to 100 the better. For some statistics PR can overshoot 100 percent; this would mean that too much of a modification has been made and if the value goes over 200 then the series is worse than it was before homogenisation, that is, it is further from the truth. A value less than zero means that an algorithm has moved the statistic in the wrong direction so that it is further from the truth on return than on release. The ranges of PR possible will be discussed alongside the statistics in question, but the range for a single statistic will be consistent across different regions and methods and therefore these values will be directly comparable. If no inhomogeneities are present in a series then the denominator of the PR formula is zero and PR is incalculable. In such cases, an algorithm, provided that it did not corrupt the series, was given a PR value of 100 for all statistics relating to that series.

PR should be considered alongside other measures to ensure that it does not give misleading conclusions. For example, a near perfect station could have a PR of 0 because no changes to the station have been made as none were very necessary, this is not as much a cause for concern as if a very corrupt station also had a PR of zero.

One way to use PR to convey information without fear of being misleading is to create percentage recovery plots, these allow the extent to which the improvement was necessary to be seen, as well as the range that the PR values themselves fall into. An example of such a plot can be seen in figure 6.1. This is the PR plot for linear trends in the ten worst (as defined in the following paragraph) stations in Wyoming scenario one after their homogenisation by MAC-D, the algorithm applied by Michele Rienzner of the University of Milan [Rienzner and Gandolfi, 2013]. It is evident that Mac-D is doing very well in this aspect of homogenisation as all returned series trends are closer to the clean series trends than the released series trends were and some trends have been moved to near perfection.

The classification of 'best' and 'worst' stations was made so that the stations most affected by the corruption process and those least affected by the corruption process could be highlighted. Good algorithms should improve the homogeneity of the worst stations, whilst not damaging the quality of the best stations, and even improving it where possible. The groupings into best and worst stations were done based on the RMSE between a clean time series and its released counterpart. Many of the best stations were already perfectly homogeneous, if more than ten such stations existed in a region then preference was given to those which the PHA also deemed to be homogeneous for the released data. As this did not narrow down the field drastically, ten were then chosen at random from all the possible candidates - these ten remained consistent across all algorithms and measures to ensure fair comparisons.

Figure 6.2 displays the locations of the best and worst stations in each region and scenario on topographical maps. These show that there is no clustering in the station classifications, which is as expected given the random nature of the corruption process.

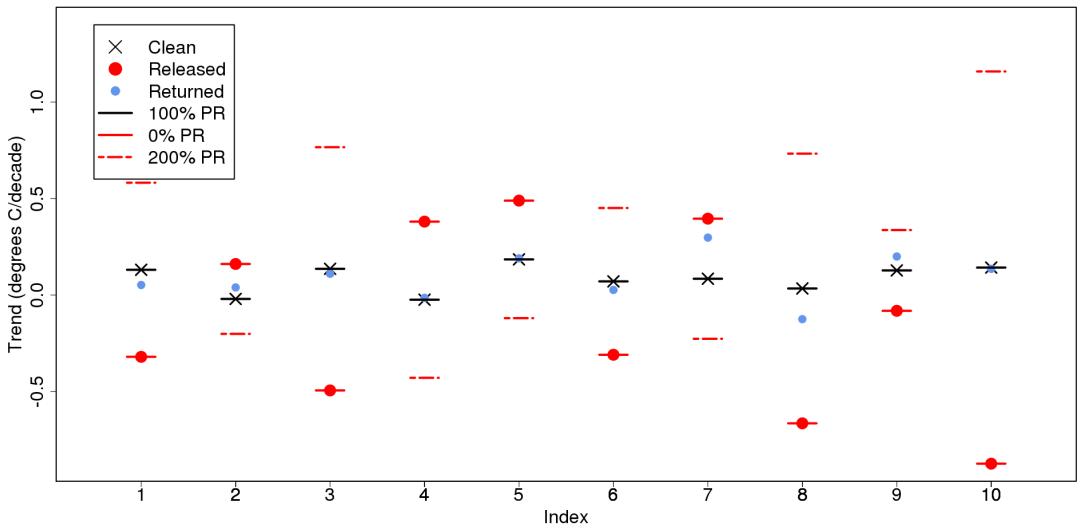

Figure 6.1. A plot to look at the percentage recovery for linear trends when using Mac-D for the ten worst stations in Wyoming scenario 1. Any points lying outside the red lines would indicate the trend has been made worse. Points between the solid red and black lines ($0 < PR < 100$) indicate the trend has been moved in the right direction. Points between the solid black and broken red lines ($PR > 100$) indicate the trend has been moved too far in the right direction, but is not as bad as before homogenisation.

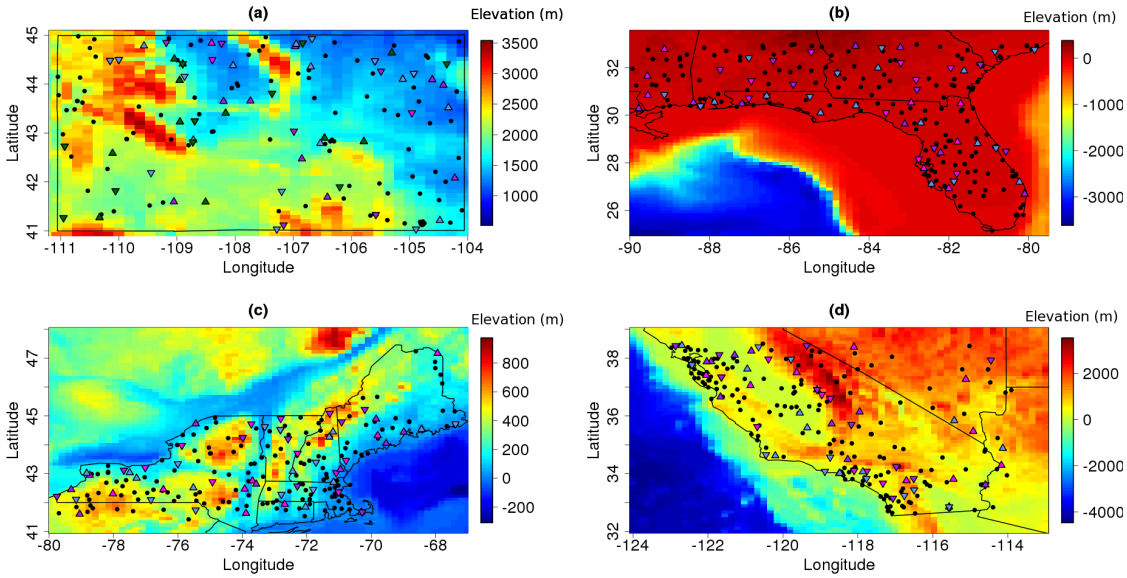

Figure 6.2. Figures to illustrate the locations of best and worst stations in (a) Wyoming, (b) the South East, (c) the North East and (d) the South West. Dots represent station locations, upward pointing triangles are the best stations and downward pointing triangles are the worst stations. Pink triangles are for scenario 1, purple are for scenario 2, blue are for scenario 3 and dark green are for scenario 4. Topographies to create these maps were obtained from the National Elevation Dataset [Gsech et al., 2002; Gsech, 2007].

6.3.2. Adjustment ability measures

Getting the mean right for temperature studies is important, as assessing changing means is valuable for impact studies. Therefore, two measures looking at the mean were assessed in this project, the bias and the RMSE, calculated between the clean and the returned series. These measures were looked at in comparison to the same measures between the clean and the released series because this allowed analysis of the impact

that homogenisation had had. Values were obtained for these statistics for the region as a whole, as well as for each of the stations in the region. For bias, the sum of absolute biases in a region was compared as this allowed the assessment of the magnitude of bias in a region and removed the possibility of the effects of stations with oppositely signed biases cancelling each other out. Plots for how bias and RMSE changed over the whole time period were provided to the algorithm developers. The percentage recovery was assessed for both bias and RMSE. PR was bounded by 100 as the upper end of its range for RMSE as it is impossible to get more than 100% recovery given that RMSE is restricted to be 0 or positive. The same restriction applies for the sum of absolute biases, but for the regional bias and bias of individual stations there was no such restriction on percentage recovery.

Bias was assessed as it is important to know if an algorithm is prone to making data consistently warmer or cooler. If there is a persistent bias at the homogenisation stage this could be addressed by post-processing (further altering) the results before carrying out climate analyses on the series, or by altering the internal workings of the algorithm itself. One of the aims of this study is to aid the development of daily temperature homogenisation algorithms and the information on the bias of returned data will be a valuable tool for this. Bias will also be helpful at the station by station level to assess the extent to which the original bias affects the bias of the returned data, i.e., is an algorithm prone to over or under compensate for original bias.

While bias compares only the means of series', RMSE was assessed as an overall measure of similarity between the clean and returned data. It can be shown to be composed of correlation, standard deviation and mean difference components. That is, the RMSE between two series x and y can be written as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(\bar{x} - \bar{y})^2 + (s_x + s_y)^2 + 2s_x s_y(1 - r_{xy})},$$

where $\bar{x}$ indicates a mean, s_x indicates a standard deviation and r_{xy} indicates a correlation [Murphy, 1988].

Both bias and RMSE were assessed using non-deseasonalised data. The primary reason for this is that, given that the bias is the difference in means between two series, if a series has been deseasonalised by having its mean subtracted then all biases will become zero, and thus, uninformative. As consistency was sought across measures where possible, in having bias assessed on non-deseasonalised data it made sense to also assess RMSE on non-deseasonalised data. A secondary reason was that, although deseasonalised data are preferred for trend analyses, as they are impacted less by inconsistent temporal and spatial sampling, non-deseasonalised data are often preferred for impact studies.

In addition to a series' behaviour on a daily basis, long term trends are also a major area of interest in climate studies as they provide information on how the climate is changing. Knowing how the climate is changing is important because without knowing this it is not possible to mitigate or adapt to these changes. Therefore, the evaluation of re-

turned trends in homogenised data is very important and this is illustrated by the fact that many other studies have focussed on trend evaluation [McCarthy et al., 2008; Venema et al., 2012; Williams et al., 2012]. Errors remaining in series because of undetected inhomogeneities can be comparable to the climate change signal [Williams et al., 2012]. Therefore, understanding the uncertainty in homogenisation is crucial when making statements about long term regional climate trends; though, it should be noted that, globally, homogenisation errors would have very little impact on the conclusion that our climate is warming [Willett et al., 2014; Karl et al., 2015].

In the present study linear trends were assessed for the time series overall and non-linear variability was calculated at the inter-annual and inter-decadal levels as being able to reproduce the low- and high-frequency variability of a series is also valuable. The overall trend is used for climate change quantification. Low-frequency variability should capture events such as ENSO and the NAO (North Atlantic Oscillation). High-frequency variability comparisons should be a good measure of whether algorithms exhibit a tendency to oversmooth series during homogenisation.

Trend assessments were always on deseasonalised data, otherwise the seasonal cycle biases the trend estimates, especially when there are missing data. The overall trends were assessed using linear regression on data aggregated to the yearly level. The reason for aggregating to this level was to reduce the variability in the data and cope with inconsistent temporal sampling as well as to reduce the autocorrelation in the data, as linear regression assumes independent inputs. For the decadal and annual long term variability calculations loess smooths were used, in a similar manner to that employed by Venema et al. [2012]. That is, for each time point in a series a prediction was made using the time points within a specified period and a fitted polynomial surface. For the annual loess this was done using $\frac{1}{42}$ of the data, which amounts to 6 months either side of the time point. For the decadal loess, $\frac{10}{42}$ of the data was used, which amounts to five years of information either side of the time point in question. In the present study a linear function was used as the basis for the loess regression to avoid the trends being allowed to vary too much over the relatively short time period, in Venema et al. [2012] a quadratic polynomial was used.

For the overall trend the linear regression coefficients were compared between clean, released and returned data, with the primary focus being on trends that had significant coefficients in any or all of these data groups. Significance here was defined at the 5% level. Trends that were significant in the clean series were expected to be significant, and of the same sign, in the returned series and trends that weren't significant in the clean series were expected not to be significant in the returned series, regardless of whether the corruption process had affected their significance. The number of trends that fell into each of these different categories was tabulated for easy comparisons across methods and regions. Percentage recoveries were also investigated for linear trend coefficients. There were no restrictions on percentage recovery for this measure as trends have no lower and upper limiting values.

For the annual and decadal variability comparisons, correlations were calculated between

the created loess smooths for equivalent stations in the clean and released and clean and returned data. These correlations were calculated having discarded the first and last six months of the data for annual smooths and the first and last five years of the data for the decadal smooths. The reason for discarding these data is that, at the ends of the time series, the loess smooths will continue on the same trajectory that the last prediction put them on, which could lead to misleading conclusions if the last trajectory was bad. The correlations between clean and released and clean and returned data were plotted to gauge how similar the loess smooths were. Percentage recovery was also used in relation to the correlations. The maximum percentage recovery was 100 for this measure as a perfect correlation of one was the upper limit.

As pointed out by Della-Marta and Wanner [2006], at the daily scale it is not sufficient to only focus homogenisation efforts on the mean of the temperature distribution. For this reason, in the present study, assessments of similarity between the clean and released and clean and returned series also focussed on variability comparisons. This was done by looking at the similarity in standard deviations in series before and after homogenisation relative to the standard deviations in the clean series and also by looking at similarity in extremes. Both these measures were analysed using non-deseasonalised data. For standard deviations deseasonalisation would not change the values, but for extremes analysis non-deseasonalised data is most beneficial, especially if this study were to be carried further to also analyse real world data where extremes have impacts on real people.

Standard deviation similarities were assessed by looking at ratios of the standard deviations for clean and released and clean and returned series to allow investigation into whether the homogenisation process had a tendency to smooth or accentuate true variability. Percentage recoveries of standard deviations were examined and there were no limits on the values these percentage recoveries could take. It is important to ensure realistic variability is retained during homogenisation to avoid the risk of creating or smoothing out climate extremes.

The climate extremes themselves were compared using scatter plots of the most extreme values on like-for-like days. That is, a homogenisation algorithm was not credited with a correct extreme if it had in fact removed the true climate extreme and created one of the same or differing magnitude on another day. As extremes are a climate artefact that can be examined with a lot more meaning at the daily data level, owing to them not being smoothed out in an aggregation process, this was an important aspect to this first benchmarking study in this area. Measurement error was allowed for in the comparison of extremes and this measurement error was deemed to be 0.14°C. Values on the same day that were within this level of precision were counted as sufficiently equal. The value for measurement error was calculated using a formula similar to that found in Brohan et al. [2006] who, knowing that each temperature recording had a random error of 0.2°C (which represents one standard deviation), stated that an average calculated using n measurements (in their case 60) would have an error of at most $\frac{0.2}{\sqrt{n}}$°C. In the present work $n = 2$ and therefore the calculation was $\frac{0.2}{\sqrt{2}} = 0.14$ (to two decimal places). The

extremes focussed on in this study were the warmest daily mean temperature recorded and the coldest daily mean temperature recorded at each station over the full period of the record.

6.4. Discussion

The author believes the subset of validation measures selected in this thesis gives a good overview of the participating algorithms' performance. However, there is inevitably more that could be investigated than could be achieved in this project.

An additional validation measure that it would be beneficial to include in a future iteration of this work would be another loess smooth comparison measure. The reason for this is that the included loess smooth comparison measure, which looks at correlations between loess smooths of the clean, released and returned data, was found to be flawed. When investigating the plots comparing correlations during the validation of the algorithms it was found that it was possible to have a higher correlation between loess smooths of the released and clean data than the returned and clean data, even if the data had been improved by homogenisation. This is illustrated in figure 6.3, which shows that the returned series is more similar to the clean series than the released series is, but its correlation is lower because it doesn't ascend and descend at the same time as the clean series as much as the released series does. As this was only realised during the validation of algorithms an alternative measure of long term variability assessment was not sought, however, a future iteration of this study would seek to evaluate this measure in a more reliable way. One suggestion is to look at the RMSE between the two smooths instead of the correlation between them.

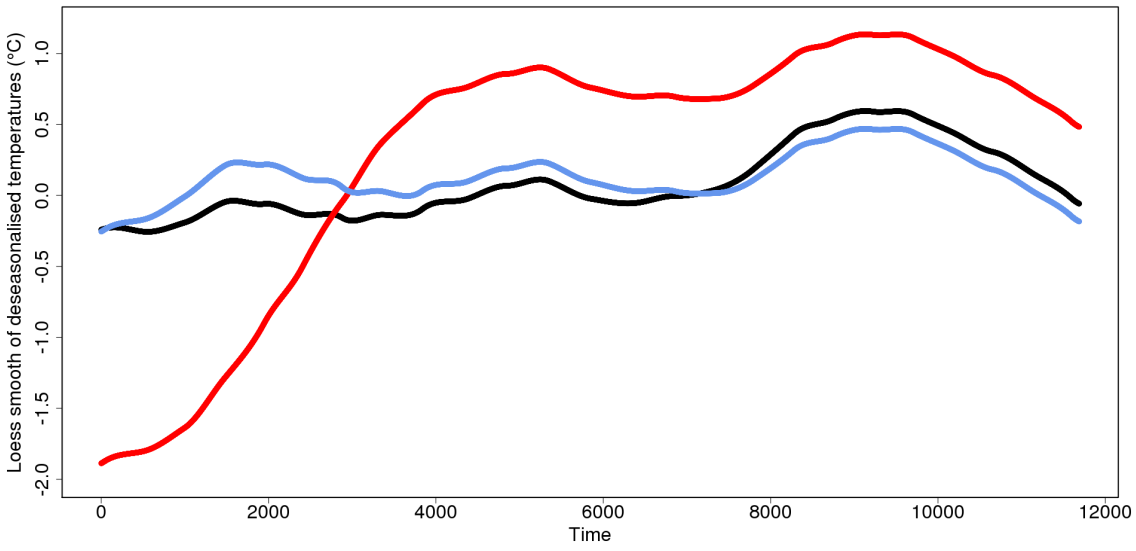

Figure 6.3. Figure to illustrate a loess smooth where the returned data (blue) could be considered to be as good or better than the released data (red) when compared to the clean data (black), but where the correlation between clean and returned loess smooths is lower than that between clean and released. Correlation between the clean and returned data loess smooth is 0.732 and between the clean and released data loess smooth is 0.915.

6.5. Summary

This chapter has introduced the validation framework that will be implemented for all algorithms in the next chapter. It highlights that considering only one aspect of algorithm performance should be avoided and for this reason details have been given for a range of measures to assess both the detection and adjustment ability of temperature homogenisation algorithms. Within these assessments focus is not only on the mean of the climatic series, but also on the impact of homogenisation on its variability and, therefore, its extremes.

Even with the range of measures given in this chapter it is by no means an exhaustive validation framework. Therefore, participants were and are encouraged to carry out further assessment of their contributions using the clean data, which has now been released to the homogenisation community. The information that homogenisers were given about how their algorithm performed is provided in electronic appendices available from `http://www.metoffice.gov.uk/hadobs/benchmarks/` and contained specific information about their algorithm performance in each region as well as summary information as to how it performed relative to the other algorithms in the study. No participants were given the data homogenised by other participants, only the summaries, but the author encourages data sharing to continue this collaborative work.

7. Benchmarking the Performance of Contributed Algorithms

Preceding chapters have laid the foundations that allow the performance of the contributed homogenisation algorithms to be assessed. Chapters three and four introduced the data and methodology for creating the benchmark clean data and chapter five explained how these data were corrupted to form the released data. Chapter six introduced the validation framework that was designed to compare the returned data contributions from homogenisers and this chapter implements that framework.

The chapter begins with an overview of the contributions submitted, their characteristics, and references where further information can be found. It then proceeds to evaluate these algorithms. At this stage any benchmark weaknesses that may affect algorithm performance are highlighted before the chapter continues with a summary of how the algorithms performed according to each validation measure. A short summary of each algorithm's strengths and weaknesses is then given in this chapter, but a more detailed summary, which provides plots and explanations for Wyoming scenarios and just plots and summaries for other scenarios, was provided to each participating homogeniser. These summary documents can be found in electronic appendices C to H. The chapter concludes with a discussion concerning the amount of uncertainty left in the data after homogenisation has occurred; a primary focus of this project.

7.1. Algorithms used

Seven distinct algorithms were run on the benchmark data. Variants of the same algorithm were also run, however, with one exception, only one realisation was evaluated for each algorithm. There follows a brief overview of each algorithm, important caveats to bear in mind and also references where more detailed information can be found. Each algorithm user was contacted to provide information on their algorithms and the subsections that follow are closely adapted from the information they provided. In particular, each algorithm user was asked to provide answers to the following questions:

1. What is the smallest distance between change points the algorithm can cope with?
2. Can the algorithm cope with gradual inhomogeneities?
3. Does the algorithm have two steps, one for detection and one for adjustment, or are the two performed simultaneously?

4. Are the adjustments made using constant offsets (thus primarily affecting the mean) or are they variable (allowing for a change in the variance as well)?

The author deliberately did not research the algorithms in detail before commencing the evaluation to avoid any biases inadvertently slipping into the analysis.

7.1.1. ACMANT

ACMANT stands for the Adapted Caussinus-Mestre detection Algorithm for homogenising Networks of Temperature series. This algorithm can be run automatically and was run for all regions and scenarios by its creator Peter Domonkos. It can be applied at the monthly or annual level, with daily data being aggregated up to the requisite level for assessment and with adjustments being evenly applied at the daily level at the final step. The smallest distance between change points it can cope with in this study is five months, as change points were analysed at the monthly level, thus a ninety day window will be more favourable than a thirty day window when evaluating the accuracy of change point locations.

Time series comparisons were carried out with composite reference series, that is a composite of multiple other stations. This composite reference series was decided using weights obtained through spatial ordinary kriging, with some minor adjustments. These minor adjustments were as follows: no station was allowed to contribute more than 40% of weight to the reference series, to avoid contributing its own inhomogeneities; negative covariances between stations were not allowed; and at least 6 stations had to be available for ordinary kriging to be applied (in practice in this study at least six stations were always available). Optimal step function fitting together with the Caussinus-Lyazrhi criterion [Caussinus and Lyazhri, 1997] was used for deciding the location and number of steps respectively and adjustment sizes were determined using ANOVA variance minimisation [Caussinus and Mestre, 2004]. The detection and adjustment were distinct steps. Only biases in the means were deliberately searched for and adjustments were constant for the period over which they acted, that is, there was no seasonal variation. Gradual inhomogeneities may have experienced some degree of correction if steps were corrected mid-trend, but no search for gradual inhomogeneities was explicitly built in.

The variant of this algorithm analysed is a slightly modified version of the freely available ”Acmant2Tmindaily” program of the ACMANT2 software package, `http://www.c3.urv.cat/data.html`. For a full description of the method the reader should consult Domonkos [2011] and Domonkos [2014]. It should be noted that since the time of analysis a new experimental version of ACMANT has been created by Peter Domonkos that works directly with daily data for certain steps of the homogenisation process.

7.1.2. Climatol

Two versions of this algorithm were evaluated as one was explicitly designed to work with daily data, whilst the other worked with data at the monthly level and then downscaled adjustments. Henceforth the two versions will be referred to as Climatol-Daily and Climatol-Monthly. A third version was also run, but was not assessed here. All versions were run for all regions and scenarios. Climatol is freely available from `www.climatol.eu` and is there accompanied by an instruction manual, which is recommended as the primary reference for this algorithm.

This algorithm is designed to run in R and many settings can be programmed by the user, thus meaning that different users will inevitably get slightly different outcomes. The default behaviour of the algorithm is to apply variable corrections, which was done for Climatol-Monthly, however, for Climatol-Daily constant adjustments were applied in each homogeneous sub-period. It should be noted here that there was a mistake in the homogenisation during the application of Climatol-Monthly for Wyoming scenarios one and two, therefore comparisons for these regions with others should be approached with caution. This mistake was a simple error of making adjustments using the wrong units and not a general flaw in the algorithm. Corrected results were provided, but only after the clean data had been released and, therefore, they could not be fairly included in the comparisons.

The minimum time between change points that is theoretically detectable is just three time steps, though at the daily scale it is unlikely that such a limit is achieved. Change points are searched for using the Standard Normal Homogeneity Test (SNHT) [Alexandersson, 1986] iteratively applied until no more change points are found when the candidate series is compared to a composite reference series, created from nearby stations. The SNHT splits the series at the most significant change point until all sections are deemed to be homogeneous sub-periods. This procedure is applied first on overlapping windows, to avoid possible masking effects of multiple change points in the series, and then on the whole series, where SNHT is more powerful. Adjustment sizes are not explicitly calculated and corrections are made as a separate step. This step entails using the homogeneous sub-periods to estimate the missing data that have been created by the splitting process. These estimates are calculated by normalising the data that are available and then filling in the missing data using weighted averages of the closest normalised data. This normalisation involves centring the distribution if constant corrections are to be applied, as in Climatol-Daily, and centring and standardising if variable corrections are to be applied, as in Climatol-Monthly. Gradual inhomogeneities are not sought out, however a step change may be inserted in the middle of a gradual inhomogeneity if a noticeable change in mean is found.

7.1.3. DAP, HOM and SPLIDHOM

Two versions of each of these algorithms were applied, but only the first was analysed. The difference between the two versions is that the first homogenises a station using its most highly correlated neighbour as the reference series and the second homogenises a station using its second most highly correlated neighbour as the reference series. For these algorithms not all stations were homogenised and released stations therefore had to be treated as returned stations for around half of the stations in each region (although this figure varied between regions and scenarios). These algorithms were applied to all scenarios for Wyoming, the South East and the South West, but only to scenario one for the North East. Results were provided for scenarios two and three of the North East at a later date, but as this was after the clean data had been released they could not be included as part of this blind benchmarking study.

For all three of these algorithms the detection and adjustment steps were performed separately and did not seek out trend inhomogeneities. The detection was performed at the annual, seasonal and monthly levels, with most weight given to the seasonal and annual evaluations. The change point was always assigned to the first day of a given month because of the automation of the homogenisation process. In practice, the change points were almost always assigned to the 1st of January unless it was very clear to the algorithm that a different month was the culprit. However, it should be kept in mind that these algorithms were tuned to European data originally and different decisions should perhaps have been implemented when applying them to US data. For these algorithms it was recommended that change points were not sought out within four to five years of another change point when analysing data at the monthly level, however, when analysing data at the daily level a much smaller time window of half a month was the minimum time between change points. Change points were searched for using multiple tests, the SNHT [Alexandersson, 1986], the Maronna and Yohai bivariate test [Potter, 1981] and the Easterling and Peterson test [Easterling and Peterson, 1995]. If a certain proportion of the tests detected a change point then that change point was deemed worthy of adjustment [Stepanek et al., 2013]. The change point detections and adjustments were two separate steps and the same change points were identified for all three homogenisation algorithms simultaneously. The adjustments were calculated for each algorithm separately and were therefore different between algorithms, although still very similar. The adjustments applied were variable meaning that changes in the means, variances and also higher order statistical moments of the temperature series could be corrected.

The Higher Order Moments (HOM) method was the first algorithm out of DAP, HOM and SPLDHOM in use in the homogenisation community and an excellent summary of it can be found in Della-Marta and Wanner [2006]. From this paper, the homogenisation process is as follows. Once the homogeneous sub-periods (HSPs) of a candidate station have been identified a reference station with data suitably spanning a change point is chosen, beginning with the most recent change point. A non-linear model is then used to determine the relationships between the candidate and reference station before and after the change point. The model before the change point is used to predict temperatures

after the change point and from this a difference series between observed and predicted temperatures is formed. These differences are binned according to the predicted temperatures in the appropriate decile of the probability distribution for the most recent time period. A smoothly varying function is then fitted between the binned decile differences to obtain an estimated adjustment for each percentile. The percentiles of the HSP needing adjustment (i.e. the period before the change point) are also found by obtaining the probability distribution of that HSP and binning the observations into deciles so that the appropriate adjustments can be applied to each point. Finally the adjustments are made and the two HSPs are merged. The process is repeated until there is only a single HSP.

The SPLIDHOM (SPLIne Daily HOMogenisation) algorithm is detailed in Mestre et al. [2011]. It is similar to HOM in that it requires predefined HSPs, but differs in its method of correction. Whereas HOM defines distributions in order to make adjustments, SPLIDHOM relies on a regression approach. In this approach a highly correlated reference station is required for each change point, just as it is for HOM, and, as in HOM, each change point is considered sequentially beginning with the most recent. Before and after a change point the regression of the candidate station (Y) on the reference station (X) is estimated, this is done using a cubic smoothing spline approach. Details of splines can be found in chapter four, and specific details about the fitting of these cubic splines can be found in Mestre et al. [2011]. If the optimal smoothing spline is defined as $\hat{m}$ then the corrected values of the candidate series can be defined as $\hat{Y}_t = Y_t + \hat{m}_{YX_{aft}-YX_{bef}}[\hat{m}_{XY_{bef}}(Y_t)]$ where the subscripts bef and aft relate to whether the spline regression estimate was obtained before or after the change point. Thus, the SPLIDHOM method does not only correct the means of the time series, but also does not require full probability distributions to be estimated before and after each change point. The study of Mestre et al. [2011] compared the performances of HOM and SPLIDHOM and found HOM to be superior for reducing RMSE and also for improving the lower quantiles, but the performance of SPLIDHOM was deemed equal to the performance of HOM for upper quantiles.

The Distribution Adjusted by Percentiles (DAP) algorithm is detailed in Stepanek et al. [2013]. It is similar to the HOM method, as it is adapted from a method for the correction of regional climate model outputs detailed in Deque [2007], which in turn was based on assumptions used in Della-Marta and Wanner [2006] for HOM. DAP calculates the percentiles for the candidate and reference series before and after a change point for each month in turn. When focusing on a particular month, one month immediately before and after is also taken into account to reduce the likelihood of sudden large step changes between months. Once the percentiles for the candidate and reference station before and after a change point have been calculated, their difference can be obtained. The differences are then smoothed to obtain an adjustment for each percentile, which can then be applied for any point of the candidate series before the change point. The process is carried out iteratively in order to obtain the most precise results.

7.1.4. MAC-D

The MAC-D method searches for Multiple Abrupt Changes in the mean value of Daily temperature series. It is able to effectively deal with seasonality, non-periodicity, missing data and autocorrelations in the data examined. It is an automated method and is documented in Rienzner and Gandolfi [2013]. It was applied only to the Wyoming scenarios in this study. MAC-D is explicitly designed to work with daily data, as the name suggests, and has a limit between change points of just 15 days in this instance. The allowable time between change points is an adjustable parameter and can be as little as ten days, but this then increases the possibility of false alarms and makes determining amplitudes harder. MAC-D is not designed to seek out trend inhomogeneities and these will therefore be either neglected or corrected as a series of steps. However, although not yet implemented, it is hoped that, in the future, post-processing of the data could take place in order to detect trends between change points or within a stairway of change points.

MAC-D needs to work on groups of closely related station series. These were made by cutting the set of stations in a scenario into groups according to their correlation with their regional signal. The regional signal for MAC-D is defined as the average disturbance from the periodic seasonality which is common to the whole dataset. Lower performance is expected for stations with a weak link to their regional signal, termed 'bad' stations, and better performance is expected for those with a stronger link, termed 'good' stations. In this analysis the groupings were not taken into account, but comparisons of performance explicitly on 'good' and 'bad' stations would be an interesting area for future study. After the series have had their seasonal periodicity and regional signal removed, autocorrelation coefficients and change points are identified via an iterative process that stops when the autocorrelation (first and second order lags) of the filtered and de-stepped series (mean adjustment only) is close to zero. At the core of each iteration, the change points are located using the Standard Normal Homogenisation Composite Method (SNHCM), [Rienzner and Gandolfi, 2011], applied to the decorrelated version of the series that still retains the change points. The decorrelated version is the version of the series with no significant autocorrelations at lags one or two. This method is a change point detection method only and the adjustments therefore had to be applied separately. In this work these adjustments were calculated as constant offsets and were therefore expected to primarily affect the mean and not higher order moments.

7.1.5. MASH

The Multiple Analysis of Series for Homogenisation (MASH) method is the oldest of the methods applied in this study. It was originally developed in the 1990s [Szentimrey, 1999], for monthly, seasonal or annual data, though the version applied here is that which can be applied to daily data and is detailed in Szentimrey [2008]. The procedure has now been automated and this allowed MASH to be run on all regions and scenarios in this study.

MASH first aggregates the daily data into monthly series in order to search for inhomogeneities through comparisons with neighbouring stations. The smallest time between change points is one month, though in this study the change point detection ability of MASH was not assessed owing to it being different from that of other algorithms. Other algorithms will typically return a single month (or day) of a detected change point, whereas MASH returns the suggested change point sizes in the monthly time series and does not attribute each change point to a single month, so a single change point can lead to multiple 'detections' [Venema et al., 2012]. However it should be noted that Venema et al. [2012] still commended MASH as one of the best homogenisation methods available and did provide an analysis of its detection ability. It should also be noted that a differing detection method did not preclude comparison of MASH with other algorithms when using the adjustment ability validation measures in the present study.

MASH's method of smoothing monthly inhomogeneities to the daily scale means that adjustments are not constant and therefore can indirectly reduce the impacts of variance changing inhomogeneities. MASH is an iterative procedure and performs detection and correction of inhomogeneities simultaneously. During this procedure confidence intervals are provided on the size and location of the suggested shifts in order to help synthesise the final series. These confidence intervals allow metadata to be used automatically, though in this study no metadata was provided to the homogenisers. In addition, the intervals could be used to provide uncertainty estimates on the homogenised data. The iterative procedure is run until no more inhomogeneities in the monthly series are identified, this will likely lead to a reduction in trend inhomogeneities, though no trend inhomogeneity search is explicitly built in.

One final cautionary note should be raised when looking at comparisons of MASH with the other evaluated algorithms; MASH does not always homogenise to the most recent period. This is the case for one of two reasons: 1. The last period is not deemed to be homogeneous. 2. There is only a single inhomogeneity and it is very close to the end of the series, thus correcting the small final segment of the series is more time effective than correcting the much longer earlier segment and gives fewer opportunities for errors to arise. When evaluating the algorithm these different reference periods were noted, but were not explicitly taken into account, thus, there may be occasions where MASH appears to have a worse performance. The reader is directed to the more detailed appendix on MASH's performance, electronic appendix H, for more information on which measures have been affected in each scenario and region.

7.2. Algorithm assessment

7.2.1. The scenarios and regions

Section two of chapter five provided a detailed description of each of the scenarios. However a summary is provided here for the reader. Scenario one represents the current

density of stations that are at least 75% complete over the period 1970-2011 and has all three types of inhomogeneity added; shelter changes, station relocations and urbanisation trends. Scenario two also has all three types of inhomogeneity added, but has an increased station density, making the density of stations in all regions equal. Scenario three has the same station density as scenario two, but does not contain any urbanisation trend inhomogeneities, thus allowing the assessment of their impact on algorithm performance. Finally, scenario four was for Wyoming only and has the same inhomogeneities as Wyoming scenario one; its contributing difference is that it has higher and more realistic station autocorrelations than the other scenarios, but similar, if slightly reduced inter-station correlations.

7.2.2. Known benchmark weaknesses

As this study was a first attempt at a substantial daily benchmarking study there were expected to be imperfections in the created data. Chapters four and five showed that these imperfections did not stop the data being valid for this study, but they should still be kept in mind when evaluating algorithm performance.

One weakness is that scenarios one to three did not have realistic autocorrelations in deseasonalised difference series, they were too low. This should, in theory, make detection easier as most algorithms assume little or no autocorrelation in these series. This was the reason for the creation of Wyoming scenario four which better mimicked observed autocorrelations. The reader should take into account these differences in scenario characteristics when drawing their own conclusions about the different algorithms assessed.

A further known benchmark characteristic was the tendency to make inter-station correlations too high. This was deemed preferable to making them too low, which would have likely resulted in a benchmark harder than reality. Having a benchmark harder than reality would make drawing conclusions about areas for algorithm improvement more difficult and would have thus reduced the benefit of this study to algorithm developers. In scenario four some very nearby stations had inter-station correlations that were a little too low, but this only affected 20% of stations within 75km of each other and the most an inter-station correlation was too low by was 0.06. 75km was chosen as the distance cut-off point for this investigation as this is the distance used in the quality control checks for the GHCND [Durre et al., 2010].

One final cautionary point arises from the error in the addition of trend inhomogeneities for scenarios one, two and four. This error meant that trend inhomogeneities created using explanatory variables began with a step change. Given that trend inhomogeneities are known to sometimes start out with a step change in reality, [Menne and Williams JR., 2009], this should not be considered too detrimental to the algorithm evaluation. However, it is expected to lead to a greater proportion of trend inhomogeneities being detected than might otherwise be the case.

7.2.3. Detection ability

Section two of chapter six introduced the framework for the validation of algorithm detection ability, which included assessing the hit rate, false alarm rate, frequency bias and critical success index of an algorithm. Figure 7.1 shows the detection abilities in Wyoming for a 60 day, figure (a), or 180 day, figure (b), detection window and figures 7.2, 7.3 and 7.4 show the same for the South East, North East and South West respectively. It is evident that in Wyoming there is more spread between algorithm performance in the different scenarios than there is in the other regions. The difference between scenario four and the other scenarios is understandable as it was designed to be more realistic and therefore more difficult. The increased difficulty of scenario four is reflected by the lower hit rates exhibited in this scenario by all algorithms shown.

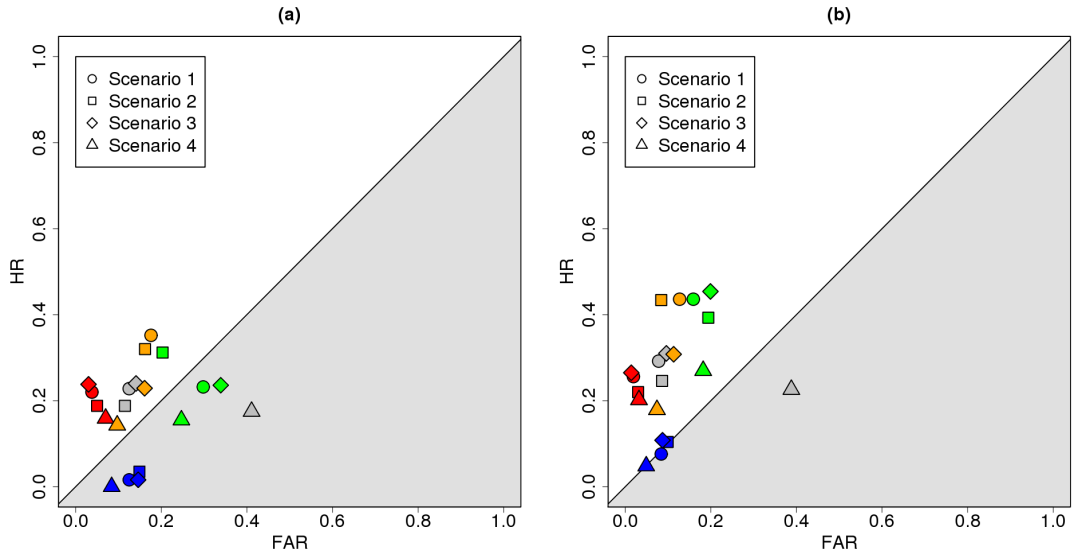

Figure 7.1. Plots showing false alarm rate against the hit rate in the Wyoming scenarios for (a) a window extending thirty days either side of the true change point and (b) a window extending ninety days either side of the true change point. MAC-D is gray, Climatol-Daily is red, Climatol-Monthly is orange, ACMANT is green and DAP, HOM and SPLIDHOM are blue. DAP, HOM and SPLIDHOM are all represented by a single blue point for each scenario because the detection approach was the same for all three of these algorithms and so the same change points were found. The shaded area below the diagonal line indicates the 'bad' area where the false alarm rate is higher than the hit rate.

The highest hit rates were always exhibited in scenario three for the South East, North East and South West when the larger window sizes were being used. Once more, this is as expected because this was arguably the easiest of the scenarios; it contained no trend inhomogeneities and also had an increased station density relative to scenario one. Trend inhomogeneities are known to often be harder to detect because they can have different start and end points with respect to different neighbouring series citepMenne2008. In this study trend inhomogeneities were also the hardest inhomogeneities to detect because they predominantly fell in the category of 'small' inhomogeneities (≤ 0.2°C) whereas the size classification of step changes varied more. In Wyoming, as long as the larger inhomogeneity window size was being considered, the highest hit rate was still found in scenario three for all algorithms with the exception of Climatol-Monthly. This was not because of a worse performance for Climatol-Monthly in scenario three, in fact it's hit rate

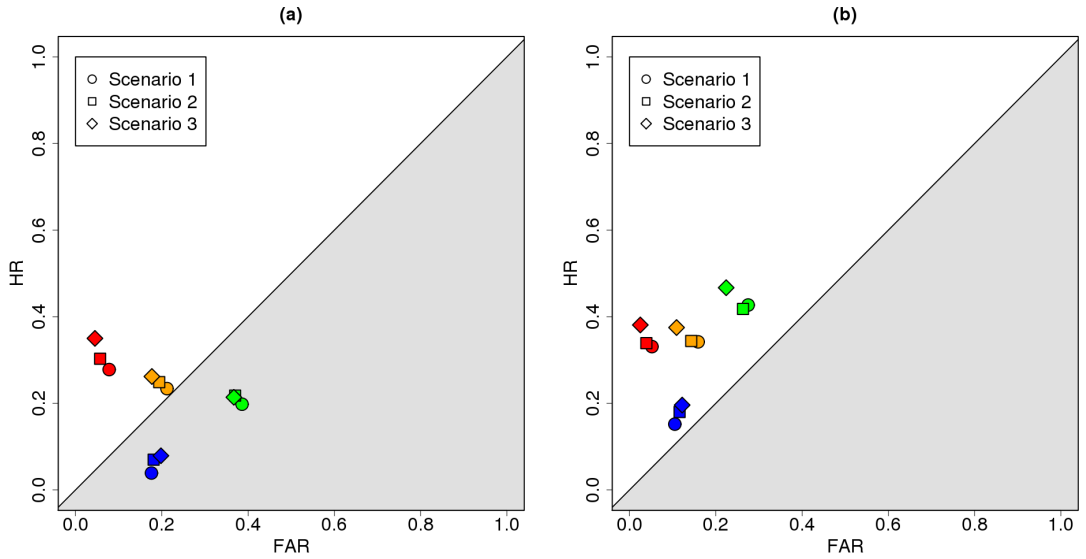

Figure 7.2. Plots showing false alarm rate against the hit rate in the South East for (a) a window extending thirty days either side of the true change point and (b) a window extending ninety days either side of the true change point. Climatol-Daily is red, Climatol-Monthly is orange, ACMANT is green and DAP, HOM and SPLIDHOM are blue. DAP, HOM and SPLIDHOM are all represented by a single blue point for each scenario because the detection approach was the same for all three of these algorithms and so the same change points were found. The shaded area below the diagonal line indicates the 'bad' area where the false alarm rate is higher than the hit rate.

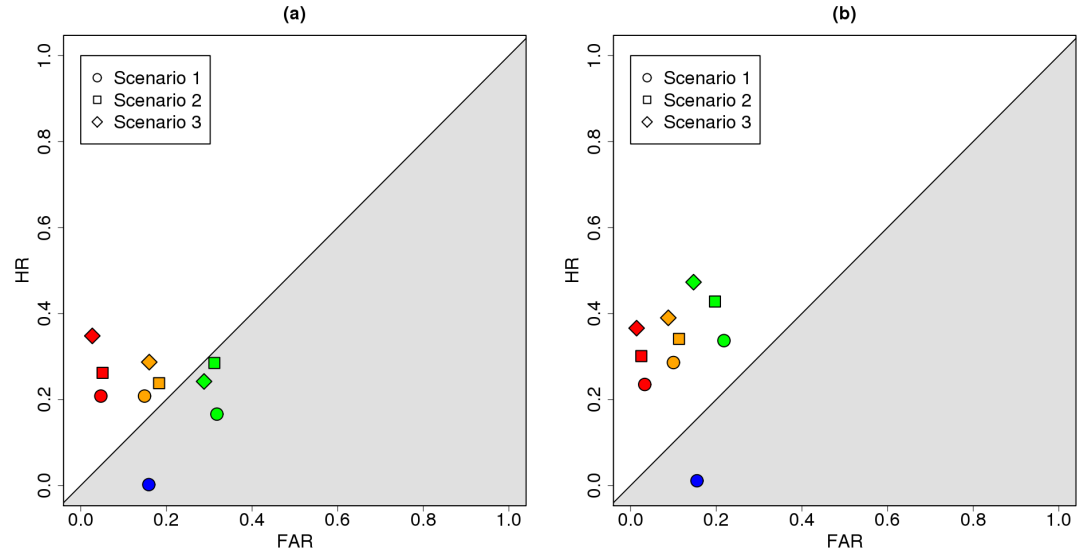

Figure 7.3. Plots showing false alarm rate against the hit rate in the North East for (a) a window extending thirty days either side of the true change point and (b) a window extending ninety days either side of the true change point. Climatol-Daily is red, Climatol-Monthly is orange, ACMANT is green and DAP, HOM and SPLIDHOM are blue. DAP, HOM and SPLIDHOM are all represented by a single blue point for scenario one because the detection approach was the same for all three of these algorithms and so the same change points were found. Scenarios two and three were not homogenised by DAP, HOM and SPLIDHOM in this region. The shaded area below the diagonal line indicates the 'bad' area where the false alarm rate is higher than the hit rate.

in this region was within the range of its hit rate for scenario three in other regions, instead there was a particularly good hit rate for Climatol-Monthly in Wyoming scenarios one and two. There is no clear reason why the hit rate was best for Climatol-Monthly in these scenarios. There was not a higher proportion of large inhomogeneities than in some other regions or scenarios, nor was there a higher proportion of constant offset

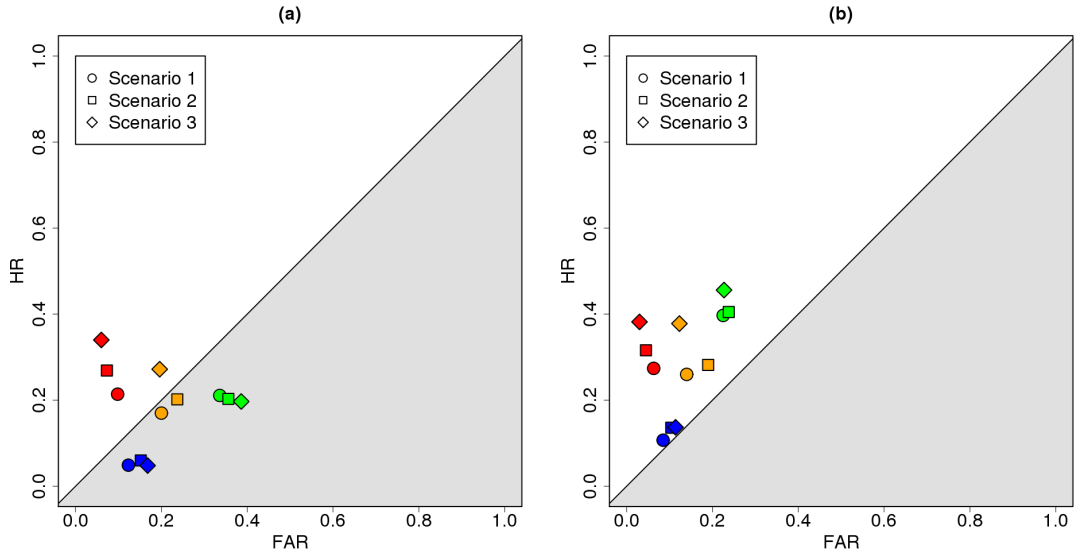

Figure 7.4. Plots showing false alarm rate against the hit rate in the South West scenarios for (a) a window extending thirty days either side of the true change point and (b) a window extending ninety days either side of the true change point. Climatol-Daily is red, Climatol-Monthly is orange, ACMANT is green and DAP, HOM and SPLIDHOM are blue. DAP, HOM and SPLIDHOM are all represented by a single blue point for each scenario because the detection approach was the same for all three of these algorithms and so the same change points were found. The shaded area below the diagonal line indicates the 'bad' area where the false alarm rate is higher than the hit rate.

inhomogeneities, which were normally easier to detect. The author would encourage further investigation into this area.

When the smaller window size was used Climatol-Daily and MAC-D kept their best performances in scenario three, as did Climatol-Monthly for all bar Wyoming, as expected from above. However, ACMANT and DAP, HOM and SPLIDHOM's performances were less consistent. As these algorithms were known to be less precise in their detection ability the author attributes the inconsistency to this cause, though further investigation to clarify this would be of interest.

Comparing across regions and scenarios there was no single region where all algorithms constantly had their highest hit rates. However, if a mean hit rate is taken within a region, that is all hit rates are summed and divided by the number of scenarios there are, then the highest average hit rate was always found in the South East, as long as this region was evaluated. This is as expected as the South East was arguably the simplest region given that it started off with a lower number of inhomogeneities per series because of many inhomogeneities being classified as unidentifiable. Given that unidentifiable inhomogeneities were more often explanatory variable inhomogeneities this meant that the South East also ended up with the greatest proportion of constant offset inhomogeneities. Constant offset inhomogeneities were overwhelmingly better detected than explanatory variable ones by all algorithms, likely because they are larger on average and less 'noisy' because their implemented size does not vary day to day.

There was no single region where the highest (worst) false alarm rate was consistently found. Looking at average false alarm rates across a region, in the same manner as

average hit rates were investigated, still reveals no consensus on the region most likely to have false alarms, though two algorithms do have most in the South East. The pairing of higher false alarm rates going with higher hit rates is not uncommon.

This pairing of higher false alarm rates and higher hit rates was also seen when comparing algorithms in general for the larger window size. ACMANT always had the highest false alarm rate of any algorithm in all regions and scenarios, but, with the exception of Wyoming scenario two, it also always had the highest hit rate. However, if the smaller window size was considered, Climatol-Daily nearly always had the highest hit rate and consistently had the lowest false alarm rate, thus making it the most precise algorithm in terms of detection. Given that Climatol-Daily retained the lowest false alarm rate even when the larger window size was used the author would commend this algorithm as the most reliable for inhomogeneity detection when false alarms are considered to have a high impact and would also recommend ACMANT when false alarms are of less concern.

The measure used in this study that takes into account hits and false alarms is the critical success index, see chapter six section 2.2. The higher this value is, the better an algorithm's detection ability is. Unsurprisingly, given the comments above, Climatol-Daily ranked top for this measure more frequently than any other algorithm did. When the smaller window size was being used, Climatol-Daily ranked top for all scenarios and regions apart from Wyoming scenarios one and two where it was beaten by its monthly counterpart. For the larger window size Climatol-Daily took the top spot four times, ACMANT took it seven times and Climatol-Monthly took it twice. This is further evidence to suggest that Climatol-Daily should be preferred when precision and a low tendency to insert false change points is desired, but ACMANT or Climatol-Monthly should not be ruled out if these two desires can be relaxed.

ACMANT should also be commended for its good ability to detect the smallest inhomogeneities, which all algorithms detected the lowest proportion of, but ACMANT least so. Climatol-Monthly and ACMANT were primarily the best algorithms for detecting urbanisation inhomogeneities too, though Climatol-Daily did also display a comparatively good performance in some scenarios and its detection ability was not as badly affected by the increased autocorrelations of scenario four as its monthly counterpart's was. In spite of this, urbanisation inhomogeneities were the least well detected overall, this is as expected given that none of the algorithms explicitly searched for this type of inhomogeneity.

The frequency bias, see chapter six section 2.2, was always below one for all regions, scenarios and algorithms apart from for MAC-D in scenario four. This means that, in nearly every case, algorithms are being too cautious about assigning change points. However, as will be explained in the following section, the algorithms that did make most changes were also those that made most stations worse in the returned data compared to the released data, implying that caution is not necessarily a bad thing.

In summary, ACMANT and Climatol-Daily are deemed to be the best algorithms for inhomogeneity detection overall because of their good hit rates, and, in Climatol-Daily's case,

low false alarm rate. Climatol-Monthly and MAC-D would not be considered bad at detection either, but the use of the detection methodology from DAP, HOM and SPLIDHOM would not be recommended. Urbanisation trend inhomogeneities are concluded to be the most difficult to locate and large inhomogeneities (greater than 1°C) are deemed to be the easiest inhomogeneities to locate. Constant offset inhomogeneities were better detected than explanatory variable inhomogeneities and this was linked to the fact that explanatory variable inhomogeneities were smaller. Scenario three, with no urbanisation inhomogeneities, was almost always the scenario that saw the best algorithm performance in terms of detection ability and scenario four, with increased autocorrelations, caused a deterioration in detection ability, though the extent of this deterioration varied between algorithms. Further summary information on the detection ability of each algorithm for each scenario is available to the reader in appendix B, where they should consult tables B.7, B.8, B.15, B.16, B.23, B.24, B.31, B.32, B.39, B.40, B.47, B.48, B.55 and B.56. All these tables were also made available to the homogenisers.

For algorithm improvement the author recommends an increased focus on working with autocorrelated data and further investigation into reliable methods to detect trend inhomogeneities and small or seasonally varying inhomogeneities. The author also recommends more algorithms that perform detection at the daily level and commends MAC-D and Climatol-Daily for their good detection abilities as truly daily algorithms.

For reference, section 4.2.2 stated that 'inhomogeneities' detected by the PHA in the clean series would not be counted as false alarms if other algorithms found them. This was to account for the fact that these could be genuine modelling errors. There was discussion about whether this gave an unfair advantage to 'PHA-like' algorithms. Having looked at validation outputs (not shown) very few of the same 'inhomogeneities' as those that the PHA identified were found by any algorithm. The algorithm most prone to finding the same 'inhomogeneities' was ACMANT, but this was also the algorithm most prone to false alarms in general and was no more 'PHA-like' than most of the other algorithms.

However, it should be noted that seven of the PHA's 'inhomogeneities' were found by at least two algorithms. Therefore, the author believes that PHA's 'inhomogeneities' were predominantly false alarms, but does not rule out the possibility of there being a few genuine change points in the created clean data. In a future iteration of the study the author would suggest not discounting false alarms if they were the same as the PHA's inhomogeneities as there is likely as much harm as good done by this approach. However, the author would recommend validation code that could tally if multiple algorithms found the same 'false alarms' so that users of the data could judge whether they wanted to try and adjust such a point or not.

7.2.4. Adjustment ability

Bias relative to the clean series

Very few stations in any region were left completely unaltered by the process of adding inhomogeneities. Regional biases were calculated as the average of the biases in the

region and varied in their sign. As expected, when looking at the sum of absolute biases, see 6.3.2, there was more bias present in the scenarios with a greater number of stations. Scenario three always had a greater sum of absolute biases than scenario two in the released scenarios. This was not surprising given that urbanisation inhomogeneities are often the main cause of small inhomogeneities and they were not present in scenario three. Looking at the percentage recovery of the sum of absolute biases, ACMANT was consistently among the top algorithms. Climatol-Daily also performed very well in general, being the top performing algorithm in all the South West scenarios.

There was predominantly a larger percentage recovery of the sum of absolute biases in scenario three, which would fit with this also being the scenario where the highest proportions of inhomogeneities were detected. In the South West and Wyoming for DAP, HOM and SPLIDHOM the largest reduction in the sum of absolute biases was instead seen in scenario two. This was likely owing to the fact that a greater proportion of large inhomogeneities were found by these algorithms in scenario two than in scenario three. The absolute sum of biases before and after homogenisation and the percentage recoveries that each algorithm achieved can be seen for each region and scenario in figure 7.5. The numerical values for the sum of absolute biases can be found, alongside other summary measures for the evaluation of bias reduction, in appendix B in tables B.1, B.9, B.17, B.25, B.33, B.41 and B.49.

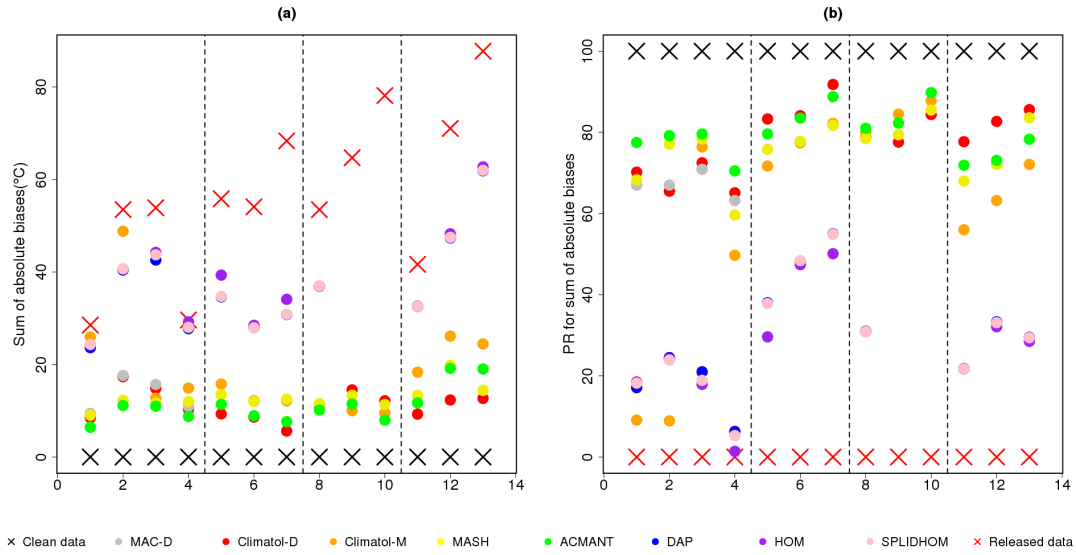

Figure 7.5. Plots to illustrate the reduction in the sum of absolute biases, relative to the clean benchmark data, for each algorithm, scenario and region. Plot (a) represents the reduction in the sum of absolute biases in °C and plot (b) shows the recovery as a percentage, with a 100% recovery being perfect and a 0% recovery meaning no change. X-axis labels are as follows; 1 - 4: Wyoming scenarios one to four; 5-7: South East scenarios one to three; 8-10: North East scenarios one to three; 11-13: South West scenarios one to three. Vertical dashed lines are added to help distinguish between these regions. Black crosses represent the clean data (always 0°C sum of absolute biases) and red crosses represent the released data relative to the clean benchmark data.

As well as tables B.1, B.9, B.17, B.25, B.33, B.41 and B.49, plots were produced that looked at bias reduction over time for each algorithm, scenario and region. Figure 7.6 contains an example of such a bias reduction plot for MAC-D for the Wyoming scenarios. As can be seen, bias is being reduced over time, but there are certain times the algorithm

is struggling with and reducing bias is evidently proving more difficult in scenario four than in the scenarios with less realistic autocorrelations.

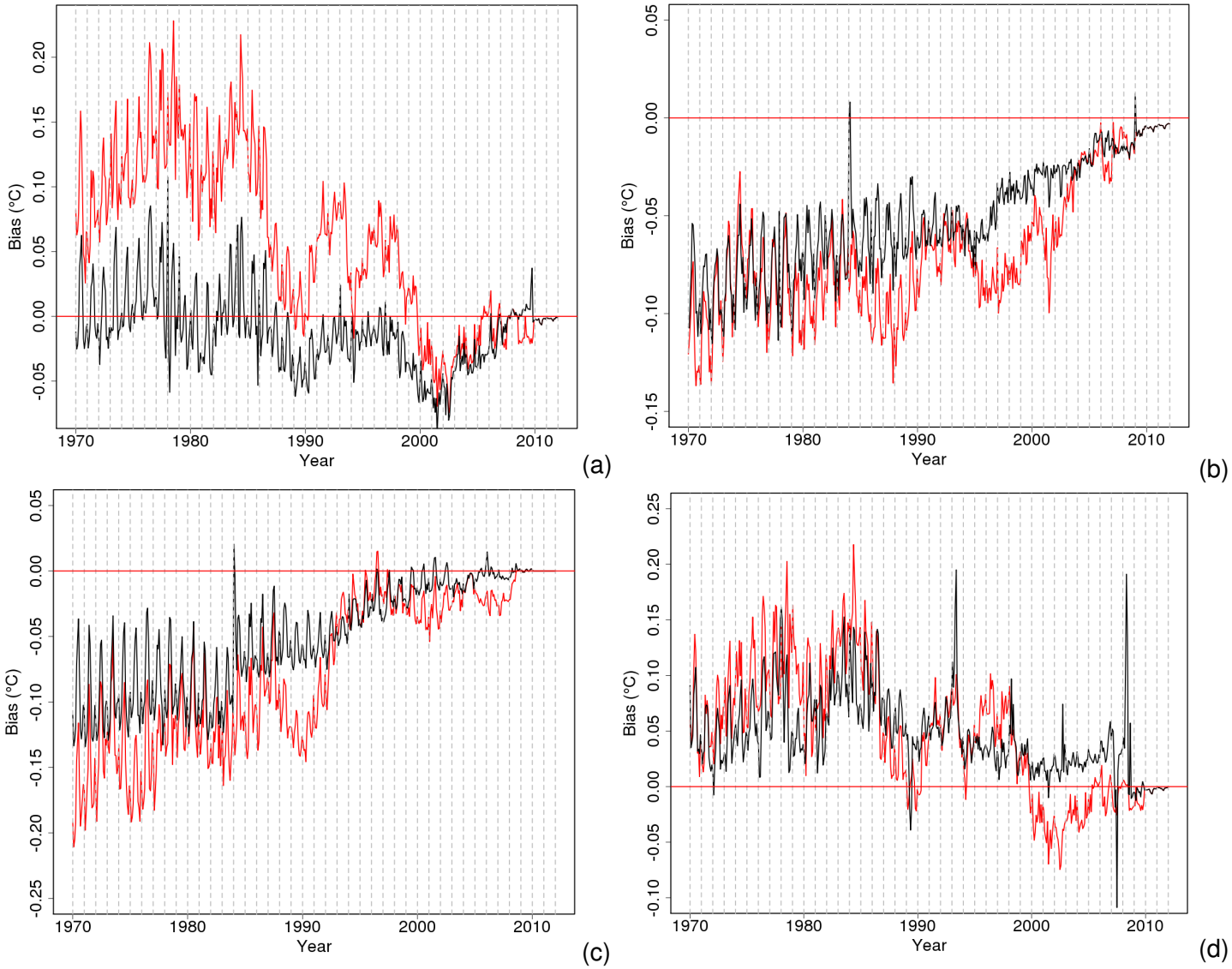

Figure 7.6. Plots to illustrate the progression of bias over time for Wyoming (a) scenario 1; (b) scenario 2; (c) scenario 3 and (d) scenario 4. Data have been aggregated to the monthly level to summarise the progression. Red lines indicate the released bias, relative to the clean benchmark data, and black lines indicate the returned bias, relative to the clean benchmark data, after MAC-D has been applied.

Best and worst stations were defined in terms of their RMSEs as was explained in section 3.2 of chapter 6. The above mentioned tables display how many of the biases for these best and worst stations each algorithm improved or worsened during the homogenisation process. Generally speaking, the best stations were largely left unaltered by homogenisation and the worst stations were improved. However, MASH did display the greatest tendency to make the best stations worse and ACMANT displayed some tendency to do the same, though not as often. The majority of algorithms made at least half of the best stations more biased during the homogenisation process of Wyoming scenario four, again illustrating the increased difficulty that the more autocorrelated data presented.

Station summary bias plots, as in figure 7.7 are a good visual way of assessing the

reduction in bias on a station by station level. It can be seen in these plots that the homogenisation process has reduced both the mean and variability (assessed using standard deviations) of the biases, both desirable outcomes. Although these plots are for the MAC-D algorithm and Wyoming scenarios, similarly good results can be seen from other methods and in other scenarios in general.

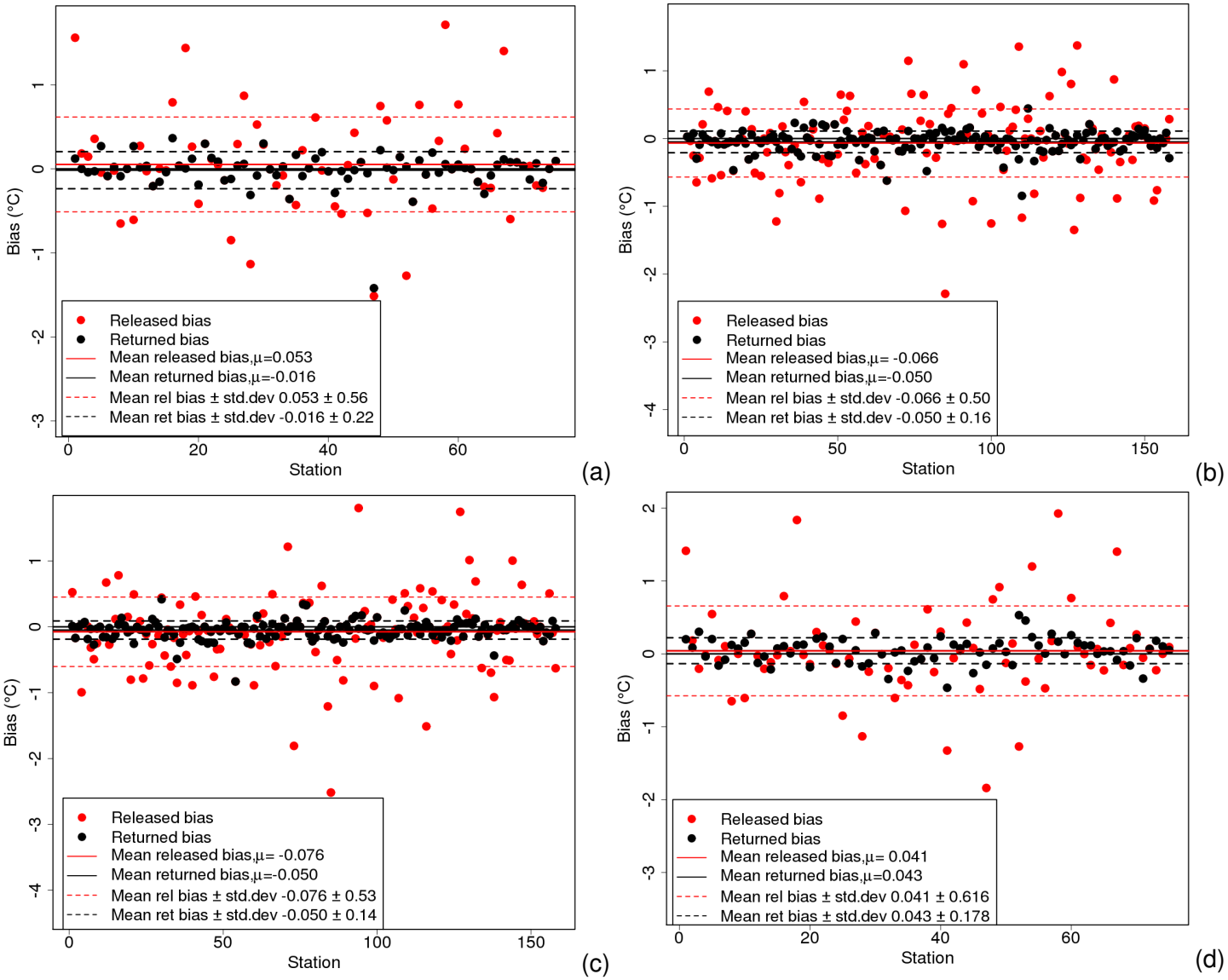

Figure 7.7. Plots to illustrate the bias, relative to the clean benchmark data, of each station in Wyoming before homogenisation (red) and after homogenisation by MAC-D (black), for (a) scenario 1; (b) scenario 2; (c) scenario 3 and (d) scenario 4.

There are inevitably some stations where the bias is made worse by the homogenisation process for all algorithms, scenarios and regions; although it is not necessarily always the same stations that each algorithm makes worse. Climatol-Daily consistently displays one of the lowest, and often the very lowest, tendency to make stations worse and MASH and ACMANT largely display the highest. These proportions of stations being made worse are related to the proportion that are left unaltered by the homogenisation process. Climatol-Daily does not alter as many stations as MASH or ACMANT, but as a result of homogenising more stations the latter two algorithms also display a greater tendency to make station biases worse. Therefore, the author recommends that if the reader seeks

an algorithm with a low rate of increasing station biases then they should use Climatol-Daily. If a greater overall bias reduction is desirable, even at the cost of some stations having lower quality after homogenisation has been applied, then an algorithm such as ACMANT could also be used. However, Climatol-Daily outperforms ACMANT even for overall bias reduction in all three South East scenarios, all three South West scenarios and Wyoming scenario four. DAP, HOM and SPLIDHOM are even more conservative than Climatol-Daily in altering station biases, leaving a greater number unchanged, but also making a similar or slightly greater proportion worse in general. Climatol-Monthly and MAC-D would both be classed as average for bias reduction, making more stations worse in general than Climatol-Daily, but not as many worse in general as ACMANT, and making more stations better than Climatol-Daily, but again, not as many as ACMANT.

There is very little tendency across the algorithms to increase the number of positively biased stations relative to perfection, but there is a noticeable lean towards returning more negatively biased stations than there were on release, though the magnitude of the biases is predominantly reduced. ACMANT displays the greatest tendency to return stations with a negative bias suggesting that caution should be applied when using stations homogenised by ACMANT for long term trend analysis.

RMSE of the returned data relative to the clean data

Similar methods were employed to evaluate an algorithm's ability to reduce RMSE as for its ability to reduce bias. RMSE is of interest in addition to bias because it can be shown to consider differences in correlations and standard deviations as well as differences in means, see chapter six section 3.2 for the formula illustrating this. Plots such as in figure 7.8 were produced for RMSE and, as expected, the RMSE reduced as the present day, which was predominantly used as an algorithm's reference period, was approached. There were spikes of lower algorithm performance exhibited in the RMSE plots over time, and these appear to be correlated with times of increased variability in temperatures between stations. Given that the algorithms applied in this study predominantly used other station series to determine the size of an adjustment, it is logical that one finds lower performance when there is increased variability as comparisons between stations are less beneficial.

The values and percentage recoveries for regional RMSEs are illustrated in figure 7.9 and can be found alongside other statistics in tables B.2, B.10, B.18, B.26, B.34, B.42, B.50 of appendix B. Climatol-Daily and ACMANT were once more the best performing algorithms in terms of percentage recovery in Wyoming, the South East and the South West, although MASH did outperform ACMANT in the South West scenario three. ACMANT was the best at regional RMSE reduction for the North East scenarios one and three, but Climatol-Monthly outperformed it in the North East scenario two. Neither Climatol-Monthly nor MASH were as consistently good as the former two algorithms though. DAP, HOM and SPLIDHOM largely lagged behind other algorithms owing to leaving a larger number of stations unchanged by the homogenisation process. MAC-D performed relatively well in the area of regional RMSE reduction for the scenarios that it sought to homogenise. The greatest consistency in RMSE reduction across scenarios was found

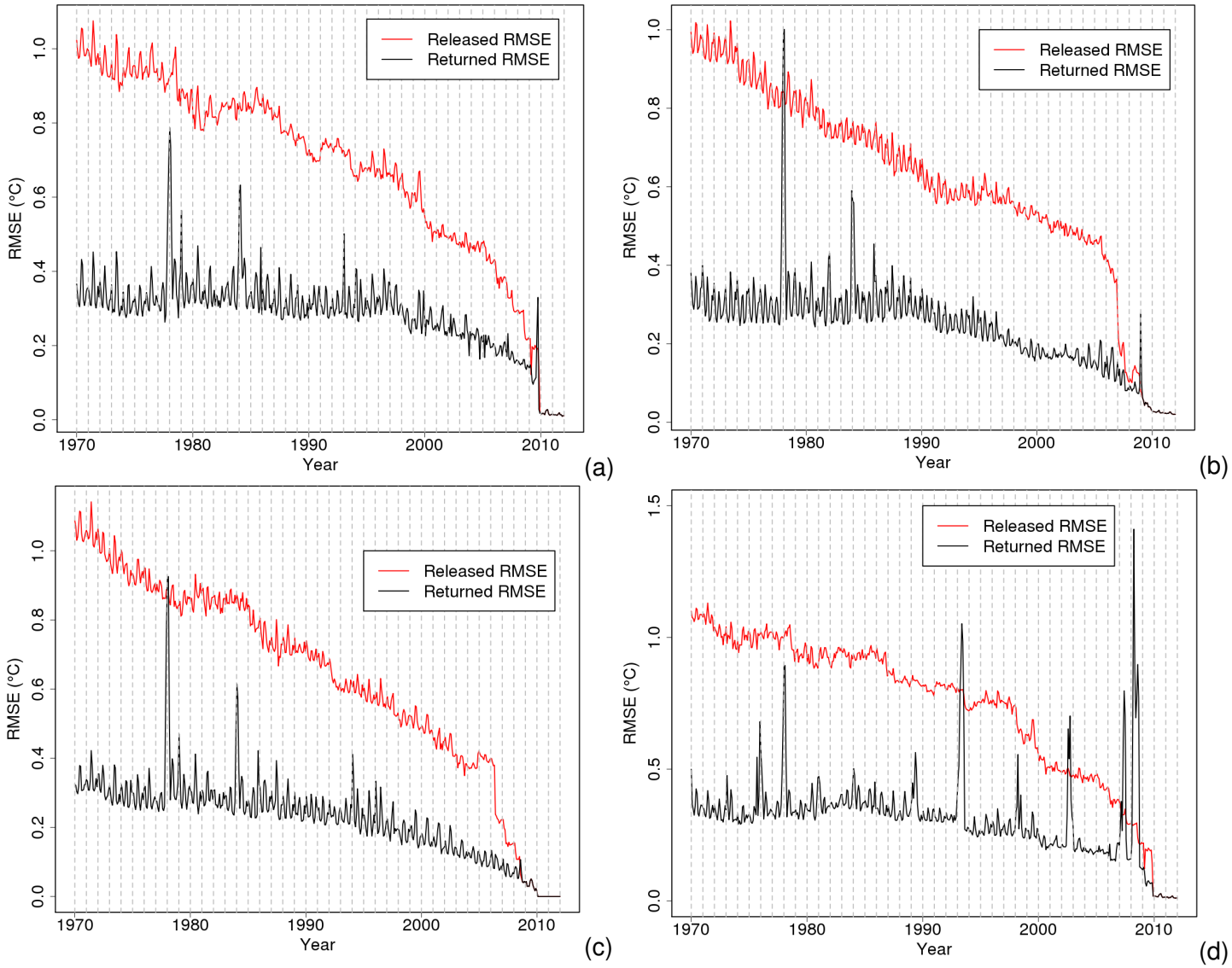

Figure 7.8. Plots to illustrate the progression of RMSE over time for Wyoming (a) scenario 1; (b) scenario 2; (c) scenario 3 and (d) scenario 4. Data have been aggregated to the monthly level to summarise the progression. Red lines indicate the released bias, relative to the clean benchmark data, and black lines indicate the returned bias, relative to the clean benchmark data, after MAC-D has been applied.

in the North East, where performance of the majority of algorithms was good.

Looking on a station by station basis the findings are similar for RMSE as they were for bias. Climatol-Daily still displays the lowest tendency to make stations worse, but also leaves a large number unchanged. ACMANT and MASH improve a larger proportion of stations, but do so at the expense of making a non-negligible number of station RMSEs worse. These tendencies are true across all scenarios and all regions. MASH consistently makes the homogeneity of all the best stations worse, but predominantly makes the homogeneity of all the worst stations better. This general tendency to improve the homogeneity of the worst stations is true for all algorithms across all scenarios.

Conclusions based on RMSE assessment measures are very similar to those drawn from bias assessment measures. Climatol-Daily is the preferred algorithm if a low tendency to make stations worse is sought. If a good overall performance is desired and the ho-

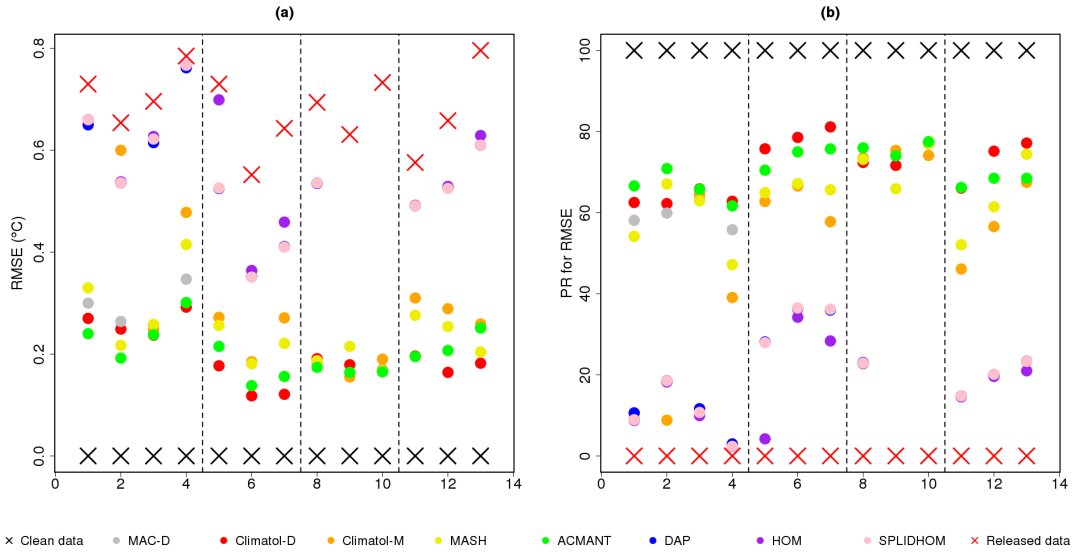

Figure 7.9. Plots to illustrate the reduction in the regional RMSE, relative to the clean benchmark data, by each algorithm for each scenario and region. Plot (a) represents the reduction in RMSE in °C and plot (b) shows the recovery as a percentage, with a 100% recovery being perfect and a 0% recovery meaning no change. X-axis labels are as follows; 1 - 4: Wyoming scenarios one to four; 5-7: South East scenarios one to three; 8-10: North East scenarios one to three; 11-13: South West scenarios one to three. Vertical dashed lines have been added to help distinguish between these regions. Black crosses represent the clean data (always a RMSE of 0°C) and red crosses represent the released data relative to the clean benchmark data.

mogeniser is willing to sacrifice the homogeneity of some good stations for the improvement of many stations then ACMANT would also be a reliable choice.

Linear Trend Recovery

As stated in chapter 6 section 3.2, long term trends are a major area of concern in climate science. The focus here is on the overall long-term linear trends as the assessments using loess smooths to investigate high and low frequency variability were found to be misleading in some cases, as discussed in section four of chapter six. The output of the loess analysis can still be found in tables B.4, B.12, B.20, B.28, B.36, B.44 and B.52 of appendix B should the reader be interested, but, as with all validation measures, these statistics should not be taken out of context.

Most interest lies in linear trends that are considered significant at some level. Trends here were defined as being significant if their regression coefficients were significant at the 5% level. In Wyoming and the South East relatively few trends in any scenario were significant at this level for the clean data, though noticeably larger proportions were significant in the released data. In the North East the vast majority of trends were significant for the clean data and there were fewer significant trends in the released data. For the South West just under half of the trends were significant in the clean data and this proportion was increased for the released data for all scenarios. In the following summary when the phrase 'truly significant trends' is used it is referring to trends that were significant in the clean data.

In Wyoming, MASH and ACMANT did the best job of returning a similar number of significant trends in the returned data as were found in the clean data, but these were not

necessarily at the stations that did have truly significant trends. MASH struggled to retain the significance of the truly significant trends in these scenarios and ACMANT also struggled slightly, but to a lesser extent. No algorithms returned significant trend values for the two truly significant trends in Wyoming scenario four, but most did a good job of reducing the number of spurious significant trends that were only present because of the inhomogeneities. Climatol-Monthly performed comparatively to MASH and ACMANT in Wyoming scenario three and relatively well in the South East scenarios. Climatol-Daily and MAC-D retained the significance and approximate value of the only significant trends in Wyoming scenarios one and two, but not in scenario three. In the South East, Climatol-Daily performed on a par with MASH and ACMANT, with all three algorithms showing good recovery of the significance and approximate values of the truly significant trends. For the North East, Climatol-Daily, Climatol-Monthly, MASH and ACMANT all did a good job of returning similar numbers of truly significant trends for all three scenarios. In terms of recovering coefficient values of the significant trends Climatol-Monthly and ACMANT were the top performing algorithms. In the South West, MASH, ACMANT, Climatol-Daily and Climatol-Monthly continued to perform well. In terms of returning significant trends that were at the same stations as in the clean data and also close in value, MASH and Climatol-Daily performed best in the South West scenarios, though all algorithms returned too many significant trends in the South West scenario two.

In all Wyoming, South East and South West scenarios DAP, HOM and SPLIDHOM constantly returned too many significant trends, largely as a result of not altering enough stations in the homogenisation process. In the North East, DAP, HOM and SPLIDHOM were only applied to scenario one, but here they returned too few significant trends, again likely as a result of not making enough changes during homogenisation.

There were no significant regional trends in any of the scenarios for Wyoming and the South East and this lack of significant trends was preserved in the released series and the series returned by all the homogenisation algorithms. In the North East, regional trends were significant in the clean, released and returned data for all scenarios. In the South West, no regional trends were significant in the clean data, but the regional trend in scenario two was made significant on release and no algorithm managed to remove this significance. MASH wrongly made the regional trend significant in scenario three of the South West, which is undesirable. The reason for the significance of trends in the North East and not the other regions is not immediately clear. However, investigation into regional linear trends from the observations that were used when creating the benchmark data shows that these too do not have a significant regional trend in Wyoming or the South East, but they do in the North East. This suggests that the model is reproducing true climate artefacts in these regions. In the South West, the observations show a significant linear trend that is not reproduced in the created data, this illustrates that this region may not be as reliable or realistic as a benchmark as the other regions. Although, it is also possible that the trend exhibited in the observations is from inhomogeneities and is not a true climate trend. It is not unsurprising that some non-significant trends are found. Globally our climate is warming, [Karl et al., 2015], but regionally, and over shorter time periods, the trends can be less pronounced.

Algorithms generally struggled with regional trend recovery most when the clean and released trends were very similar in value. This was shown in Wyoming scenario two where percentage trend recoveries were very small and in the South East scenario two, where percentage trend recoveries were much too large. In the South East scenario two, the trends were in fact the same in the clean and released data to three decimal places, which leads to a weakness in percentage trend recovery being highlighted. This weakness is that if statistics are similar in clean and released data then the denominator of the percentage recovery expression is incredibly small meaning that very large percentage trend recovery values are quickly obtained, even if an algorithm's performance is still reasonable. When such a situation arises it is helpful to look at a visual representation of trend recovery in terms of the clean, released and returned °C/decade trend in addition to the percentage trend recovery value. Visualisations of trends for the clean, released and returned data are therefore shown in figure 7.10 alongside the percentage recovery values for these trends. The numerical values of the trends can also be seen in appendix B in tables B.3, B.11, B.19, B.27, B.35, B.43 and B.51.

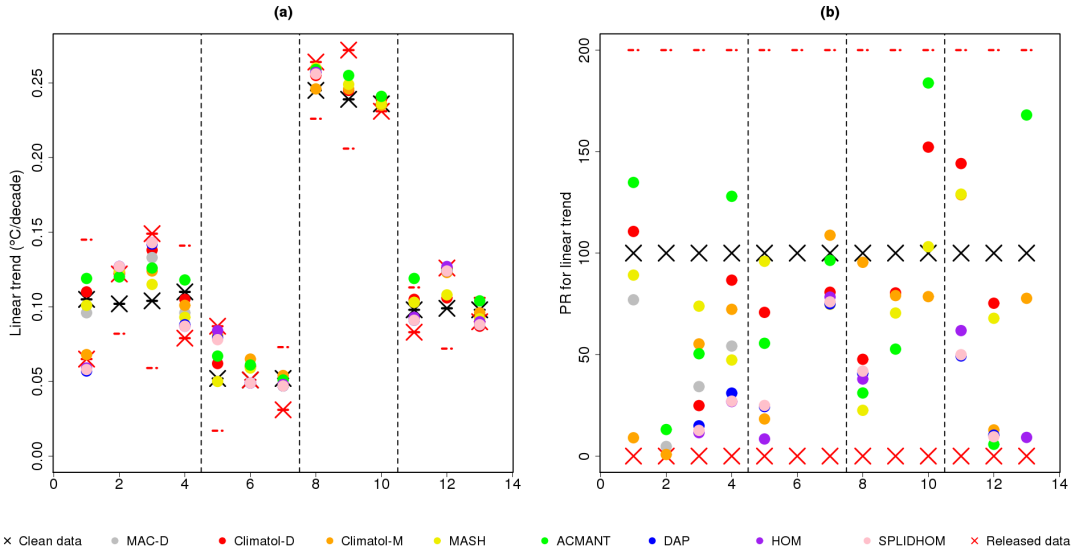

Figure 7.10. Plots to illustrate the recovery of regional linear trends, relative to the clean benchmark data, by each algorithm for each scenario and region. Plot (a) represents the recovery of the linear trend in °C/decade and plot (b) shows the recovery as a percentage. The Y-axis in plot (b) has been restricted to only show percentage recovery values between 0% and 200%, that is, only values that display no change or some change for the better. Therefore if certain algorithms aren't represented for a particular scenario in plot (b) this means that the algorithm returned a regional linear trend that was more dissimilar to the true regional linear trend than it was on release. X-axis labels are as follows; 1 - 4: Wyoming scenarios one to four; 5-7: South East scenarios one to three; 8-10: North East scenarios one to three; 11-13: South West scenarios one to three. Vertical dashed lines have been added to help distinguish between these regions. Black crosses represent the clean data, red crosses represent the released data and red dashes represent the 200% recovery point, beyond which trends have been moved in the right direction, but to such an extent that they are now more dissimilar to the clean trend than they were on release.

Comparing across scenarios and regions there is no one algorithm that is consistently best or worst at regional trend recovery. MASH is nearly always in the top three, but ACMANT, Climatol-Daily and Climatol-Monthly also have occasions of being the best. For the North East where the regional trends are significant there is still no one consistently best algorithm.

There were always more positive than negative trends in the clean data for all scenarios and regions. The addition of inhomogeneities added more negative trends in, and, in general the process of homogenisation then corrected the data sufficiently to make these trends positive once more, though occasionally the balance was redressed too far and not enough negative trends were returned. Interestingly, ACMANT did not return enough negative trends in the majority of the Wyoming, South East and South West scenarios, in spite of its tendency to return negatively biased stations. This suggests that ACMANT may be biasing stations more negatively towards the beginning of series, which would then lead to an increased likelihood of creating positive trends, though the author has not investigated this theory. In the North East ACMANT displayed no such problem of returning too many stations with positive trends, though all bar one of the trends in this region were positive in the clean data anyway.

In summary, no one algorithm stands out as being the best for trend recovery, though Climatol-Daily still shows the lowest tendency to make stations worse, which is a desirable quality. MASH and ACMANT are still improving the greatest number of stations, but at the cost of making more worse as well. Climatol-Monthly should be approached with caution as it can have a relatively high tendency to make station trends worse, though this is not consistent across all regions, but primarily evidenced in the South East, where trends are smallest on average.

Variability similarity assessed by standard deviation comparisons between clean, released and returned data

Recovering clean station variabilities is the area that algorithms found most difficult. This is not a surprising result as most were not designed to homogenise moments higher than the mean. The three algorithms that were designed for the homogenisation of higher order moments, DAP, HOM and SPLIDHOM, still struggled in this area; they were changing station variabilities, but they were not necessarily changing them for the better.

On release there were more stations that were too variable than not variable enough. In general all algorithms kept this balance. However, this does not mean that the same stations fell in the same categories of 'too variable' and 'too uniform' in the released and returned data. That is to say, the majority of stations did have their variabilities changed by the homogenisation process and for Climatol-Daily, Climatol-Monthly, ACMANT and MAC-D this always resulted in more station variabilities being improved than made worse. For DAP, HOM and SPLIDHOM there were always more variabilities made worse than made better by homogenisation. For MASH more variabilities were made worse than better in all Wyoming scenarios, but only scenario one of the North East and scenario two of the South West. In the summary document provided to homogenisers, plots, as in figure 7.11, were included that allowed the quick evaluation of whether there was a tendency to return stations that were too variable or too uniform. Figure 7.11 is for the Wyoming scenarios for MAC-D where it can be seen that variabilities are being brought closer to the observed variabilities in general, with little tendency to consistently return too variable or too uniform stations. Further information on algorithm variability recovery can be found in appendix B, in tables B.5, B.13, B.21, B.29, B.37, B.45 and B.53.

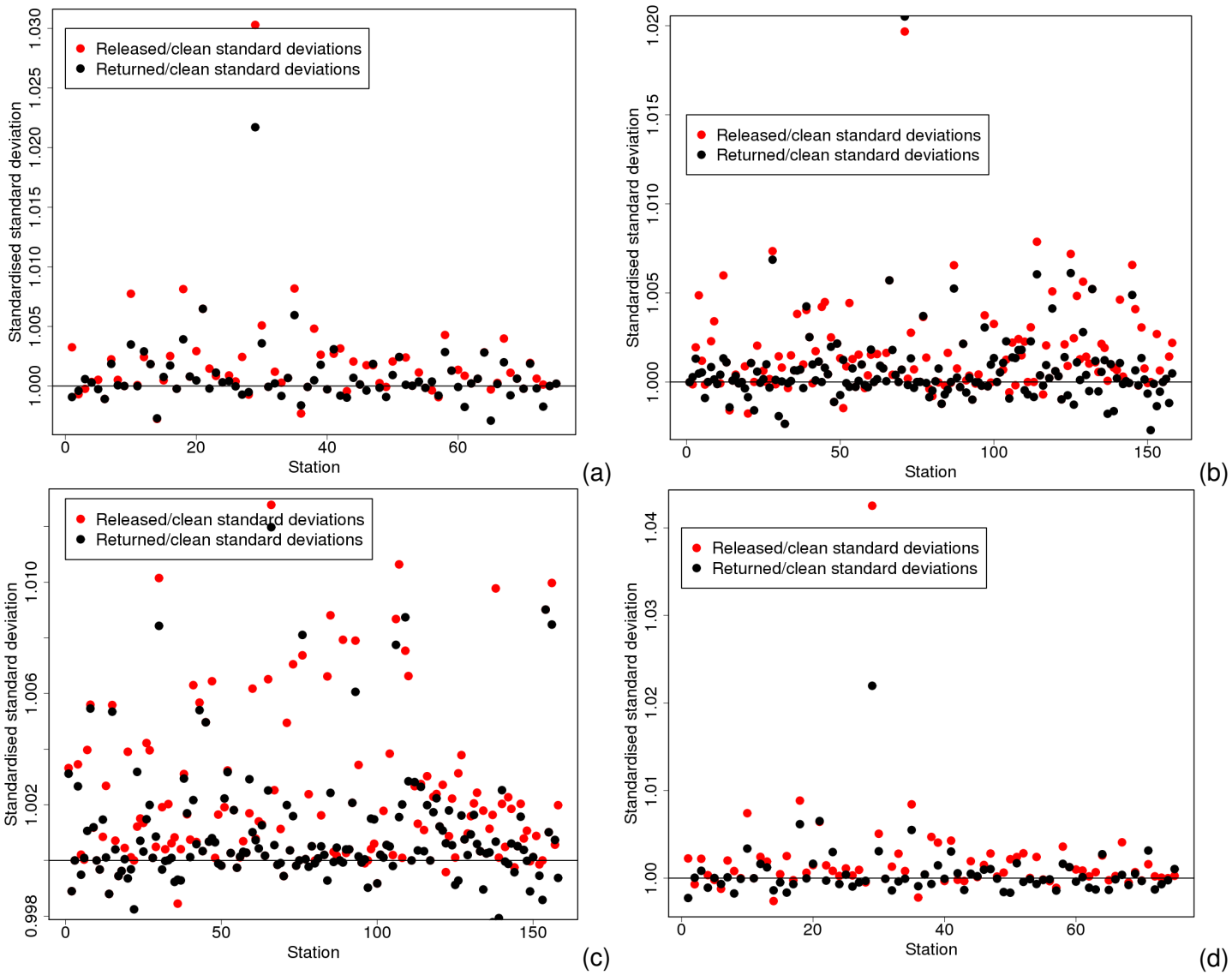

Figure 7.11. Plots to illustrate the ratios of released to clean (red) and returned to clean, for MAC-D, (black) standard deviations for each station in Wyoming (a) scenario 1; (b) scenario 2; (c) scenario 3 and (d) scenario 4.

Out of MAC-D, Climatol-Daily, Climatol-Monthly and ACMANT, ACMANT is best for all regions and scenarios apart from the North East scenario two where it comes second to Climatol-Monhtly. For this reason the author recommends ACMANT as the best algorithm for improving station variabilities with the cautionary note that it does also make a non-negligible number of variabilities worse.

Extreme Value Recovery

At the daily level extremes are not smoothed out by an aggregation process, which is desirable as extremes have impacts felt at the societal scale. However, because extremes are single observations, measurement uncertainty must be taken into account. If returned extremes matched the extremes in the clean data exactly then they were referred to as being 'exact', if they matched the extremes in the clean data within ± 0.14°C then they were referred to as being 'exact to measurement precision'. The value of 0.14°C was obtained from [Brohan et al., 2006] and its justification is provided in chapter 6 section 3.2.

Tables B.6, B.14, B.22, B.30, B.38, B.46 and B.54 provide information on each algorithm's ability to recover extremes and the following paragraphs draw on information from these tables for their conclusions.

Without exception, more of the cold extremes were corrupted by the process of adding inhomogeneities than the warm extremes and in general they were made cooler than in the clean data. The reason for this phenomenon is that more of the cold extremes occurred earlier in the time series, where there was a greater chance of an inhomogeneity being in progress owing to inhomogeneities propagating backwards. The slightly greater tendency to colder extremes can be explained by the fact that inhomogeneities that resulted in a warming for the more recent time period were implemented by cooling the less recent time period, and warming inhomogeneities were fractionally more common. The algorithm that was best at recovering cold extremes to be exact was Climatol-Daily for most of the Wyoming and all of the South East and South West scenarios and also the North East scenario three. Climatol-Monthly was best at recovering cold extremes exactly for the North East scenarios one and two. ACMANT was best at recovering cold extremes exactly for Wyoming scenario one and best at recovering them exact to measurement precision for all Wyoming scenarios and the North East scenarios one and three and the South West scenario three. MASH was best at recovering cold extremes to measurement precision in the South West scenarios one and two.

For warm extremes, ACMANT was best at recovering them exactly for Wyoming scenarios one to three, the South East scenario two and the North East scenario three. It was also best at recovering the warm extremes exact to measurement precision for all the South East scenarios, all the North East scenarios and Wyoming scenarios one to three. For Wyoming scenario four, DAP, HOM and SPLIDHOM were best at returning warm extremes exactly as in the clean data and exact to measurement precision, though this was achieved by just not changing any extremes from the released data. Climatol-Daily was best at recovering extremes exactly for the South East scenarios one and three, best or joint best for all the South West scenarios and joint best for the North East scenario two. Climatol-Daily was also best at recovering warm extremes exact to measurement precision for the South East scenario one and all three South West scenarios. Climatol-Monthly was joint best for recovering extremes exactly for the North East scenario two and the South West scenario one and was best for the North East scenario one.

In terms of not making extremes worse there is not a single algorithm that does best in all regions and scenarios. However, for cold extremes Climatol-Daily is primarily the algorithm least prone to making extremes worse and for warm extremes Climatol-Daily also shows a low tendency to make extremes worse, but so do DAP, HOM and SPLIDHOM. MASH was the algorithm that most commonly improved extremes from their values in the released data, but not by enough to make them exact to measurement precision. However, MASH was also the algorithm which made most extremes worse and should therefore be used with caution. It should also be noted though that some extremes MASH 'made worse' were owing to different reference periods being used in the homogenisation process.

Overall the conclusions are the same for this assessment measure as for the majority of others; Climatol-Daily and ACMANT would be the authors recommended algorithms for extreme value recovery, but the majority of other algorithms should also be considered relatively suitable for this task. This is an encouraging result when the majority of algorithms were not specifically designed to work with daily data.

7.3. Summary of algorithm performance

The conclusion of this work has to be, as expected, that different algorithms have different strengths and weaknesses. Therefore, someone with data to homogenise should clarify which aspects of their data they are most interested in before they choose a homogenisation method. The following subsections attempt to give an overview of the strengths and weaknesses of each of the algorithms that were contributed to this study to allow an informed decision on a fit for purpose algorithm to be made. The reader should bear in mind that these statements are true when the algorithms are applied to the created benchmark data, but the author believes the same characteristics would be exhibited when the methods are applied to real world data. An interesting extension to this study would be to apply these algorithms to the real world data the benchmarks were created from. This would allow quantification of whether similar features appear to be exhibited in the observations as in the created benchmarks. For example, whether similar numbers of inhomogeneities were found by these algorithms in the observations and the benchmarks.

7.3.1. ACMANT

ACMANT is considered by the author to be one of the best homogenisation algorithms contributed to this study. ACMANT consistently made a large number of stations better during the homogenisation process. Although, this came at the cost of making a non-negligible number of stations worse as well, often including stations that were perfect on release. Its ability to recover variabilities and extreme values is commendable across most regions, though it does struggle more in the South West. The precision of the detections made by this algorithm could be improved as a larger window size was necessary for this algorithm to have a good detection ability in general and even then it displayed a high false alarm rate. However, its detection ability is praiseworthy because of the consistently higher rate of detection for small inhomogeneities than that found for other algorithms. The increase in station density does not cause a consistent change in performance in ACMANT, in Wyoming and the South West little improvement is made, but in the South East and North East algorithm performance is generally better in scenario two than scenario one. The impact of having no trend inhomogeneities in scenario three relative to scenario two results in better detection ability in general, but has little affect on adjustment ability. ACMANT's adjustment ability for scenario four in the presence of increased autocorrelations is commendable with little degradation in performance being

shown. The detection ability of ACMANT in scenario four is not as good as in scenario one suggesting that this could be considered an area of improvement for this algorithm.

7.3.2. Climatol-Daily

Alongside ACMANT, Climatol-Daily is considered by the author to be one of the best algorithms. It shows little tendency to make stations worse during the homogenisation process, although this comes at the cost of leaving a non-negligible number of stations unchanged, which may be considered undesirable for some applications. It shows a good and precise detection ability and is to be highly commended for its adjustment ability too, which homogenised some stations to perfection, a trait that no other algorithm can boast. Generally speaking Climatol-Daily's performance was not affected by a change in station density, though this is not true in the South West where more stations did lead to better algorithm performance. In the South West, the presence of trend inhomogeneities did not lead to a consistent change in algorithm performance, but in other regions there was some evidence of trend inhomogeneities hindering Climatol-Daily's ability to homogenise as well as it could. The presence of autocorrelations did not have a detrimental effect on Climatol-Daily's adjustment ability, which is praiseworthy, though its detection ability for scenario four in Wyoming was lower than scenario one, which was the equivalent scenario with lower autocorrelations.

7.3.3. Climatol-Monthly

Climatol-Monthly's performance can best be described as being average. It does not show as much of a tendency to leave stations unchanged as its daily counterpart, but it does show more of a tendency to make them worse, though rarely the greatest tendency to do so. Generally speaking the increase in the station density from scenario one to scenario two does lead to a slight improvement in algorithm performance, but not to such an extent that Climatol-Monthly should be considered unreliable when used to homogenise a smaller station network. Trend inhomogeneities lead to a lower detection ability for Climatol-Monthly, though they are understandably harder to detect. The adjustment ability of Climatol-Monthly is generally not affected by the presence of trend inhomogeneities, although its performance is worse in the South West scenario two than scenario three. Given that scenario one of Wyoming was not correctly homogenised by this algorithm it is difficult to quantify the effect of autocorrelations on its adjustment performance. However, detection ability is degraded by the presence of autocorrelations and the adjustment measures that could be compared also suggest that this is an area this algorithm struggles with.

7.3.4. DAP, HOM and SPLIDHOM

These algorithms generally do not make a large proportion of stations worse according to most validation measures, which is commendable. However, they also suffer from a low detection ability meaning that many stations are left completely unchanged by the homogenisation process. Owing to this the author recommends that the primary area for development for these algorithms is their detection methodology. Looking at the number of stations that are changed, there are always a greater proportion improved than made worse for all three of these algorithms for nearly all adjustment ability measures, which is commendable. However, for station variabilities that are changed by these algorithms, the vast majority are made worse and this was more often due to station variabilities being wrongly increased than wrongly decreased. Therefore, this should be considered a necessary area for algorithm improvement before these algorithms are run on any data where higher order moments are of interest. The detection ability itself was hindered by largely assigning change points to the first day of the year, this meant that with the smaller window the false alarm rate was always higher than the hit rate, though this was generally not true for the larger window size. The increase in station density from scenario one to scenario two does not have a consistent effect, it improves hit rates, but also increases false alarm rates; the adjustment ability is improved for Wyoming and the South West, but not the South East and couldn't be assessed in the North East owing to only scenario one being homogenised. The absence of trend inhomogeneities in scenario three relative to scenario two does not lead to a consistent change in algorithm performance in general, which is commendable. Autocorrelations on the other hand do noticeably affect the performance of these algorithms with fewer inhomogeneities being detected when the autocorrelations in the underlying data were increased. Working with autocorrelated data could therefore be considered another area of improvement for these algorithms.

7.3.5. MAC-D

Overall MAC-D is ranked as being average with reference to most algorithm assessment measures. It is rarely best or worst for any scenario or measure, but instead displays a consistency that is commendable if the characteristics of the data to be homogenised are unsure. Its performance is not unduly changed by changes in station density or by the presence or lack of artificial trend inhomogeneities. Its ability to recover station variabilities is one of the best of all algorithms analysed, which is praiseworthy when it is not specifically designed to homogenise moments higher than the mean. MAC-D's performance is degraded by the presence of autocorrelations in the data, but this is realised in more stations being made worse, not fewer being improved. This algorithm is not affected substantially by a change in window size for inhomogeneity detection making its accuracy in locating change points commendable.

7.3.6. MASH

The performance of MASH could be classified differently depending on what the aims of the homogenisation exercise are. MASH is commonly the algorithm most prone to reducing station biases, RMSEs and trends, but this comes at the cost of consistently making a non-negligible number of stations worse during the homogenisation process. It rarely leaves any stations unchanged by homogenisation and this means that even perfect stations are commonly corrupted by this algorithm. Looking at significant trends, its ability to remove spurious significance present only because of inhomogeneities is commendable. MASH does not perform well at recovering clean station variabilities, being the most prone to making stations too variable on return for all bar the South East scenario one and the South West scenario two. It struggles with extreme value recovery as well, rarely being best in this area of evaluation, and being worst for all Wyoming scenarios. Looking at the difference in scenarios MASH should be commended for not having a performance overly dependent on underlying data characteristics. The increase in station density from scenario one to scenario two caused no noticeable change in algorithm performance for most regions, though the performance was better for scenario two in Wyoming. The presence of trend inhomogeneities did not lead to a consistent change in algorithm performance. The presence of autocorrelations in Wyoming scenario four does lead to a slight degradation in algorithm performance, but not as substantial as for some other algorithms. Therefore, if the underlying characteristics of data to be homogenised are unknown and the homogeniser is prepared to sacrifice the homogeneity of some stations for the improvement of others MASH could be considered a suitable candidate for the effort.

7.4. Uncertainty remaining after homogenisation

An advantage of a benchmarking study with synthetic data is that the truth is completely known beforehand, and thus, the uncertainty remaining in the data after homogenisation can be quantified. In this study this was assessed by looking at the bias and RMSE for a region between clean and returned data after homogenisation relative to bias and RMSE for a region between clean and released data before homogenisation. It was also assessed by comparing regional trends before and after homogenisation. The reason for looking at regional data is that climate studies are often carried out with a focus bigger than a single station. Therefore, although single stations may be improved or degraded by homogenisation the overall impact on climate conclusions is likely to be at the regional level. A further reason for focusing on the regional level is that information on a station by station level has already been provided above to aid homogenisers in their algorithm development and this section now aims to aid climate scientists in their climate change conclusions.

7.4.1. Regional Bias and RMSE

RMSE and bias allow the overall similarity of behaviour between clean and returned series' to be assessed. Bias is advantageous as it reports the differences in means between the clean and released and clean and returned regions, which is of interest if a change in mean temperatures over time is being assessed and the uncertainty in that change is being sought. However, for bias, the effects of multiple badly homogenised stations can cancel out and for this reason regional RMSE, which must be positive, was also investigated to allow the quantification of the magnitude of errors remaining in a series.

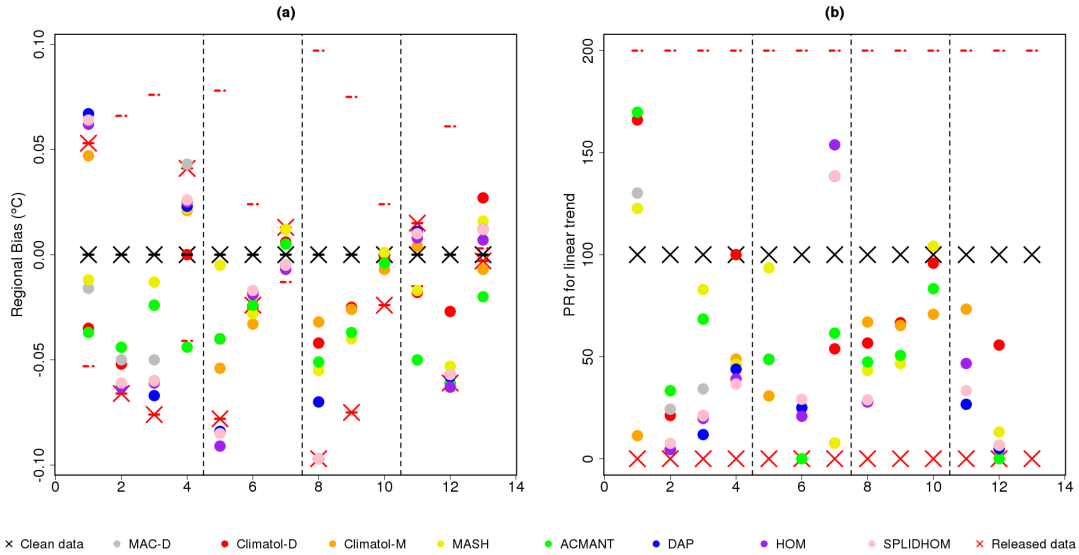

Figure 7.12. Plots to illustrate the reduction in regional bias, relative to the clean benchmark data, by each algorithm for each scenario and region. Plot (a) represents the change in regional bias in °C and plot (b) shows the recovery as a percentage, with a 100% recovery being perfect and a 0% recovery meaning no change. The Y-axis in plot (b) has been restricted to only show percentage recovery values between 0% and 200%, that is, only values that display no change or some change for the better. Therefore if certain algorithms aren't represented for a particular scenario in plot (b) this means that the algorithm returned a regional bias that was more dissimilar to the true regional bias than it was on release. X-axis labels are as follows; 1 - 4: Wyoming scenarios one to four; 5-7: South East scenarios one to three; 8-10: North East scenarios one to three; 11-13: South West scenarios one to three. Vertical dashed lines have been added to help distinguish between these regions. Points outside the red lines indicate an algorithm has returned a regional bias larger than it was on release. Black crosses represent the clean data, red crosses represent the released data and red dashes represent the 200% recovery point, beyond which regional biases have been moved in the right direction, but to such an extent that the regional bias is now greater than on release.

The mean bias for a region is simply calculated as the mean of the individual station biases. The biases across the regions and scenarios varied in magnitude, but were never larger than 0.1°C as can be seen in figure 7.12. It was not atypical for some regional mean biases to be increased during homogenisation; this is why some coloured points do not appear in 7.12 (b). This happened in Wyoming scenario one for DAP, HOM and SPLIDHOM; MACD and ACMANT (which also changed the sign of the bias) in Wyoming scenario four; DAP, HOM and SPLIDHOM for the South East scenario one; Climatol-Daily, Climatol-Monthly and MASH for the South East scenario two; Climatol-Daily, MASH and ACMANT for the South West scenario one, where the sign of the bias was also changed; HOM in the South West scenario two and all algorithms in the South West scenario three,

where the bias only started off as -0.003°C. However, in only one case, ACMANT for the South West scenario one, was the bias increased enough that its value would change from 0°C to 0.1°C when reporting to the closest tenth of a degree.

The regional RMSE is calculated as $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(Clean_i - Released_i)^2}{n}}$ where n is the number of data points in the region for which the record is not missing and $Clean_i$ and $Released_i$ (or $Returned_i$) are the mean values on day i. This means that all stations are effectively put into one long string of time points before this calculation takes place. Regional released RMSEs were always between 0.55°C and 0.80°C, with no pattern in where the largest regional RMSEs were found. The regional RMSEs were never increased, therefore it can be concluded that the uncertainty due to inhomogeneities in the data is, in general, lower after homogenisation than before. For the returned regional RMSEs, values ranged from 0.10°C to 0.77°C. Climatol-Daily and ACMANT were the best at reducing regional RMSEs, with values below 0.3°C for the data these algorithms returned in all bar two cases. HOM was the worst algorithm for reducing regional RMSE in general. Therefore, in terms of quantifying remaining uncertainty, when using the best algorithms the uncertainty for RMSE at least can be assumed to have been reduced by 50% in most cases, this can be seen in figure 7.9.

7.4.2. Regional Trends

Enabling the calculation of reliable climate trends is one of the primary reasons for homogenisation. Of all the regions and scenarios studied it was only in the North East where significant trends were found in the clean data and these were between 0.236 and 0.245°C/decade in magnitude. The data corruption process made this range increase to 0.231 to 0.272°C/decade, but without exception all algorithms brought the regional trends closer to the clean trends during homogenisation in this region, thus reducing the uncertainty remaining in the data. For the South West scenario two, the corruption process made the regional trend significant and no algorithm managed to remove this significance, but Climatol-Daily came closest, reducing the trend from 0.126°C/decade to 0.106°C/decade, though this is still 0.007°C/decade larger than the true regional trend. The largest difference in trends between clean and released data was 0.045°C/ decade found in Wyoming scenario three, this was made smaller by all homogenisation algorithms. The biggest difference between a clean and returned trend is found in Wyoming scenario one for DAP and is 0.048°C/ decade, but this is the only time a discrepancy of this size is found. In no case is the sign of a regional trend ever changed by homogenisation as was illustrated in figure 7.10. In general, trends are improved by homogenisation and therefore the uncertainty in regional trends can be said to be reduced. However, uncertainty in the significance of trends remains as the South West scenario two illustrates that just because all algorithms return a significant trend this does not mean that the underlying clean data exhibits a significant trend.

7.5. Discussion

This thesis was able to assess the eight contributed homogenisation algorithms according to the measures defined in chapter six. This allowed the quantification of their performance relative to the created clean data, which is beneficial as there was a known answer. However, as already stated, the created data are not a perfect replication of the real world. The author would therefore recommend that an extension to this study be formed by applying the homogenisation algorithms to the real world data from which the created data were formed. The evaluation framework would have to be altered, as eluded to in section three of this chapter, but conclusions could still be drawn from the results of applying the algorithms to the real data by comparisons with the results from applying them to the benchmarks. For example, if a similar number of small, medium and large inhomogeneities were found in the real world data compared to the number found in the benchmark data, it would be reasonable to suppose that a similar proportion had also been missed. If it could be established that a similar proportion had been missed, then the uncertainty remaining in the real world data after homogenisation may then also be supposed to be similar to that remaining in the benchmark data after homogenisation.

If a similar study to this were to be formed that looked only at certain algorithms then more bespoke comparisons could be made. For example, performance could be compared between the 'good' and 'bad' stations for MAC-D and validation measures could be adapted to cope with series that were not homogenised to the most recent period for MASH. Neither of these extensions were carried out here as the evaluation sought to compare all algorithms as much as possible instead of highlighting specific ones.

Two further extensions to this study if it were to be carried out again would relate to greater uncertainty quantification. The first would help to quantify uncertainty in the quality of the benchmarks by keeping a tally of inhomogeneities found by different algorithms. If all algorithms found the same supposedly false alarm then it would help the benchmark creator to know that there may be a bug in the code. Equally multiple algorithms getting the same hit would help to further quantify which inhomogeneities are easiest to find. The second extension would be a lot more time consuming as it is the suggestion that an ensemble of benchmarks be produced with the same underlying properties, but slightly different inhomogeneities in each. For example, there could be 100 Wyoming scenario ones, all with the three inhomogeneity types and the same number of stations, but with a new set of inhomogeneities generated each time. This ensemble approach would allow uncertainty estimates to be placed on the detection measures. Adjustment measure uncertainties would still be difficult to quantify, but uncertainties could be given for improvements in regional biases, RMSEs and trends. For example, it may be possible to state that a hypothetical algorithm 'A' always reduces RMSE by at least 50% whereas algorithm 'B' sometimes reduces it by 90%, but other times only by 10%.

The overlap of algorithms with the study of Venema et al. [2012] is ACMANT, Climatol-monthly and MASH, though it is not necessarily the exact same version of the algorithm that was run in this study and Venema et al. [2012]. In Venema et al. [2012] detection

ability was carried out at the annual level and both ACMANT and Climatol performed well. The hit rates exhibited in this thesis are a little lower than those found in Venema et al. [2012] which is to be expected because of the greater time restriction employed here for a detection to count as a hit. Also, the false alarm rates are higher in this study, which is once more to be expected. In terms of adjustment ability, when Venema et al. [2012] looked at the CRMSE between clean and released and clean and returned data they commended the performance of both ACMANT and MASH and showed plots which suggest that neither of these algorithms showed such a tendency to make stations worse during homogenisation at the monthly level as they did at the daily level. Although Venema et al. [2012] do note that when looking at network wide CRMSE ACMANT does not perform as well and lags behind MASH, something not exhibited in this thesis. No comment on Climatol's performance with respect to temperature series CRMSEs could be found. Looking at linear trend comparisons Venema et al. [2012] found that Climatol had a tendency to greatly decrease the magnitude of any trend in temperature series; it still reduced trend magnitudes in this thesis, but here it was right to do so as the corruption process of the data predominantly made trends larger than they were in the clean data. MASH outperformed ACMANT in trend recovery in Venema et al. [2012], though both did well. In this thesis MASH and ACMANT still both performed well in trend recovery and MASH still outperformed ACMANT in terms of the number of station trends improved in all bar Wyoming scenario one. However, MASH did also make more trends worse than ACMANT did. The similarities and differences between these benchmarking studies shows the benefit of multiple homogenisation studies; they reinforce some information, but challenge other pieces of information, they also show that at different temporal (and spatial) scales algorithms may perform differently.

7.6. Summary

This chapter has used the validation framework laid out in chapter six to evaluate the performance of the eight homogenisation algorithms contributed to this blind benchmarking study. The results have been encouraging in that the algorithms do improve the homogeneity of stations in general. However, the evaluation has also shown that there is to date, as anticipated, no algorithm that is completely reliable when seeking to homogenise daily temperature data. The two algorithms that stand out as having the best performance are ACMANT and Climatol-Daily, though both of these have weaknesses too.

Areas for algorithm improvement that have been identified by this study are: a greater capacity to deal with autocorrelated data; improved detection ability for small, seasonally varying or gradual inhomogeneities; and a need to homogenise moments higher than the mean. The following chapter will provide the authors suggestions for areas of future work and a summary of the conclusions from this and each of the previous chapters.

8. Conclusions and Future Work

This study has investigated the performance of homogenisation algorithms on daily temperature data. Chapters one and two gave the motivations behind such a study and an overview of previous work in the areas of homogenisation and benchmarking. Chapters three and four introduced how realistic clean artificial data can be created, detailing the method employed in this study. Chapter five introduced the inhomogeneity structures that were added on to these clean data, the reasons for the specific inhomogeneity choices and the creation of four scenarios to investigate the impacts of different data and station network characteristics on homogenisation algorithm performance. Chapter six explained the validation framework to evaluate algorithm performance in this work. The results of implementing this framework on the eight homogenisation algorithms contributed to this study were presented in chapter seven, where summaries of their performances were also given.

This final chapter draws on the conclusions from the preceding chapters to highlight the achievements of this study. It also draws on the conclusions and discussions from previous chapters to give an outline of where the author recommends future work in this area should be directed and what should be done differently if the study were to be repeated.

8.1. Conclusions

This project has been the first comparison study of homogenisation algorithms on daily temperature data. It was necessary as previous studies have focused on monthly or annual data and have largely looked at smaller station networks than those which were developed here for four regions in North America.

This project benefited from using a modelling approach that allowed the incorporation of other climatic variables into the creation of daily temperature time series. The GAM is able to model artefacts such as seasonal cycles and long term trends that in the past have commonly had to be removed before temperature data could be modelled. The GAM used here is also good because, as its inputs come from reanalysis data and location variables, it can create stations where none previously existed and is not hindered by missing data. Although the focus in this research was on North America, the methodology of using a GAM to produce synthetic daily temperature series could be generalised to other regions of the globe. However, if using the GAM for other regions of the globe, the input variables should be reviewed so as to incorporate those most pertinent to the study region.

Another advantage of using a GAM is that realistic inhomogeneities can be created by perturbing the inputs to the model. This can create seasonally varying inhomogeneities and inhomogeneities which are dependent on other climatic variables, both these types of inhomogeneity are known to be observed in reality. As well as this, the GAM does not prevent the addition of constant offset inhomogeneities, allowing a consistency between this study and previous homogenisation studies. The power of the GAM was exploited to create four different inhomogeneity scenarios, three exploring both step and trend changes and one exploring only step changes. Different station densities could also be investigated because of the GAM's ability to create extra stations, and series autocorrelations were explored in scenario four that was created only for Wyoming.

Overall the change in station density was not found to impact algorithm performance to the same extent that changes in autocorrelations were. The absence of trend inhomogeneities was also not found to impact algorithm performance to the same extent as autocorrelations. However, detection ability was predominantly better in the absence of trend inhomogeneities, as trend inhomogeneities usually fell into the category of 'small' inhomogeneities, that were consistently less well detected.

The validation framework incorporated assessments for both algorithm detection ability, which has been the primary focus of most studies in the past, and algorithm adjustment ability. It is important that these two aspects are viewed as complementary and not competitive as good performance in one aspect does not guarantee good performance in the other.

ACMANT is commended as being the best algorithm for detecting small inhomogeneities, though for ACMANT to have a good detection ability in any region or scenario the larger of the two detection windows was needed, and even then a high false alarm rate was exhibited. Climatol-Daily should also be commended for a good detection ability in general, and for this algorithm the precision of these detections is praiseworthy with a smaller change point window size being sufficient. Overall, these two algorithms would be upheld as the best contributions to this study, for both detection and adjustment ability.

Climatol-Daily homogenised some stations to perfection, which no other algorithm succeeded in doing. It also consistently made very few stations worse, though it left a non-negligible number unchanged. ACMANT on the other hand improved a great number of stations, but made a non-negligible number worse. MASH also improved the homogeneity of a great number of stations, but left very few unchanged, meaning that even perfect stations were commonly corrupted by this algorithm. DAP, HOM and SPLIDHOM had the opposite problem and left too many stations completely unchanged. MAC-D was only applied in Wyoming, but here displayed average performance, with a good number of stations improved, though also a non-negligible number made worse. Climatol-Monthly's performance could also be described as average as it too showed a good ability to improve stations, but also made a non-negligible number of them worse.

As anticipated, the conclusions in the area of algorithm improvement are that more work is required to create or adapt homogenisation algorithms to cope with autocorrelated

data, small and seasonally varying inhomogeneities and trend inhomogeneities.

With these conclusions, the author believes the objectives of this study have been met. A suite of realistic benchmark datasets have been created and different inhomogeneity structures have been explored. The homogenisation community were successfully engaged and feedback on algorithms was provided to each participating homogeniser as well as now being made available to the homogenisation community as a whole. Areas for future work are identified below, but the author believes this study to have been an incredibly beneficial first study into the performance of algorithms on daily temperature data.

8.2. Discussion and Future Work

The above section gave areas for improvement of algorithms and this section details the areas for improvement and extension of the benchmarks and validation measures. The first of these areas correlates well with the need for algorithms that can cope with autocorrelated data, in that there is a need for better modelling of autocorrelated data. Scenario four of this study created data that had better autocorrelations than the rest of the scenarios, but they were still not completely realistic. Autocorrelations really need to match those found in observations for deseasonalised difference series, as this is the level at which inhomogeneities are commonly sought out. However, even in scenario four, autocorrelations tailed off too soon in these series.

A further area of improvement for the benchmarks would be to create more realistic inter-station correlations. These were, in general, too high in the created data. However, the highest inter-station correlations were reduced when scenario four was created. This suggests that in a future iteration of this study it would be possible to simultaneously improve autocorrelations and inter-station correlations. One possible way of doing this would be to use a spatio-temporal model on top of the GAM. That is, use the GAM model as it is to produce the mean predictions, but then create a model for the perturbations from these means that incorporates the spatial and temporal patterns. This could, for example, be carried out in R's Spatio Temporal package using the differences between the predicted means and observations as inputs to the model, along with any necessary covariates, such as the time and location variables.

A way of expanding this benchmarking study using either the existing model formulation or the new spatio-temporal formulation suggested above would be to look at other regions of the globe. The four regions chosen in North America did have different climates, but they did not explore all climates exhibited on the Earth. The author would recommend investigating regions such as Europe, where the results could be more directly compared to Venema et al. [2012] and also at least one region that has very little seasonal cycle or very predictable weather patterns to see what effects these aspects have on both model and algorithm capabilities. As stated in chapter four and in section one of this chapter, if different areas of the globe were to be considered, the variables included in the model

should be reviewed to ensure that the best model for each region is produced.

Even with the existing modelled regions it would be possible to expand the study by looking at expanding the scenarios. For example, similar released scenarios could be produced, but with changes to the underlying data. Such changes could include incorporating different variables or data from different reanalyses, downscaling from reanalyses in a different way or exploring variants of the existing smooth functions, for example by changing their degrees of freedom or basis functions. These changes could be carried out individually or in varying combinations.

As well as the possible extension from changing the underlying models, to create more scenarios from different clean data, this study could also be extended by looking at changes to the released scenarios. Chapter five suggested the creation of a metadata scenario. In the present study such a scenario was not created as others were deemed to be higher priority. However, seeing how much algorithms could use metadata, even if only as a validation of the change points they find, would be an interesting extension to this work. If the use of such data led to an increase in performance then it would further highlight the need for the recovery and digitisation of old station metadata and the reliable recording of new station metadata.

Existing scenarios could also be extended or new ones created by looking at a wider range of inhomogeneities or a change in the size distribution of inhomogeneities. In the present study constant offset inhomogeneities had their sizes chosen from a discrete distribution and explanatory variables had their perturbations chosen from a discrete distribution as well. In a future study both of these distributions could be made continuous. As the discussion section in chapter five states, having a continuous distribution, such as a Normal distribution, for the constant offset inhomogeneities' sizes would likely make them more similar to the sizes of explanatory variable inhomogeneities, thus making comparisons between inhomogeneities added in the two different ways simpler. Specifying a continuous distribution for urbanisation inhomogeneity sizes as a trend per year instead of an overall trend would remove the need for the jump in allowable sizes at the urbanisation length of fifteen years, which is not a very realistic artefact of the current study.

Still relating to urbanisation inhomogeneities, a future iteration of this study would ensure that the bug which made all explanatory variable urbanisation inhomogeneities begin with a step change is removed. Still in the area of inhomogeneity addition, it was found that platform inhomogeneities, those that act over a short period and are then corrected, were not frequent enough. Adding in more platform inhomogeneities could make the created benchmarks more realistic, as could adding clustered inhomogeneities, which affect multiple stations at similar times in a similar manner. A further inhomogeneity that it would be beneficial to investigate would be negative trend inhomogeneities. Negative trend inhomogeneities would be more likely to mask a true climate signal, so assessing algorithm performance on the removal of such artificial trends would be beneficial.

As stated in chapter seven, a very interesting extension to this study would be to apply the contributed homogenisation algorithms to the real world data as well. This would be

beneficial as comparisons between the results from applying the homogenisation algorithms to the real world data and the benchmark data could lead to better quantification of the likely inhomogeneities and uncertainties remaining in the observations after homogenisation has taken place.

For the validation aspect of the study the author would recommend the comparison of algorithm performance on long and short term variability recovery. This was designed to have been implemented in this study by using loess smooths and then assessing the correlations between these smooths between clean, released and returned data. Correlations were found to have the possibility of being misleading as a measure, because they could be higher for released data, that were more dissimilar to the clean data, than for the improved returned data and, therefore, this aspect of validation was not discussed in the study, though the results can still be seen in tables B.4, B.12, B.20, B.28, B.36, B.44 and B.52. The author believes that the loess smooths could still be used by looking at the RMSE between these smooths and not the correlations. RMSE removes the drawback that was found with correlations and could be used in conjunction with the RMSE between the clean, released and returned series at the daily level. At the daily level the RMSE gives an error between two series for the day to day variability, but when using the RMSE on loess smooths it would give an error between the lower frequency variability, as desired.

A further extension to the validation aspect of this study would be to provide some measure of uncertainty on the statistics returned that quantify algorithm performance. This was not done in this study as the suggested manner of including uncertainties would require that all algorithms be run multiple times on a created ensemble of the existing scenarios. Such an ensemble would have, for example, 100 realisations of each of the existing thirteen released station groups. The benefit of having an ensemble would be that the hit rate, false alarm rate, regional trend recovery etc. could be recorded for each algorithm for each ensemble member. The range that these values took could then be used to provide uncertainty bounds on the measure in question. A variation of this was done in chapter seven of this study where reference was given to how the same algorithm performed across different regions and scenarios. However, to be able to provide uncertainty estimates in each scenario for each region would give the algorithm creators even more information on the reliability of their algorithm. The code that was used to create the inhomogeneities is available on request and therefore there is no reason why an ensemble of the existing inhomogeneous data scenarios shouldn't be created in the future.

Even without an ensemble of the same scenarios and regions, quantification of some benchmark uncertainties can be provided. Uncertainties in the benchmarks arise from whether the clean data really are clean. In the discussion section of chapter seven it was suggested that a tally could be kept of inhomogeneities found by all the participating algorithms. If such a tally were to be kept then it could be used to see if multiple algorithms were finding the same 'false alarms' and also whether they were finding the same hits. The former would indicate possible benchmark problems, whereas the latter would give

information on the easiest inhomogeneities to find.

Something a little like this method was carried out using the PHA, where 'inhomogeneities' that the PHA found in the clean data were not counted as false alarms in the returned data. It could be argued that if only one other algorithm finds the same 'inhomogeneity' as the PHA then they are just sharing a false alarm, however, if multiple algorithms start sharing false alarms then further investigation would be warranted.

The PHA should, of course, not be held up as the perfect algorithm. This study showed that it had a notable false alarm rate when inhomogeneities had to be found within a month of their true date, though Venema et al. [2012] found a low false alarm rate when the detection simply had to be in the right year. If this study were to be carried out again the author would therefore suggest that the inhomogeneity tallying method be carried out as well as, or even instead of the PHA comparison when searching for any possible problems in the clean data.

The different findings from Venema et al. [2012] about the PHA and about some points of other algorithms used in this thesis shows that only two benchmarking studies is not enough. More work needs to be done to continue the process of assessing and improving homogenisation algorithms. This thesis is but one step in this direction and it is the author's hope that this discussion section, which draws on those from previous chapters, has provided ideas for further steps to continue this important work.

8.3. Summary

Overall the author concludes that this study has been beneficial as the first comparison study of algorithm performance on daily temperature benchmarks. It shows that there are certainly differences between the currently available algorithms and has identified areas for improvement in each of these algorithms and for algorithms in general. It has also evaluated the created benchmark data and provided recommendations for improving this in a future research project. All the data used for this project are freely available at `http://www.metoffice.gov.uk/hadobs/benchmarks/`, where electronic appendices giving more detailed information on the performance of each algorithm can also be found. All the code used in this project is also available on request.

Appendices

A. Instructions to Homogenisers

This appendix contains the main emails that were sent to homogenisers when inviting their participation in this study and then giving them further instructions on how to participate including the location of the data.

A.1. Early October - Invitation to participate

Dear Homogeniser,

You may have met me at the Budapest workshop on homogenisation (May 2014), but if not let me introduce myself. I am a PhD student at the University of Exeter working on creating benchmarks for the homogenisation of daily surface temperature data. I would like to involve as many people as possible in this work and am therefore looking for daily homogenisation algorithm developers and users who may be able to test their chosen algorithm on my data.

Ideally I would like a number of different algorithms to be tested on my data to maximise the use of this blind study. I would like the homogenised version of the data to be returned along with details of the size, timing and type (if possible) of any adjustments made. The results of this work will form a part of my PhD thesis and therefore I am subject to time constraints and would be grateful if you were able to return your results to me by the end of November. If this target date would stop you participating then please do get in touch.

The benchmark series I am creating are grouped into three 'worlds' each of which covers the same four regions in North America and these regions range in size from 75 to 230 stations. If you would be interested in participating please respond to this email and I will ensure that you are informed when the data are released (target 10th October 2014). The data will be mimicking GHCN data (ASCII format), but will not contain any quality flags as they have been created in such a manner that users should consider them to be pre-quality controlled. These data are investigating different inhomogeneity characteristics, but I am also hoping for a slightly later (end of October 2014) and smaller release that will look at different climate characteristics of data.

Details of the benchmark creation, validation of algorithm performance and assessment of the usefulness of my benchmarks will not only form my thesis, but I hope a number of other publications in the coming years. Your participation in this process will lead to an acknowledgement or co-authorship of any relevant publications and should you wish

to collaborate more and co-author further papers together I would be interested to hear from you on this matter.

Please feel free to forward this message on to anyone else you think may be interested in this work and do not hesitate to contact me if you have any queries about the process.

I have attached a presentation similar to the one I delivered in Budapest to give you a further overview of my research.

I look forward to hearing from you soon,

Thank you in advance,

Rachel Warren

NB – The presentation referred to in this email is not included as part of this appendix, but detailed the data sources (GHCND, 20CR and the Australian Bureau of Meteorology); the inhomogeneities focused on (station relocations, shelter changes and urbanisation); the number of worlds (three at the time, though this was later increased to be four); the notion of adding in inhomogeneities by changing model inputs or using constant offsets and the fact that both detection and adjustment ability would be assessed.

A.2. Mid October 2014 - Further instructions on participation

Dear Homogenisers,

Thank you for the interest that you have shown in running your homogenisation algorithms on my data and for the further comments you have made to me regarding this project.

I am afraid the final stage of creating the benchmarks has taken a little longer than anticipated and therefore they will now not be ready until the end of next week/ the start of the following one (24th/27th October). As this is a delay of two weeks in the release date I will of course change the date when I would request you return your results to me by, this is therefore now **Friday 12th December**.

In the meantime I hope that the following information will aid you in any preparations you may need to make for your algorithms:

There are three different realisations of the benchmarks - world 1, world 2 and world 3.

Each world contains four regions of North America, these regions are classified as:

1. Wyoming: Latitude: (41,45) and Longitude: (-111,-104.2), Number of stations in world 1= 75, Number of stations is worlds 2 and 3= 158 2. South East: Latitude: (25.1,33) and Longitude: (-90,-79.7), Number of stations in world 1= 153, Number of stations in worlds 2 and 3= 210 3. North East: Latitude: (41.4,47.3) and Longitude: (-79.8,-67.3), Number of stations in world 1= 148, Number of stations in worlds 2 and 3= 210 4. South West:

Latitude: (32.6,38.5) and Longitude: (-123,-113.9), Number of stations in world 1= 151 , Number of stations in worlds 2 and 3= 222

The numbers of stations in these regions may be reduced very slightly as a result of final analyses, but will remain around these values and will not increase.

To maximise the usefulness of this study if you are unable to run your algorithm on all regions in all worlds it will be more beneficial if you use all worlds in a single region than if you run your algorithm on just a single world. If you are choosing just a single region please will you prioritise Wyoming with world 1 in this region being the highest priority, then world 2 and then world 3.

If you feel able to notify me of the regions and worlds you are intending to use please do so in order that I may know if any regions would be unused.

I attach an example data file and an example metadata file for Wyoming, both of these are examples only, but follow the GHCND database format and are of the same format to the data I am producing, each line begins with the station code, then the year and the month of the observations and then the letters TMEN indicating that the variable that has been recorded is mean temperature. The temperatures are recorded in tenths of a degree C and missing data values are indicated by -9999.

The data will be hosted on a web page, the link to which I will give you at the time of release. If you pass this link on to anyone else you think might be interested in participating in this work please will you inform me so that I am able to contact them.

Thank you once more for your participation in this work, I am incredibly grateful.

Best wishes,

Rachel Warren

A.3. Late October 2014 - Release of scenarios one to three

Dear All,

Thank you for your patience with the final stages of the data release. I am delighted to announce that the data are now available for download from the following link:

`http://www.metoffice.gov.uk/hadobs/benchmarks/`.

Thank you for your participation in this project. All the information that you should need is available from the above link. As I said before, if you pass this link on to anyone else you think may be interested in this project then please will you inform me so that I am able to contact them directly.

If you encounter any problems with the data or have any further questions please do not hesitate to contact me.

Many thanks and best wishes,

Rachel Warren

A.4. Mid December 2014 - Thanks for participation and release of scenario four

Dear All,

Thank you very much for returning the data for all or some of the regions and worlds I released, this is much appreciated.

For worlds one to three the deadline for the return is today - if you are planning on returning data, but have not yet done so please will you contact me so that I know to expect further contributions and what time scale I can expect these on, though I believe I already have almost of all of what I was expecting and for this I am very grateful as it should allow the next stage of the project to proceed on schedule.

There has also been a further update to the website, this introduces Wyoming world 4 - it is just 75 stations (so mimics Wyoming world 1) and its aim is to explore choices within the benchmark creation model. If you would be able to look at this world as well it would be beneficial to the study - I will of course extend the deadline for this world until the 9th January to allow you more time to work on it.

Thank you once more for your participation, I will keep you updated on the progress made,

Best wishes,

Rachel

B. Tables to summarise algorithm performance

The numbers in the following tables for non-biased stations/ stations with zero RMSE don't necessarily match with the numbers of homogeneous stations in table two from chapter five. This is because the numbers in chapter five were the number of stations with no identifiable inhomogeneities, whereas the numbers here are the numbers of stations with no inhomogeneities at all.

Table B.1. A summary of algorithm performance using bias, which is defined as the difference in means between the clean and released or clean and returned series and is therefore measured in °C. Absolute bias refers to the value obtained by taking the modulus of the bias thereby forcing it to be positive. When percentage recovery is referred to the letters indicate the following: I = Improved; a PR less than 75 indicating the improvement is not large enough and in brackets between 125 and 200, which indicates a bias better than before, but that has overshot the true value. GI = Greatly improved; a PR between 75 and 125. MW = Made worse; a PR of less than 0 or greater than 200 indicating that homogenisation increased the station bias, potentially by 'correcting' it too far. U = Unchanged; PR of 0, values in brackets in this column indicate that the bias is unchanged because the station was already unbiased. For the best and worst stations the groupings are simply Improved: PR between 0 and 200, Unchanged: PR = 0 and Made worse: PR is less than 0 or greater than 200. Values in brackets in the column referring to non-biased stations indicate the quantity of stations that have an absolute bias less than 0.05°C, which therefore effectively have a bias of zero when rounded to point one degree precision.

Scenario	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
WYW1	Released	0.053	37 (0.44)	36 (-0.34)	2 (18)	28.52	-	-	-	-	-	-	-	-	-	-	0.01	0.27
	MACD	-0.016	38 (0.11)	35 (-0.35)	2 (27)	9.42	25	10 (7)	19 (2)	12	0	9	1	10	0	0	0.01	-0.16
	Climatol-D	-0.035	27 (0.11)	46 (-0.12)	2 (29)	8.49	32	5 (2)	27 (2)	7	0	10	0	9	1	0	0.01	-0.11
	Climatol-M	0.047	36 (0.41)	37 (-0.30)	2 (18)	25.93	1	51 (0)	15 (2)	6	2	8	0	10	0	0	0.01	0.25
	MASH	-0.012	32 (0.13)	43 (-0.12)	0 (30)	9.03	38	16 (4)	0 (0)	17	5	0	5	8	0	2	-0.003	0.06
	ACMANT2	-0.037	24 (0.08)	50 (-0.09)	1 (33)	6.43	35	13 (10)	2 (1)	14	4	2	4	10	0	0	0.02	-0.10
	DAP1	0.067	41 (0.35)	33 (-0.28)	1 (16)	23.64	2	25 (4)	36 (1)	7	0	9	1	6	3	1	0.02	0.28
	HOM1	0.062	39 (0.37)	35 (-0.28)	1 (16)	24.25	3	21 (1)	42 (1)	7	0	9	1	5	4	1	0.02	0.30
	SPLIDHOM1	0.064	39 (0.37)	35 (-0.28)	1 (16)	24.34	2	22 (2)	42 (1)	6	0	9	1	5	4	1	0.02	0.31

Table B.2. A summary of algorithm performance using RMSE, which is the root mean squared error between clean and returned series or clean and released series (see chapter 6 section 3.2), reported in °C. Percentage recovery here cannot be greater than 100 as it is not possible to overshoot perfection because RMSE is constrained to be positive. The categories of PR are: I = Improved; PR between 0 and 75; GI = Greatly improved; a PR of between 75 and 100. MW = Made worse; a PR of less than 0, and U = Unchanged; a PR of 0 (values in brackets indicate no improvement possible because of perfection).

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
WYW1	Released	0.726	2 (4)	(0, 0.11)	(1.17, 1.95)	-	-	-	-	-	-	-	-	-	-	-
	MACD	0.304	2 (5)	(0, 0.26)	(0.10, 1.50)	17	30	7	19 (2)	0	9	1	10	0	0	58.10
	Climatol-D	0.272	2 (6)	(0, 0.11)	(0.11, 1.20)	21	25	0	27 (2)	0	10	0	9	1	0	62.54
	Climatol-M	0.659	2 (4)	(0, 0.11)	(1.04, 1.78)	0	56	2	15 (2)	1	7	2	10	0	0	9.22
	MASH	0.333	0 (0)	(0.09, 0.23)	(0.17, 1.70)	12	46	17	0 (0)	1	0	9	8	0	2	54.15
	ACMANT2	0.242	1 (1)	(0, 0.26)	(0.11, 0.80)	14	49	9	2 (1)	2	2	6	10	0	0	66.60
	DAP1	0.649	1 (3)	(0, 0.23)	(0.45, 1.79)	0	26	12	36 (1)	0	9	1	6	1	3	10.62
	HOM1	0.662	1 (3)	(0, 0.23)	(0.44, 1.95)	0	24	8	42 (1)	0	9	1	5	4	1	8.75
	SPLIDHOM1	0.661	1 (3)	(0, 0.23)	(0.44, 1.95)	0	24	8	42 (1)	0	9	1	5	4	1	8.93

Table B.3. A summary of algorithm performance on linear trend recovery. Note that 'significant trends preserved' (the value in brackets in the 'significant trends' column) refers to where a trend that is significant in the clean data is also significant in the released or returned data. This value is red only when the trend's significance is preserved and when its value is also preserved (with a 0.05°C buffer to allow for slight changes). All trends are in the units of °C/decade. The range of percentage recovery values are the same as for bias and therefore table 1 should be seen for PR classifications used here.

Scenario	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
WYW1	Clean	-0.027	0.279	71	0.103	4	-0.022	1	-	-	-	-	-	-	-	-	-	-	0.105	-
	Released	-0.874	0.669	50	0.216	25	-0.238	31 (1)	-	-	-	-	-	-	-	-	-	-	0.065	-
	MACD	-0.125	0.328	64	0.123	11	-0.045	6 (1)	32	10 (5)	7	19 (2)	0	1	9	10	0	0	0.096	76.99
	Climatol-D	-0.112	0.287	69	0.123	6	-0.043	4 (1)	39	2 (4)	1	27 (2)	0	0	10	9	0	1	0.110	110.66
	Climatol-M	-0.787	0.627	50	0.206	25	-0.206	30 (1)	0	52 (0)	6	15 (2)	1	1	8	10	0	0	0.068	9.03
	MASH	-0.203	0.236	68	0.119	7	-0.069	0 (0)	45	11 (5)	14	0 (0)	3	7	0	10	0	0	0.101	89.15
	ACMANT2	-0.008	0.291	73	0.123	2	-0.005	2 (0)	44	12 (6)	9	2 (2)	5	3	2	10	0	0	0.119	134.77
	DAP1	-0.671	0.491	52	0.159	23	-0.192	21 (1)	3	24 (3)	8	36 (1)	0	1	9	7	0	3	0.057	-19.17
	HOM1	-0.874	0.571	52	0.180	23	-0.212	25 (1)	4	20 (2)	6	42 (1)	0	9	1	6	0	4	0.059	-14.35
	SPLIDHOM1	-0.874	0.567	52	0.180	23	-0.214	25 (1)	3	20 (3)	6	42 (1)	0	1	9	6	0	4	0.058	-17.53

Table B.4. A summary of algorithm performance when considering similarity in inter-annual and inter-decadal variability using correlations of loess smooths. Loess smooths were compared between clean and released and clean and returned data using Spearman rank correlations and it is these correlations that were used in the calculation of percentage recovery. PR values cannot be greater than 100 and therefore are constrained to be the same as for RMSE. Colour coding of table rows to represent algorithms is the same as for other tables.

Scenario	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
WYW1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.623	0.888
	27	17	10	19 (2)	0	9	1	9	0	1	34	14	6	19 (2)	0	9	1	9	0	1	0.15 (47)	0.018 (23)	0.910	0.982
	31	13	2	27 (2)	0	10	0	9	1	0	39	7	0	27 (2)	0	10	0	9	1	0	0.05 (20)	NA	0.918	0.983
	0	52	6	15 (2)	1	7	2	9	1	0	0	56	2	15 (2)	1	7	2	10	0	0	0.001 (35)	0.0001 (23)	0.655	0.904
	39	21	15	0 (0)	4	0	6	9	0	1	48	17	10	0 (0)	4	0	6	10	0	0	0.062 (47)	0.009 (45)	0.961	0.989
	43	13	16	2 (1)	2	2	6	9	0	1	52	8	12	2 (1)	1	2	7	9	0	1	0.56 (47)	0.58 (47)	0.957	0.987
	4	24	10	36 (1)	0	9	1	7	3	0	1	30	8	36 (1)	0	9	1	7	0	3	1.16 (33)	0.11 (33)	0.694	0.918
	4	18	10	42 (1)	0	9	1	6	4	0	5	20	7	42 (1)	0	9	1	6	4	0	33 (1.21)	0.13 (33)	0.678	0.911
	4	19	9	42 (1)	0	9	1	6	4	0	4	22	6	42 (1)	0	9	1	6	4	0	1.17 (33)	0.12 (33)	0.676	0.910

Table B.5. A summary of algorithm performance on variability recovery. Variability between stations was compared using ratios of standard deviations relative to the clean series. The variability increases and decreases columns are relative to the released series; that is if returned stations were made more (less) variable than the released series what was the percentage recovery of this change? The groupings are the same as for bias, but without the 'unchanged' option as this is moot when these columns pertain specifically to variabilities that have been changed. The sums of the numbers in these columns do not equal the values in columns three and four as those pertain to the variability relative to the clean series.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
WYW1	Released	56	17	-	-	-	-	-	-	-	-	-	-	-	-	-
	MACD	43	30	5	0	7	10	25	6	21 (2)	0	9	1	9	0	1
	Climatol-D	42	31	0	3	6	13	18	6	29 (2)	0	10	0	9	1	0
	Climatol-M	53	20	1	2	12	2	29	12	17 (2)	0	7	3	6	0	4
	MASH	40	35	0	2	28	10	16	19	0 (0)	3	0	7	5	0	5
	ACMANT2	43	31	2	4	12	9	31	14	3 (1)	3	2	5	8	0	2
	DAP1	51	23	0	0	21	3	6	8	37 (1)	0	9	1	2	3	5
	HOM1	50	24	0	0	18	1	6	7	43 (1)	0	9	1	1	4	5
	SPLIDHOM1	50	24	0	0	19	2	5	6	43 (1)	0	9	1	1	4	5

Table B.6. A summary of algorithm performance on recovery and preservation of extremes. Extremes were here compared on like for like days. That is, if an algorithm did not preserve the day of the extreme it was not credited with preserving it at all. Measurement error is important here as single days are being focussed on, whereas for all other statistics aggregation of some kind has occurred and therefore random measurement error would be expected to have cancelled out. The measurement error here was calculated as 0.14°C from Brohan et al. [2006] and the number of values exact to measurement precision is indicated in brackets. Where extremes are referred to as being 'too warm' or 'too cool' here the implication is that they are more than 0.14°C away from the clean value.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
WYW1	Released	46 (56)	-	-	-	-	-	-	13 (26)	-	-	-	-	-	-
	MACD	48 (59)	2	3	1	3	7	0	22 (42)	9	3	2	5	10	4
	Climatol-D	48 (62)	2	4	0	5	6	0	20 (44)	6	5	3	3	11	3
	Climatol-M	46 (56)	4	4	0	5	6	0	14 (26)	19	7	0	15	8	0
	MASH	43 (56)	3	0	3	2	3	8	16 (38)	9	1	9	13	1	4
	ACMANT2	49 (62)	0	4	1	4	3	1	23 (48)	4	3	3	8	5	4
	DAP1	47 (57)	0	7	1	1	9	0	12 (25)	3	22	2	4	18	1
	HOM1	45 (55)	1	7	1	2	9	0	13 (26)	2	23	1	3	18	2
	SPLIDHOM1	46 (56)	1	7	1	1	9	0	12 (24)	2	24	1	3	18	3

Table B.7. A table to summarise algorithm detection ability when a window of thirty days either side of a change point was used. Colour coding is the same as for other tables, apart from that blue now represents DAP1, HOM1 and SPLIDHOM1 as all three of these algorithms had the same change point detection method applied and therefore yielded the same results. Values in brackets for hits, false alarms and CSI indicate the values that are obtained if you count multiple hits and multiple false alarms within a single window. The value not in brackets for CSI is when only multiple false alarms in windows are counted, but not multiple hits as the justification for the latter is more complicated. All values for hit rate (HR) and false alarm rate (FAR) are calculated from non-bracketed quantities. Abbreviations used in this table are as follows: CO = constant offset; EV = explanatory variables; SC = shelter changes; SR = station relocations.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
WYW1	57 (58)	40 (63)	193	279	0.228	0.125	0.467	0.182 (0.185)	51.3%	9.88%	76.7%	26.0%	2.06%	20.7%	26.3%	20.5%
	55 (55)	12 (12)	195	307	0.220	0.038	0.261	0.210 (0.210)	51.3%	8.72%	83.3%	24.4%	0%	19.8%	26.3%	18.2%
	88 (88)	56 (58)	162	263	0.352	0.176	0.568	0.286 (0.286)	62.8%	22.7%	96.7%	46.3%	2.06%	35.1%	40.0%	25.0%
	58 (58)	95 (103)	192	224	0.232	0.298	0.626	0.164 (0.164)	29.5%	20.3%	40.0%	35.0%	3.09%	19.80%	24.2%	29.5%
	4 (4)	40 (42)	246	279	0.016	0.125	0.179	0.014 (0.014)	1.28%	1.74%	3.33%	1.63%	1.03%	1.80%	0%	4.55%

Table B.8. A table to summarise algorithm detection ability when a window of ninety days either side of a change point was used. See comments for table 7 for a further explanation of the columns.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
WYW1	73 (77)	24 (45)	177	284	0.292	0.078	0.467	0.247 (0.258)	60.3%	15.1%	93.3%	35.0%	2.06%	27.0%	33.7%	25.0%
	64 (64)	6 (6)	186	302	0.256	0.019	0.261	0.250 (0.250)	53.8%	12.8%	83.3%	29.3%	3.09%	23.4%	30.5%	20.5%
	109 (109)	39 (41)	141	269	0.436	0.127	0.568	0.375 (0.375)	71.7%	30.8%	96.7%	60.2%	6.18%	44.1%	46.3%	36.4%
	109 (109)	49 (55)	141	259	0.436	0.159	0.626	0.357 (0.357)	64.1%	34.3%	70.0%	64.2%	9.28%	44.1%	45.3%	38.6%
	19 (19)	26 (27)	231	282	0.076	0.084	0.179	0.069 (0.069)	14.1%	4.65%	20.0%	9.76%	1.03%	6.31%	6.32%	13.6%

Table B.9. As in table 1, but for Wyoming scenarios 2 and 3.

Scenario	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
WYW2	Released	-0.066	63 (0.34)	84 (-0.38)	11 (36)	53.49	-	-	-	-	-	-	-	-	-	-	0	0.35
	MACD	-0.050	52 (0.09)	99 (-0.13)	7 (66)	17.63	55	26 (10)	41 (7)	19	0	6	4	10	0	0	-0.01	-0.09
	Climatol-D	-0.052	51 (0.09)	96 (-0.13)	11 (70)	17.36	55	13 (10)	58 (11)	11	0	10	0	10	0	0	0	-0.11
	Climatol-M	-0.063	66 (0.29)	82 (-0.36)	11 (36)	48.79	0	106 (1)	36 (10)	5	0	9	1	10	0	0	0	-0.34
	MASH	-0.044	45 (0.06)	113 (-0.09)	0 (74)	12.27	78	30 (12)	0 (0)	38	0	0	10	10	0	0	-0.01	0.01
	ACMANT2	-0.044	38 (0.06)	116 (-0.06)	4 (75)	11.14	77	23 (19)	10 (4)	25	0	4	6	10	0	0	-0.03	-0.10
	DAP1	-0.063	65 (0.23)	85 (-0.30)	8 (40)	40.41	16	46 (4)	73 (8)	11	0	8	2	9	1	0	-0.01	-0.31
	HOM1	-0.063	67 (0.23)	83 (-0.30)	8 (41)	40.59	12	46 (6)	76 (8)	10	0	8	2	9	1	0	-0.01	-0.32
	SPLIDHOM1	-0.061	68 (0.23)	82 (-0.31)	8 (40)	40.69	12	49 (4)	75 (8)	10	0	8	2	9	1	0	-0.01	-0.32
WYW3	Released	-0.076	61 (0.34)	88 (-0.37)	9 (30)	53.87	-	-	-	-	-	-	-	-	-	-	0	-0.53
	MACD	-0.050	46 (0.09)	106 (-0.11)	63 (6)	15.68	54	28 (14)	36 (6)	20	0	7	3	10	0	0	-0.01	-0.09
	Climatol-D	-0.060	39 (0.07)	111 (-0.11)	8 (68)	14.84	52	13 (16)	62 (8)	7	0	9	1	10	0	0	0	-0.09
	Climatol-M	-0.024	46 (0.10)	105 (-0.08)	7 (74)	12.70	68	20 (22)	27 (7)	14	0	8	2	9	0	1	-0.02	0.12
	MASH	-0.013	50 (0.10)	108 (-0.06)	0 (95)	11.65	82	29 (21)	0 (0)	26	1	0	9	10	0	0	0	0.14
	ACMANT2	-0.024	41 (0.09)	113 (-0.07)	4 (84)	10.99	81	22 (21)	8 (4)	22	0	4	6	9	0	1	-0.03	0.13
	DAP1	-0.067	61 (0.26)	88 (-0.30)	9 (36)	42.58	13	45 (7)	73 (9)	11	0	10	0	8	1	1	0	-0.40
	HOM1	-0.061	64 (0.27)	85 (-0.32)	9 (34)	44.22	12	41 (7)	79 (9)	10	0	10	0	8	1	1	0	-0.42
	SPLIDHOM1	-0.060	64 (0.27)	85 (-0.31)	9 (35)	43.71	11	42 (7)	78 (9)	11	0	10	0	8	1	1	0	-0.42

Table B.10. As in table 2, but for Wyoming scenarios 2 and 3.

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
WYW2	Released	0.654	11 (15)	(0, 0)	(1.25, 2.53)	-	-	-	-	-	-	-	-	-	-	-
	MACD	0.264	7 (13)	(0, 0.29)	(0.10, 0.89)	36	54	20	41 (7)	0	6	4	10	0	0	59.90
	Climatol-D	0.249	18 (11)	(0, 0)	(0.06, 0.55)	44	41	4	58 (11)	0	10	0	10	0	0	62.23
	Climatol-M	0.600	14 (11)	(0, 0.07)	(1.12, 2.28)	0	108	4	36 (10)	0	9	1	10	0	0	8.81
	MASH	0.217	0 (2)	(0, 0.11)	(0.12, 0.72)	31	94	33	0 (0)	0	0	10	10	0	0	67.07
	ACMANT2	0.192	9 (4)	(0, 0.16)	(0.09, 0.59)	40	88	16	10 (4)	0	4	6	10	0	0	70.87
	DAP1	0.536	8 (12)	(0, 0.24)	(0.43, 1.42)	6	58	13	73 (8)	0	8	2	9	0	1	18.55
	HOM1	0.538	8 (12)	(0, 0.23)	(0.47, 1.42)	6	56	12	76 (8)	0	8	2	9	0	1	18.25
	SPLIDHOM1	0.536	8 (11)	(0, 0.23)	(0.42, 1.42)	6	56	13	75 (8)	0	8	2	9	0	1	18.55
WYW3	Released	0.696	9 (13)	(0, 0.02)	(1.34, 2.73)	-	-	-	-	-	-	-	-	-	-	-
	MACD	0.252	6 (13)	(0, 0.20)	(0.10, 0.58)	32	62	22	36 (6)	0	7	3	10	0	0	63.84
	Climatol-D	0.237	8 (14)	(0, 0.05)	(0.10, 0.58)	42	43	3	62 (8)	0	9	1	10	0	0	65.88
	Climatol-M	0.246	7 (13)	(0, 0.13)	(0.10, 1.88)	41	75	8	27 (7)	0	8	2	9	0	1	64.67
	MASH	0.258	0 (4)	(0.02, 0.12)	(0.16, 1.84)	37	97	24	0 (0)	1	0	9	10	0	0	62.93
	ACMANT2	0.238	4 (8)	(0, 0.14)	(0.11, 1.84)	43	89	14	8 (4)	0	4	6	10	0	0	65.79
	DAP1	0.615	9 (13)	(0, 0.02)	(1.34, 2.39)	4	62	10	73 (9)	0	10	0	8	1	1	11.65
	HOM1	0.627	9 (14)	(0, 0.02)	(1.34, 2.50)	3	58	9	79 (9)	0	10	0	8	1	1	9.94
	SPLIDHOM1	0.622	9 (14)	(0, 0.02)	(1.34, 2.39)	3	59	9	78 (9)	0	10	0	8	1	1	10.73

Table B.11. As in table 3, but for Wyoming scenarios 2 and 3.

Scenario	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
WYW2	Clean	-0.044	0.253	152	0.108	6	-0.027	1	-	-	-	-	-	-	-	-	-	-	0.102	-
	Released	-0.510	0.756	118	0.234	40	-0.205	65 (0)	-	-	-	-	-	-	-	-	-	-	0.122	-
	MACD	-0.158	0.378	146	0.136	12	-0.052	17 (1)	61	24 (10)	15	41 (7)	0	4	6	10	0	0	0.121	4.74
	Climatol-D	-0.158	0.476	146	0.139	12	-0.060	15 (1)	69	12 (4)	4	58 (11)	0	0	10	10	0	0	0.123	-6.54
	Climatol-M	-0.441	0.725	120	0.220	38	-0.185	61 (0)	0	107 (0)	5	36 (10)	0	1	9	10	0	0	0.122	0.78
	MASH	-0.058	0.585	152	0.131	6	-0.036	3 (0)	88	29 (18)	23	0 (0)	0	0	10	10	0	0	0.124	-8.88
	ACMANT2	-0.027	0.378	152	0.108	6	-0.027	4 (1)	87	21 (22)	4	14 (10)	0	6	4	10	0	0	0.120	13.09
	DAP1	-0.373	0.675	127	0.196	31	-0.155	50 (0)	15	49 (5)	8	73 (8)	0	2	8	8	1	1	0.127	-20.72
	HOM1	-0.373	0.674	127	0.196	31	-0.153	51 (0)	13	48 (5)	8	76 (8)	0	2	8	7	2	1	0.127	-23.84
	SPLIDHOM1	-0.373	0.673	126	0.197	32	-0.150	50 (0)	13	50 (5)	7	75 (8)	0	2	8	8	1	1	0.127	-19.74
WYW3	Clean	-0.065	0.342	153	0.109	5	-0.039	1	-	-	-	-	-	-	-	-	-	-	0.104	-
	Released	-0.788	1.102	123	0.249	35	-0.205	60 (0)	-	-	-	-	-	-	-	-	-	-	0.149	-
	MACD	-0.114	0.459	149	0.146	9	-0.072	12 (0)	69	22 (11)	14	36 (6)	0	3	7	10	0	0	0.133	34.23
	Climatol-D	-0.086	0.459	153	0.145	5	-0.049	14 (0)	63	10 (11)	4	62 (8)	0	1	9	10	0	0	0.138	24.88
	Climatol-M	-0.117	0.352	153	0.131	5	-0.065	4 (1)	84	16 (11)	13	27 (7)	0	2	8	9	1	0	0.124	55.25
	MASH	-0.172	0.371	152	0.121	6	-0.043	4 (1)	99	25 (10)	24	0 (0)	1	0	9	9	0	1	0.115	73.85
	ACMANT2	-0.089	0.297	156	0.129	2	-0.053	4 (1)	84	26 (17)	19	8 (4)	0	6	4	9	0	1	0.126	50.42
	DAP1	-0.707	1.011	127	0.222	31	-0.171	48 (0)	15	45 (6)	10	73 (9)	0	0	10	8	1	1	0.142	14.91
	HOM1	-0.709	1.015	127	0.220	31	-0.174	49 (0)	15	42 (4)	9	79 (9)	0	0	10	8	1	1	0.143	11.53
	SPLIDHOM1	-0.707	1.013	127	0.218	31	-0.171	47 (0)	15	43 (4)	9	78 (9)	0	0	10	8	1	1	0.143	12.66

Table B.12. As in table 4, but for Wyoming scenarios 2 and 3.

Scen.	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
WYW2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.677	0.905
	54	32	24	41 (7)	0	6	4	9	0	1	68	27	15	41 (7)	0	6	4	10	0	0	0.30 (23)	0.02 (67)	0.923	0.984
	54	25	10	58 (11)	0	10	0	10	0	0	78	9	2	58 (11)	0	10	0	10	0	0	0.183 (12)	0.004 (118)	0.929	0.985
	0	106	5	36 (10)	0	9	1	10	0	0	0	111	1	36 (10)	0	9	1	10	0	0	0.232 (87)	0.057 (85)	0.707	0.919
	92	48	18	0 (0)	0	0	10	10	0	0	99	38	21	0 (0)	0	0	10	10	0	0	0.480 (31)	0.012 (41)	0.968	0.991
	92	39	13	10 (4)	0	4	6	9	0	1	102	30	12	10 (4)	0	4	6	10	0	0	0.418 (31)	0.007 (26)	0.974	0.992
	18	50	9	73 (8)	0	8	2	9	1	0	18	53	6	73 (8)	0	8	2	9	1	0	43 (0.161)	145 (0.057)	0.760	0.938
	18	49	7	76 (8)	0	8	2	9	1	0	19	50	5	76 (8)	0	8	2	9	0	1	43 (0.187)	145 (0.059)	0.762	0.938
	17	52	6	75 (8)	0	8	2	9	1	0	19	51	5	75 (8)	0	8	2	9	1	0	43 (0.158)	5 (0.066)	0.761	0.938
WYW3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.694	0.897
	54	43	19	36 (6)	0	7	3	9	0	1	66	37	13	36 (6)	0	7	3	10	0	0	0.240 (138)	0.019 (51)	0.932	0.985
	65	16	7	62 (8)	0	9	1	9	0	1	79	7	2	62 (8)	0	9	1	10	0	0	0.240 (138)	0.012 (83)	0.938	0.985
	74	36	12	27 (7)	0	8	2	9	0	1	91	25	8	27 (7)	0	8	2	10	0	0	0.145 (73)	0.025 (6)	0.962	0.989
	103	39	16	0 (0)	1	0	9	8	0	2	102	40	16	0 (0)	1	0	9	10	0	0	0.167 (47)	0.011 (6)	0.974	0.991
	88	40	18	8 (4)	0	4	6	9	0	1	99	32	15	8 (4)	0	4	6	10	0	0	0.226 (47)	0.020 (75)	0.973	0.991
	13	51	12	73 (9)	0	10	0	9	1	0	20	50	6	73 (9)	0	10	0	9	1	0	119 (0.238)	17 (0.064)	0.765	0.926
	11	49	10	79 (9)	0	10	0	9	1	0	20	46	4	79 (9)	0	10	0	9	1	0	119 (0.217)	17 (0.062)	0.757	0.923
	12	48	11	78 (9)	0	10	0	9	1	0	19	47	5	78 (9)	0	10	0	9	1	0	119 (0.241)	17 (0.067)	0.759	0.924

Table B.13. As in table 5, but for Wyoming scenarios 2 and 3.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
WYW2	Released	110	37	-	-	-	-	-	-	-	-	-	-	-	-	-
	MACD	93	58	3	9	16	23	45	14	48 (7)	0	6	4	6	0	4
	Climatol-D	95	52	2	8	8	30	31	10	69 (11)	0	10	0	8	0	2
	Climatol-M	104	44	0	6	16	4	66	20	46 (10)	0	9	1	7	0	3
	MASH	97	61	3	6	62	10	34	43	0 (0)	0	0	10	2	0	8
	ACMANT2	92	62	1	14	23	33	48	25	14 (4)	0	4	6	7	0	3
	DAP1	94	56	0	0	30	8	18	21	81 (8)	0	8	2	3	1	7
	HOM1	90	60	0	0	29	8	15	22	84 (8)	0	8	2	2	1	7
	SPLIDHOM1	94	56	0	0	27	10	19	19	83 (8)	0	8	2	3	1	6
WYW3	Released	130	19	-	-	-	-	-	-	-	-	-	-	-	-	-
	MACD	108	44	0	2	22	31	46	15	42 (6)	0	7	3	10	0	0
	Climatol-D	119	31	1	3	14	30	36	4	70 (8)	0	9	1	10	0	0
	Climatol-M	90	61	1	1	28	20	41	33	34 (7)	0	8	2	8	0	2
	MASH	103	55	2	3	55	19	48	31	0 (0)	1	0	9	7	0	3
	ACMANT2	113	41	2	3	36	31	59	15	12 (4)	0	4	6	10	0	0
	DAP1	113	36	0	1	35	6	20	14	82 (9)	0	10	0	3	1	6
	HOM	109	40	0	0	28	6	23	13	88 (9)	0	10	0	4	1	5
	SPLIDHOM1	109	40	0	0	33	12	9	17	87 (9)	0	10	0	4	1	5

Table B.14. As in table 6, but for Wyoming scenarios 2 and 3.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
WYW2	Released	101 (113)	-	-	-	-	-	-	38 (69)	-	-	-	-	-	-
	MACD	102 (126)	2	11	1	7	11	0	40 (91)	8	12	4	18	18	7
	Climatol-D	104 (127)	1	13	0	3	12	2	48 (101)	4	9	1	16	23	4
	Climatol-M	101 (113)	14	8	0	11	12	0	40 (69)	24	14	0	30	21	0
	MASH	102 (129)	7	3	4	7	3	5	38 (87)	14	3	7	24	4	19
	ACMANT2	106 (140)	4	2	1	6	4	1	46 (121)	2	2	4	15	4	10
	DAP1	101 (113)	1	20	2	3	19	0	38 (70)	4	27	7	7	39	4
	HOM1	101 (113)	1	20	2	3	19	0	40 (73)	5	28	3	3	40	6
	SPLIDHOM1	101 (113)	1	20	2	3	19	0	38 (72)	6	28	3	5	40	4
WYW3	Released	114 (125)	-	-	-	-	-	-	51 (68)	-	-	-	-	-	-
	MACD	117 (131)	6	6	1	4	9	1	56 (89)	5	10	2	25	18	9
	Climatol-D	116 (132)	6	5	0	1	14	0	62 (100)	4	6	1	19	23	5
	Climatol-M	115 (138)	1	4	2	7	6	0	58 (99)	9	5	5	20	9	11
	MASH	101 (124)	3	4	5	9	4	9	49 (101)	11	3	8	24	0	11
	ACMANT2	121 (139)	6	3	0	2	5	3	58 (121)	4	4	2	13	3	11
	DAP1	114 (126)	2	11	0	17	2	0	51 (72)	10	25	1	7	39	4
	HOM1	114 (125)	2	12	0	1	18	0	52 (74)	8	26	3	4	39	4
	SPLIDHOM1	114 (126)	2	12	0	1	17	0	53 (73)	8	27	2	4	39	5

Table B.15. As in table 7, but for Wyoming scenarios 2 and 3.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
WYW2	87 (87)	71 (138)	376	547	0.188	0.115	0.470	0.145 (0.145)	39.7%	9.15%	86.0%	22.5%	0%	17.8%	23.1%	11.6%
	87 (87)	31 (31)	376	587	0.188	0.050	0.246	0.176 (0.176)	40.4%	8.83%	88.4%	20.7%	1.51%	22.0%	19.4%	10.5%
	148 (148)	100 (105)	315	518	0.320	0.162	0.528	0.261 (0.261)	54.1%	21.8%	90.7%	43.7%	6.06%	30.9%	37.1%	23.3%
	94 (94)	193 (214)	369	425	0.203	0.312	0.641	0.139 (0.139)	26.0%	17.7%	39.5%	29.3%	6.06%	22.5%	19.4%	17.4%
	16 (16)	92 (95)	447	526	0.035	0.149	0.230	0.029 (0.029)	6.16%	2.21%	9.30%	4.95%	0.51%	4.71%	2.69%	2.33%
WYW3	107 (109)	84 (146)	338	513	0.240	0.141	0.566	0.181 (0.184)	57.7%	9.09%	88.5%	22.8%	1.82%	19.3%	29.5%	NA
	106 (106)	18 (18)	339	579	0.238	0.030	0.274	0.229 (0.229)	60.6%	7.47%	88.5%	22.8%	1.21%	19.7%	28.5%	NA
	102 (102)	96 (101)	343	501	0.229	0.161	0.454	0.187 (0.187)	49.6%	11.0%	65.6%	25.6%	3.64%	18.5%	28.0%	NA
	105 (105)	202 (231)	340	394	0.236	0.339	0.752	0.155 (0.155)	36.5%	17.9%	36.1%	28.8%	12.1%	22.3%	25.1%	NA
	7 (7)	87 (93)	438	510	0.016	0.146	0.225	0.013 (0.013)	2.92%	0.97%	1.64%	2.74%	0%	2.10%	0.97%	NA

Table B.16. As in table 8, but for Wyoming scenarios 2 and 3.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
WYW2	114 (115)	52 (114)	349	555	0.246	0.086	0.470	0.198 (0.199)	47.9%	13.9%	90.7%	32.9%	1.01%	24.1%	28.5%	17.4%
	102 (102)	18 (18)	361	589	0.220	0.030	0.246	0.212 (0.212)	45.2%	11.4%	88.4%	27.0%	2.02%	24.6%	22.6%	15.1%
	201 (201)	51 (54)	262	556	0.434	0.084	0.528	0.389 (0.389)	65.1%	33.4%	97.7%	62.2%	10.6%	44.5%	47.8%	31.4%
	182 (182)	118 (129)	281	489	0.393	0.194	0.641	0.307 (0.307)	51.4%	33.8%	88.4%	54.5%	11.6%	40.3%	41.4%	32.6%
	48 (48)	60 (63)	415	547	0.104	0.099	0.230	0.091 (0.091)	19.9%	5.99%	39.5%	12.2%	2.02%	11.0%	11.83%	5.81%
WYW3	138 (140)	56 (117)	307	529	0.310	0.096	0.567	0.246 (0.248)	69.3%	14.0%	98.4%	34.2%	1.82%	26.5%	36.2%	NA
	118 (118)	9 (9)	327	576	0.265	0.014	0.274	0.260 (0.260)	65.0%	9.42%	91.8%	25.6%	3.64%	22.7%	30.9%	NA
	137 (137)	66 (68)	308	519	0.308	0.113	0.454	0.267 (0.267)	60.6%	17.5%	75.4%	37.4%	5.45%	26.1%	36.2%	NA
	202 (203)	116 (136)	243	468	0.454	0.199	0.753	0.347 (0.348)	70.8%	34.1%	80.3%	54.3%	20.6%	42.9%	48.3%	NA
	48 (48)	51 (54)	397	534	0.108	0.087	0.225	0.096 (0.096)	27.7%	3.25%	29.5%	13.2%	0.61%	9.66%	12.1%	NA

Table B.17. As in table 1, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
WYW4	Released	0.041	34 (0.48)	39 (-0.34)	2 (12)	29.62	-	-	-	-	-	-	-	-	-	-	0.01	0.20
	MACD	0.043	50 (0.14)	25 (-0.15)	0 (15)	10.90	26	14 (6)	4 (0)	25	2	1	7	10	0	0	0.01	0.02
	Climatol-D	0	37 (0.14)	38 (-0.13)	0 (18)	10.33	23	10 (6)	21 (0)	15	1	4	5	10	0	0	0.01	0.05
	Climatol-M	0.021	37 (0.22)	38 (-0.18)	0 (22)	14.90	24	17 (7)	0 (0)	27	3	0	7	8	0	2	0.01	0.27
	MASH	0.022	38 (0.18)	37 (-0.14)	0 (26)	11.97	29	22 (4)	0 (0)	20	3	0	7	7	0	3	0.03	0.20
	ACMANT2	-0.044	28 (0.10)	47 (-0.13)	0 (24)	8.75	31	6 (9)	6 (0)	23	1	1	8	9	0	1	-0.06	-0.15
	DAP1	0.023	34 (0.43)	39 (-0.33)	2 (12)	27.74	0	18 (0)	53 (2)	2	0	10	0	4	6	0	0.01	0.12
	HOM1	0.025	34 (0.44)	39 (-0.34)	2 (12)	29.20	0	17 (0)	54 (2)	2	0	10	0	3	7	0	0.01	0.13
	SPLIDHOM1	0.026	34 (0.44)	39 (-0.33)	12 (2)	28.04	0	17 (0)	54 (2)	2	0	10	0	3	7	0	0.01	0.13
SEW1	Released	-0.078	64 (0.34)	86 (-0.39)	3 (44)	55.82	-	-	-	-	-	-	-	-	-	-	0	0.04
	Climatol-D	-0.040	47 (0.03)	103 (-0.08)	3 (105)	9.31	72	10 (9)	52 (3)	7	0	10	0	10	0	0	0	-0.04
	Climatol-M	-0.054	31 (0.12)	120 (-0.10)	2 (80)	15.77	64	19 (17)	23 (2)	28	0	5	5	10	0	0	-0.03	0.09
	MASH	-0.005	74 (0.09)	79 (-0.09)	0 (91)	13.49	77	28 (15)	0 (0)	33	3	0	7	10	0	0	0	0.16
	ACMANT2	-0.040	45 (0.06)	106 (-0.08)	2 (96)	11.37	75	20 (15)	13 (2)	28	0	4	6	10	0	0	-0.02	-0.02
	DAP1	-0.084	55 (0.20)	95 (-0.25)	3 (59)	34.60	25	45 (2)	65 (3)	13	0	10	0	10	0	0	0	0
	HOM1	-0.091	56 (0.23)	94 (-0.28)	3 (56)	39.32	21	43 (2)	69 (3)	15	0	10	0	8	1	1	0	-0.06
	SPLIDHOM1	-0.085	56 (0.19)	94 (-0.25)	3 (58)	34.73	25	45 (2)	65 (3)	13	0	10	0	10	0	0	0	-0.01

Table B.18. As in table 2, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
WYW4	Released	0.785	2 (4)	(0, 0.14)	(1.20, 2.38)	-	-	-	-	-	-	-	-	-	-	-
	MACD	0.347	0 (2)	(0.03, 0.44)	(0.21, 0.64)	10	28	33	4 (0)	0	1	9	10	0	0	55.78
	Climatol-D	0.292	0 (1)	(0.09, 0.21)	(0.16, 0.29)	15	22	17	21 (0)	0	4	6	10	0	0	62.80
	Climatol-M	0.478	0 (0)	(0.12, 0.29)	(0.29, 1.76)	6	31	38	0 (0)	0	0	10	8	0	2	39.07
	MASH	0.415	0 (0)	(0.14, 0.40)	(0.27, 1.67)	4	40	31	0	0	0	10	8	0	2	47.18
	ACMANT2	0.301	0 (0)	(0.12, 0.37)	(0.17, 1.13)	16	35	18	6 (0)	0	1	9	10	0	0	61.67
	DAP1	0.762	2 (4)	(0, 0.14)	(1.20, 2.38)	0	16	4	53 (2)	0	10	0	4	6	0	2.96
	HOM1	0.767	2 (4)	(0, 0.14)	(1.20, 2.38)	0	13	6	54 (2)	0	10	0	3	7	0	2.28
	SPLIDHOM1	0.767	2 (4)	(0, 0.14)	(1.20, 2.38)	0	14	5	54 (2)	0	10	0	3	7	0	2.36
SEW1	Released	0.730	3 (18)	(0, 0.03)	(1.36, 3.06)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.177	3 (27)	(0, 0.03)	(0.04, 0.23)	36	39	3	52 (3)	0	10	0	10	0	0	75.75
	Climatol-M	0.272	2 (14)	(0, 0.17)	(0.08, 1.52)	43	60	25	23 (2)	0	5	5	10	0	0	62.74
	MASH	0.256	0 (9)	(0.01, 0.12)	(0.10, 1.51)	48	64	41	0 (0)	0	0	10	10	0	0	64.95
	ACMANT2	0.215	2 (12)	(0, 0.09)	(0.11, 0.63)	43	74	21	13 (2)	0	4	6	10	0	0	70.47
	DAP1	0.525	3 (18)	(0, 0.03)	(0.57, 1.64)	9	65	11	65 (3)	0	10	0	10	0	0	28.11
	HOM1	0.699	3 (18)	(0, 0.03)	(0.57, 3.97)	8	59	14	69 (3)	0	10	0	8	1	1	4.21
	SPLIDHOM1	0.526	18 (3)	(0, 0.03)	(0.57, 1.65)	9	64	12	65 (3)	0	10	0	10	0	0	27.93

Table B.19. As in table 3, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
WYW4	Clean	-0.038	0.242	72	0.117	3	-0.025	2	-	-	-	-	-	-	-	-	-	-	0.110	-
	Released	-1.000	1.045	52	0.222	23	-0.238	28 (0)	-	-	-	-	-	-	-	-	-	-	0.079	-
	MAC	-0.096	0.305	69	0.110	6	-0.042	3 (0)	31	16 (6)	18	4 (0)	3	6	1	10	0	0	0.096	54.19
	Climatol-D	-0.092	0.303	69	0.121	6	-0.036	6 (0)	29	10 (8)	7	21 (0)	4	2	4	10	0	0	0.105	86.67
	Climatol-M	-0.081	0.303	69	0.116	6	-0.032	4 (0)	30	21 (7)	17	0 (0)	6	4	0	10	0	0	0.101	72.30
	MASH	-0.147	0.246	65	0.117	7	-0.069	2 (0)	28	23 (10)	14	0 (0)	4	0	6	10	0	0	0.093	47.33
	ACMANT2	-0.036	0.244	72	0.126	3	-0.025	1 (0)	41	16 (3)	15	0 (0)	4	6	0	10	0	0	0.118	127.96
	DAP1	-0.996	1.045	53	0.216	22	-0.230	25 (1)	0	17 (1)	2	53 (2)	0	10	0	4	6	0	0.088	31.02
	HOM1	-0.996	1.045	53	0.217	22	-0.218	25 (1)	0	17 (0)	2	54 (2)	0	10	0	3	7	0	0.087	26.86
	SPLIDHOM1	-0.996	1.045	52	0.219	23	-0.203	25 (1)	1	16 (0)	2	54 (2)	0	10	0	3	7	0	0.087	27.07
SEW1	Clean	-0.056	0.188	141	0.060	12	-0.032	4	-	-	-	-	-	-	-	-	-	-	0.052	-
	Released	-0.820	0.866	110	0.188	43	-0.166	68 (2)	-	-	-	-	-	-	-	-	-	-	0.087	-
	Climatol-D	-0.089	0.197	143	0.069	10	-0.025	9 (3)	70	12 (14)	2	52 (3)	0	0	10	10	0	0	0.062	70.79
	Climatol-M	-0.038	0.237	152	0.083	1	-0.038	18 (1)	65	18 (23)	22	23 (2)	0	5	5	10	0	0	0.081	18.29
	MASH	-0.040	0.217	137	0.063	16	-0.017	6 (3)	91	22 (12)	28	0 (0)	3	0	7	10	0	0	0.05	95.96
	ACMANT2	-0.062	0.273	152	0.070	1	-0.062	9 (1)	84	13 (21)	20	13 (2)	0	6	4	10	0	0	0.067	55.56
	DAP1	-0.369	0.838	123	0.121	30	-0.089	48 (2)	24	48 (3)	10	65 (3)	0	10	0	10	0	0	0.079	24.37
	HOM1	-0.372	0.825	124	0.129	24	-0.101	53 (2)	21	46 (2)	12	69 (3)	0	10	0	8	1	1	0.084	8.43
	SPLIDHOM1	-0.380	0.824	124	0.120	29	-0.092	49 (2)	24	47 (4)	10	65 (3)	0	10	0	10	0	0	0.078	24.94

Table B.20. As in table 4, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
WYW4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.624	0.882
	22	21	28	4 (0)	2	1	7	7	0	3	22	30	19	4 (0)	1	1	8	10	0	0	0.310 (33)	0.052 (45)	0.888	0.973
	26	15	13	21 (0)	1	4	5	8	0	2	31	16	7	21 (0)	1	4	5	10	0	0	0.195 (29)	0 (23)	0.912	0.981
	26	24	25	0 (0)	4	0	6	7	0	3	29	30	16	0 (0)	4	0	6	10	0	0	0.423 (47)	0 (66)	0.904	0.973
	33	19	23	0 (0)	1	0	9	7	0	3	36	25	14	0 (0)	0	0	10	10	0	0	0.148 (28)	0.019 (45)	0.933	0.982
	37	15	23	0 (0)	3	0	7	6	0	4	39	18	18	0 (0)	1	0	9	9	0	1	0.558 (47)	0.063 (47)	0.932	0.980
	0	13	7	53 (2)	0	10	0	4	6	0	0	17	3	53 (2)	0	10	0	4	6	0	0.021 (30)	0.007 (30)	0.658	0.894
	0	11	8	54 (2)	0	10	0	2	7	1	0	16	3	54 (2)	0	10	0	3	7	0	0.020 (30)	0.007 (30)	0.651	0.891
	0	12	7	54 (2)	0	10	0	2	7	1	0	16	3	54 (2)	0	10	0	3	7	0	0.020 (30)	0.006 (30)	0.656	0.893
SEW1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.557	0.829
	78	18	2	52 (3)	0	10	0	10	0	0	87	9	2	52 (3)	0	10	0	10	0	0	0.001 (150)	0.017 (92)	0.943	0.989
	70	39	18	23 (2)	0	5	5	10	0	0	85	22	21	23 (2)	0	5	5	10	0	0	0.255 (140)	0.057 (129)	0.921	0.984
	76	46	31	0 (0)	2	0	8	10	0	0	97	23	33	0 (0)	0	0	10	10	0	0	0.190 (88)	0.068 (49)	0.958	0.991
	78	34	26	13 (2)	0	4	6	9	0	1	85	33	20	13 (2)	0	4	6	10	0	0	0.253 (88)	0.047 (88)	0.942	0.986
	19	55	11	65 (3)	0	10	0	9	0	1	30	48	7	65 (3)	0	10	0	10	0	0	0.192 (88)	0.143 (34)	0.729	0.921
	19	55	7	69 (3)	0	10	0	9	1	0	25	47	9	69 (3)	0	10	0	8	1	1	0.309 (86)	0.149 (88)	0.717	0.913
	20	56	9	65 (3)	0	10	0	10	0	0	31	47	7	65 (3)	0	10	0	10	0	0	1.412 (29)	0.524 (56)	0.728	0.921

Table B.21. As in table 5, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged from released (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
WYW4	Released	55	18	-	-	-	-	-	-	-	-	-	-	-	-	-
	MACD	32	43	3	5	8	9	24	22	4 (0)	1	1	8	9	0	1
	Climatol-D	44	31	2	5	3	9	26	9	21 (0)	3	4	3	10	0	0
	Climatol-M	45	30	2	5	18	4	29	17	0 (0)	4	0	6	7	0	3
	MASH	36	39	0	1	25	6	19	24	0 (0)	1	0	9	8	0	2
	ACMANT2	45	30	0	7	9	8	29	16	6 (0)	1	1	8	10	0	0
	DAP1	51	22	0	0	11	2	3	4	55 (2)	0	10	0	0	6	4
	HOM	53	20	0	0	11	3	4	1	56 (2)	0	10	0	1	7	2
	SPLIDHOM1	54	19	0	0	11	3	3	2	56 (2)	0	10	0	1	7	2
SEW1	Released	114	36	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	97	53	2	5	11	33	38	9	55 (3)	0	10	0	10	0	0
	Climatol-M	101	50	6	6	40	14	42	20	25 (2)	1	5	4	8	0	2
	MASH	86	67	3	7	35	36	34	38	0 (0)	4	0	6	9	0	1
	ACMANT2	90	61	2	14	19	32	51	20	15 (2)	4	4	2	10	0	0
	DAP1	90	60	0	3	31	5	26	20	68 (3)	0	10	0	4	0	6
	HOM	87	63	1	1	30	6	22	21	72 (3)	0	10	0	4	1	5
	SPLIDHOM1	92	58	1	1	34	7	22	20	68 (3)	0	10	0	4	0	6

Table B.22. As in table 6, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
WYW4	Released	56 (63)	-	-	-	-	-	-	11 (22)	-	-	-	-	-	-
	MACD	42 (55)	2	1	6	1	2	8	15 (31)	16	3	11	8	3	3
	Climatol-D	52 (59)	0	2	4	0	3	7	16 (36)	10	11	0	7	10	1
	Climatol-M	42 (48)	0	3	1	0	4	19	15 (33)	14	12	0	5	9	2
	MASH	14 (15)	0	1	2	2	0	55	16 (29)	21	3	6	6	1	9
	ACMANT2	46 (56)	0	3	0	1	4	11	10 (16)	8	7	2	7	7	3
	DAP1	56 (63)	0	6	0	0	6	0	11 (22)	2	25	2	1	21	2
	HOM1	56 (63)	0	6	0	0	6	0	11 (23)	1	25	1	0	22	3
	SPLIDHOM1	56 (63)	0	6	0	0	6	0	11 (24)	0	25	1	1	21	3
SEW1	Released	84 (101)	-	-	-	-	-	-	47 (67)	-	-	-	-	-	-
	Climatol-D	100 (134)	0	6	0	4	9	0	72 (107)	6	9	1	11	15	4
	Climatol-M	79 (126)	0	7	0	9	10	1	59 (95)	9	8	3	14	11	13
	MASH	84 (124)	8	3	1	9	3	5	48 (98)	11	5	4	11	7	17
	ACMANT2	97 (134)	4	3	0	6	5	1	60 (104)	8	5	5	11	5	15
	DAP1	87 (106)	2	15	0	5	22	3	50 (77)	12	14	2	17	26	5
	HOM1	86 (106)	2	17	0	2	23	3	52 (79)	12	15	1	14	26	6
	SPLIDHOM1	87 (108)	2	15	0	3	22	3	50 (74)	14	13	2	19	25	6

Table B.23. As in table 7, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
WYW4	44 (46)	132 (293)	208	189	0.175	0.411	1.309	0.081 (0.084)	38.5%	8.05%	75.0%	15.9%	0%	11.7%	27.4%	10.9%
	40 (40)	22 (23)	212	299	0.159	0.070	0.243	0.145 (0.145)	38.5%	5.75%	65.6%	14.3%	1.06%	14.4%	18.9%	13.0%
	36 (36)	31 (31)	216	290	0.143	0.097	0.259	0.127 (0.127)	32.1%	6.32%	40.6%	17.5%	1.06%	12.6%	20.0%	6.52%
	39 (39)	79 (89)	213	241	0.155	0.247	0.498	0.114 (0.114)	29.5%	9.20%	50.0%	17.5%	1.06%	12.6%	18.9%	15.2%
	0 (0)	27 (27)	252	294	0	0.084	0.104	0 (0)	0%	0%	0%	0%	0%	0%	0%	0%
SEW1	101 (101)	40 (42)	262	470	0.278	0.078	0.379	0.249 (0.249)	63.5%	3.26%	95.6%	58.0%	3.04%	24.1%	47.1%	7.32%
	85 (85)	108 (117)	278	402	0.234	0.212	0.516	0.177 (0.177)	47.3%	6.98%	73.3%	43.2%	6.09%	19.1%	37.8%	11.0%
	72 (73)	198 (246)	291	311	0.198	0.389	0.849	0.118 (0.120)	29.7%	13.0%	40.0%	26.1%	13.5%	21.0%	20.2%	17.1%
	14 (14)	90 (92)	349	420	0.039	0.176	0.281	0.031 (0.031)	8.78%	0.47%	15.6%	5.68%	0.87%	4.32%	5.04%	1.22%

Table B.24. As in table 8, but for Wyoming scenario 4 and the South East scenario 1.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
WYW4	57 (64)	120 (155)	195	189	0.226	0.388	1.309	0.108 (0.120)	50.0%	10.3%	87.5%	23.0%	0%	17.1%	32.6%	15.2%
	51 (51)	11 (12)	201	298	0.202	0.032	0.243	0.194 (0.194)	51.3%	6.32%	84.4%	18.3%	1.06%	19.8%	23.2%	15.2%
	45 (45)	23 (23)	207	286	0.179	0.074	0.259	0.164 (0.164)	39.7%	8.05%	50.0%	21.4%	2.13%	16.2%	24.2%	8.70%
	68 (68)	56 (60)	184	252	0.270	0.182	0.498	0.221 (0.221)	52.6%	15.5%	71.9%	34.1%	2.13%	22.5%	34.7%	21.7%
	12 (12)	15 (15)	240	294	0.048	0.049	0.104	0.045 (0.045)	11.5%	1.74%	18.8%	4.76%	0%	4.50%	4.21%	6.52%
SEW1	120 (120)	26 (27)	143	471	0.331	0.052	0.379	0.321 (0.321)	68.2%	8.84%	100%	65.9%	7.39%	30.9%	50.4%	12.2%
	124 (124)	79 (82)	239	418	0.342	0.159	0.536	0.279 (0.279)	56.8%	18.6%	86.7%	59.1%	14.3%	33.3%	43.7%	22.0%
	155 (158)	136 (164)	208	359	0.427	0.275	0.849	0.294 (0.298)	64.9%	27.4%	82.2%	68.2%	25.2%	43.2%	49.6%	31.7%
	55 (55)	52 (53)	308	445	0.152	0.105	0.281	0.132 (0.132)	29.1%	5.58%	44.4%	29.5%	3.91%	13.6%	20.2%	11.0%

Table B.25. As in table 1, but for the South East scenarios 2 and 3.

Scen.	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
SEW2	Released	-0.024	88 (0.28)	111 (-0.27)	11 (67)	54.08	-	-	-	-	-	-	-	-	-	-	0	0.33
	Climatol-D	-0.025	55 (0.03)	141 (-0.05)	14 (156)	8.61	96	13 (13)	72 (11)	5	0	10	0	10	0	0	0	-0.04
	Climatol-M	-0.033	54 (0.05)	149 (-0.06)	7 (151)	12.20	97	18 (20)	47 (7)	21	0	7	3	9	0	1	-0.01	-0.18
	MASH	-0.028	77 (0.04)	133 (-0.07)	0 (155)	12.01	103	38 (18)	0 (0)	51	0	0	10	9	0	1	0	-0.14
	ACMANT2	-0.024	67 (0.03)	134 (-0.05)	9 (148)	8.91	109	26 (22)	22 (9)	22	0	8	2	10	0	0	-0.01	-0.03
	DAP1	-0.018	82 (0.15)	117 (-0.14)	11 (89)	27.96	34	62 (7)	80 (11)	16	0	10	0	9	1	0	0	0.13
	HOM1	-0.019	82 (0.15)	117 (-0.14)	11 (89)	28.46	35	61 (6)	81 (11)	16	0	10	0	9	1	0	0	0.15
	SPLIDHOM1	-0.017	84 (0.145)	115 (-0.137)	89 (11)	27.93	36	62 (5)	80 (11)	16	0	10	0	9	1	0	0	0.14
SEW3	Released	0.013	104 (0.34)	98 (-0.33)	8 (80)	68.33	-	-	-	-	-	-	-	-	-	-	0	0.25
	Climatol-D	0.006	98 (0.03)	100 (-0.02)	12 (180)	5.62	113	7 (6)	71 (8)	5	0	10	0	10	0	0	0	0.01
	Climatol-M	0.012	101 (0.07)	104 (-0.05)	5 (174)	12.14	114	14 (8)	45 (5)	24	1	5	4	8	0	2	-0.01	0.38
	MASH	0.012	117 (0.06)	93 (-0.05)	0 (161)	12.51	115	24 (13)	0 (0)	58	0	0	10	9	0	1	0	0.14
	ACMANT2	0.005	101 (0.04)	104 (-0.03)	5 (160)	7.64	118	15 (11)	31 (5)	30	0	6	4	10	0	0	-0.02	0.07
	DAP1	-0.005	102 (0.15)	102 (-0.16)	6 (103)	30.78	48	64 (9)	69 (6)	14	0	9	1	10	0	0	0	0.08
	HOM1	-0.007	102 (0.16)	102 (-0.17)	6 (106)	34.08	47	55 (5)	80 (6)	17	0	9	1	10	0	0	0	0.07
	SPLIDHOM1	-0.005	103 (0.14)	101 (-0.16)	6 (103)	30.84	48	62 (9)	69 (6)	16	0	9	1	10	0	0	0	0.17

Table B.26. As in table 2, but for the South East scenarios 2 and 3.

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
SEW2	Released	0.552	11 (37)	(0, 0)	(1.28, 2.21)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.118	14 (58)	(0, 0)	(0.02, 0.24)	70	55	2	72 (11)	0	10	0	10	0	0	78.55
	Climatol-M	0.185	7 (45)	(0, 0.07)	(0.04, 1.51)	62	77	17	47 (7)	0	7	3	10	0	0	66.57
	MASH	0.181	0 (28)	(0, 0.07)	(0.04, 1.42)	54	97	59	0 (0)	0	0	10	9	0	1	67.17
	ACMANT2	0.138	9 (40)	(0, 0.08)	(0.05, 0.37)	59	98	22	22 (9)	0	8	2	10	0	0	74.99
	DAP1	0.352	11 (37)	(0, 0)	(0.16, 1.42)	9	97	13	80 (11)	0	10	0	9	1	0	36.31
	HOM1	0.364	11 (37)	(0, 0)	(0.17, 1.42)	8	97	13	81 (11)	0	10	0	9	1	0	34.21
	SPLIDHOM1	0.351	11 (37)	(0, 0)	(0.16, 1.42)	9	98	12	80 (11)	0	10	0	9	1	0	36.49
SEW3	Released	0.643	8 (37)	(0, 0.02)	(1.25, 2.37)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.121	12 (67)	(0, 0.02)	(0.05, 0.20)	95	32	4	71 (8)	0	10	0	10	0	0	81.14
	Climatol-M	0.271	5 (49)	(0, 0.08)	(0.05, 2.87)	87	51	22	45 (5)	0	5	5	8	0	2	57.78
	MASH	0.221	0 (18)	(0.02, 0.13)	(0.12, 1.73)	77	64	69	0 (0)	0	0	10	9	0	1	65.64
	ACMANT2	0.156	5 (48)	(0, 0.16)	(0.12, 0.77)	75	70	29	31 (5)	0	6	4	10	0	0	75.71
	DAP1	0.411	6 (38)	(0, 0.02)	(0.24, 1.63)	17	104	14	69 (6)	0	9	1	10	0	0	35.97
	HOM1	0.459	6 (37)	(0, 0.02)	(0.19, 2.13)	13	95	16	80 (6)	0	9	1	9	0	1	28.36
	SPLIDHOM1	0.410	6 (38)	(0, 0.02)	(0.23, 1.63)	18	104	13	69 (6)	0	9	1	10	0	0	36.20

Table B.27. As in table 3, but for the South East scenarios 2 and 3.

Scenario	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
SEW2	Clean	-0.060	0.201	196	0.056	14	-0.017	4	-	-	-	-	-	-	-	-	-	-	0.051	-
	Released	-0.975	0.722	142	0.157	68	-0.163	91 (1)	-	-	-	-	-	-	-	-	-	-	0.051	-
	Climatol-D	-0.056	0.174	197	0.065	13	-0.017	9 (3)	105	7 (10)	5	72 (11)	0	0	10	10	0	0	0.060	6623.36
	Climatol-M	-0.041	0.206	204	0.068	6	-0.022	9 (3)	105	12 (15)	24	47 (7)	0	3	7	10	0	0	0.065	10618.51
	MASH	-0.031	0.264	203	0.062	7	-0.011	10 (3)	121	33 (17)	39	0 (0)	0	10	0	10	0	0	0.059	5792.2
	ACMANT2	-0.038	0.281	201	0.064	9	-0.013	6 (3)	119	20 (17)	23	22 (9)	0	2	8	10	0	0	0.061	7227.60
	DAP1	-0.435	0.547	155	0.096	55	-0.076	55 (2)	38	63 (5)	14	79 (11)	0	0	10	9	0	1	0.049	-851.33
	HOM1	-0.446	0.547	157	0.094	53	-0.080	56 (2)	38	63 (5)	13	80 (11)	0	0	10	9	0	1	0.049	-884.09
	SPLIDHOM1	-0.433	0.547	155	0.095	55	-0.078	56 (2)	38	63 (5)	14	79 (11)	0	0	10	9	0	1	0.049	-1148.76
SEW3	Clean	-0.067	0.198	189	0.060	21	-0.016	3	-	-	-	-	-	-	-	-	-	-	0.052	-
	Released	-0.676	0.729	127	0.170	83	-0.184	91 (1)	-	-	-	-	-	-	-	-	-	-	0.031	-
	Climatol-D	-0.103	0.200	184	0.058	26	-0.025	4 (3)	112	5 (6)	7	71 (8)	0	0	10	10	0	0	0.048	80.72
	Climatol-M	-0.628	0.313	201	0.058	9	-0.024	5 (3)	116	10 (10)	24	45 (5)	0	5	5	10	0	0	0.054	108.80
	MASH	-0.259	0.178	187	0.055	23	-0.029	5 (3)	120	18 (11)	61	0 (0)	0	10	0	10	0	0	0.047	74.74
	ACMANT2	-0.151	0.216	198	0.055	12	-0.025	2 (2)	122	13 (7)	32	31 (5)	0	4	6	10	0	0	0.051	96.41
	DAP1	-0.558	0.454	149	0.101	61	-0.087	51 (2)	47	65 (5)	18	69 (6)	0	1	9	10	0	0	0.047	74.99
	HOM	-0.581	0.481	148	0.106	62	-0.097	53 (2)	42	60 (8)	14	80 (6)	0	1	9	10	0	0	0.048	78.34
	SPLIDHOM1	-0.520	0.454	148	0.102	62	-0.085	50 (2)	45	64 (6)	20	69 (6)	0	1	9	10	0	0	0.047	75.97

Table B.28. As in table 4, but for the South East scenarios 2 and 3.

Scenario	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
SEW2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.568	0.849
	97	23	7	72 (11)	0	10	0	9	0	1	115	11	1	72 (11)	0	10	0	10	0	0	0.087 (170)	0 (207)	0.962	0.994
	100	34	18	47 (7)	0	7	3	10	0	0	121	21	14	47 (7)	0	7	3	10	0	0	0.056 (138)	0.013 (126)	0.962	0.993
	121	44	45	0 (0)	0	0	10	9	0	1	132	36	42	0 (0)	0	0	10	10	0	0	0.242 (149)	0.029 (149)	0.971	0.993
	106	43	30	22 (9)	0	8	2	9	0	1	122	35	22	22 (9)	0	8	2	10	0	0	0.178 (88)	0.036 (149)	0.957	0.990
	32	73	15	79 (11)	0	10	0	9	1	0	39	73	8	79 (11)	0	10	0	9	1	0	191 (0.195)	104 (0.100)	0.773	0.948
	33	71	15	80 (11)	0	10	0	9	1	0	38	73	8	80 (11)	0	10	0	9	1	0	170 (0.290)	104 (0.054)	0.765	0.945
	34	71	15	79 (11)	0	10	0	9	1	0	40	72	8	79 (11)	0	10	0	9	1	0	170 (0.218)	104 (0.058)	0.772	0.948
SEW3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.439	0.811
	105	23	3	71 (8)	0	10	0	10	0	0	122	6	3	71 (8)	0	10	0	10	0	0	0.010 (82)	0.002 (54)	0.963	0.993
	117	25	15	45 (5)	0	5	5	9	0	1	126	19	15	45 (5)	0	5	5	10	0	0	0.179 (60)	0.022 (153)	0.965	0.992
	119	50	41	0 (0)	0	0	10	9	0	1	127	35	48	0 (0)	0	0	10	10	0	0	0.239 (149)	0.025 (149)	0.961	0.992
	114	29	31	31 (5)	0	6	4	6	0	4	121	30	23	31 (5)	0	6	4	10	0	0	0.374 (76)	0.022 (170)	0.955	0.988
	45	77	13	69 (6)	0	9	1	9	1	0	55	72	8	69 (6)	0	1	9	10	0	0	176 (0.385)	42 (0.053)	0.719	0.932
	42	69	13	80 (6)	0	9	1	8	0	2	47	64	10	80 (6)	0	9	1	10	0	0	176 (0.408)	135 (0.050)	0.686	0.920
	45	76	14	69 (6)	0	9	1	9	0	1	54	71	10	69 (6)	0	9	1	10	0	0	176 (0.376)	135 (0.051)	0.718	0.932

Table B.29. As in table 5, but for the South East scenarios 2 and 3.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged from released (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
SEW2	Released	160	39	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	135	61	3	7	15	46	45	11	83 (11)	0	10	0	10	0	0
	Climatol-M	141	62	2	9	43	35	49	18	54 (7)	0	7	3	7	0	3
	MASH	125	85	1	11	65	24	61	48	0 (0)	0	0	10	8	0	2
	ACMANT2	133	68	1	9	37	38	67	27	31 (9)	0	8	2	9	0	1
	DAP1	133	66	1	1	54	7	23	33	91 (11)	0	10	0	2	1	7
	HOM	135	64	1	4	53	5	25	30	92 (11)	0	10	0	1	1	8
	SPLIDHOM1	130	69	1	2	53	4	26	33	91 (11)	0	10	0	2	1	7
SEW3	Released	157	45	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	116	82	2	4	8	53	50	14	79 (8)	0	10	0	10	0	0
	Climatol-M	114	91	1	6	21	33	67	32	50 (5)	1	5	4	7	0	3
	MASH	127	83	2	9	59	33	61	46	0 (0)	0	0	10	7	0	3
	ACMANT2	121	84	1	9	24	40	70	30	36 (5)	1	6	3	9	0	1
	DAP1	126	78	1	0	56	8	33	37	75 (6)	0	3	7	5	0	5
	HOM	129	75	0	2	60	8	20	34	86 (6)	0	9	1	1	0	9
	SPLIDHOM1	126	78	1	1	56	9	32	36	75 (6)	0	3	7	4	0	6

Table B.30. As in table 6, but for the South East scenarios 2 and 3.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
SEW2	Released	132 (150)	-	-	-	-	-	-	70 (95)	-	-	-	-	-	-
	Climatol-D	146 (184)	1	8	0	3	13	1	92 (145)	5	15	3	17	19	6
	Climatol-M	146 (183)	3	6	1	4	8	5	83 (145)	7	14	3	20	14	7
	MASH	142 (183)	3	6	3	5	4	6	76 (127)	9	13	10	30	6	15
	ACMANT2	153 (186)	5	2	2	7	3	5	89 (139)	7	9	8	27	7	13
	DAP1	137 (165)	4	17	1	2	21	0	84 (118)	15	15	4	24	30	4
	HOM1	137 (162)	6	16	1	3	22	0	74 (113)	14	18	5	24	31	5
	SPLIDHOM1	139 (165)	3	16	2	3	21	0	80 (119)	15	15	5	22	28	6
SEW3	Released	141 (154)	-	-	-	-	-	-	88 (106)	-	-	-	-	-	-
	Climatol-D	167 (197)	2	5	1	1	4	0	114 (163)	8	10	1	15	13	0
	Climatol-M	163 (191)	1	6	2	2	4	4	101 (149)	14	10	4	19	12	2
	MASH	137 (187)	3	2	3	8	1	6	86 (136)	21	4	8	22	9	10
	ACMANT2	159 (199)	2	2	3	1	1	2	100 (155)	8	9	8	17	7	6
	DAP1	155 (173)	7	13	1	3	13	0	93 (130)	22	15	3	15	20	5
	HOM1	147 (169)	5	17	0	3	16	0	93 (127)	19	17	3	18	19	7
	SPLIDHOM1	149 (172)	7	13	1	5	12	0	91 (126)	24	13	3	15	19	10

Table B.31. As in table 7, but for the South East scenarios 2 and 3.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
SEW2	135 (135)	37 (442)	310	611	0.303	0.057	0.376	0.277 (0.277)	62.5%	7.66%	97.9%	64.7%	4.30%	31.3%	44.2%	12.2%
	111 (111)	126 (136)	334	533	0.249	0.194	0.528	0.191 (0.191)	49.5%	7.66%	76.6%	47.9%	6.45%	24.5%	34.1%	14.8%
	97 (97)	240 (292)	348	408	0.218	0.370	0.835	0.132 (0.132)	33.7%	21.7%	40.4%	43.7%	9.32%	18.8%	26.8%	20.9%
	31 (31)	117 (123)	414	531	0.070	0.181	0.328	0.055 (0.055)	12.5%	3.07%	10.6%	20.2%	0.72%	4.17%	12.3%	5.22%
SEW3	155 (155)	29 (33)	288	620	0.350	0.045	0.424	0.326 (0.326)	77.2%	1.22%	98.2%	71.3%	1.59%	29.6%	43.4%	NA
	116 (116)	115 (120)	327	534	0.262	0.177	0.533	0.206 (0.206)	53.3%	4.47%	67.3%	48.5%	5.16%	24.1%	29.5%	NA
	95 (96)	238 (277)	348	410	0.214	0.367	0.844	0.132 (0.133)	36.5%	9.35%	30.9%	34.6%	12.3%	18.9%	25.4%	NA
	36 (36)	127 (133)	407	520	0.081	0.196	0.386	0.063 (0.063)	16.8%	1.22%	16.4%	16.9%	1.59%	6.67%	6.59%	NA

Table B.32. As in table 8, but for the South East scenarios 2 and 3.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
SEW2	151 (151)	25 (30)	294	614	0.339	0.039	0.376	0.331 (0.331)	66.8%	10.7%	100%	72.3%	6.45%	33.3%	47.1%	19.1%
	153 (153)	91 (98)	292	547	0.344	0.143	0.528	0.282 (0.282)	59.8%	16.5%	89.4%	63.0%	12.9%	34.9%	41.3%	25.2%
	186 (188)	168 (207)	259	471	0.418	0.263	0.835	0.285 (0.287)	63.6%	26.4%	85.1%	76.5%	19.7%	41.1%	47.8%	35.7%
	80 (80)	74 (77)	365	565	0.180	0.116	0.328	0.153 (0.153)	32.6%	7.66%	42.6%	40.3%	4.30%	17.7%	23.9%	11.3%
SEW3	169 (169)	16 (20)	274	624	0.381	0.025	0.424	0.372 (0.372)	82.7%	2.44%	98.2%	78.7%	3.17%	33.0%	46.2%	NA
	166 (166)	70 (71)	277	570	0.375	0.109	0.533	0.323 (0.323)	73.1%	8.94%	89.1%	67.6%	9.92%	33.3%	43.9%	NA
	207 (210)	143 (166)	236	496	0.467	0.224	0.844	0.340 (0.343)	79.7%	20.3%	78.2%	75.7%	24.2%	40.7%	56.1%	NA
	88 (87)	77 (81)	355	560	0.199	0.122	0.386	0.168 (0.168)	41.1%	2.85%	45.6%	38.2%	4.37%	16.3%	25.4%	NA

Table B.33. As in table 1, but for the North East scenarios 1 and 2.

Scenario	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
NEW1	Released	-0.097	58 (0.33)	85 (-0.40)	3 (25)	53.47	-	-	-	-	-	-	-	-	-	-	0	0.27
	Climatol-D	-0.042	51 (0.05)	90 (-0.10)	5 (88)	11.37	73	7 (7)	50 (3)	6	0	10	0	10	0	0	0	0
	Climatol-M	-0.032	57 (0.05)	86 (-0.09)	3 (87)	10.65	83	11 (7)	30 (3)	12	1	9	0	10	0	0	0	0.03
	MASH	-0.055	56 (0.04)	100 (-0.10)	0 (69)	11.52	81	28 (11)	0 (0)	26	2	0	8	10	0	0	0	-0.03
	ACMANT2	-0.051	36 (0.04)	108 (-0.08)	2 (77)	10.15	77	24 (8)	19 (2)	16	1	7	2	10	0	0	0	-0.02
	DAP1	-0.070	58 (0.23)	86 (-0.27)	2 (35)	36.89	13	62 (0)	54 (2)	15	0	8	2	10	0	0	-0.01	0.12
	HOM1	-0.097	58 (0.23)	86 (-0.27)	2 (35)	36.96	15	61 (0)	55 (2)	13	0	8	2	10	0	0	0	0.10
	SPLIDHOM1	-0.069	58 (0.23)	86 (-0.27)	2 (36)	36.96	14	61 (0)	55 (2)	14	0	8	2	10	0	0	-0.01	0.13
NEW2	Released	-0.075	83 (0.29)	118 (-0.34)	6 (39)	64.68	-	-	-	-	-	-	-	-	-	-	0	-0.20
	Climatol-D	-0.025	79 (0.06)	122 (-0.08)	6 (112)	14.51	91	14 (6)	73 (6)	17	0	10	0	10	0	0	0	-0.10
	Climatol-M	-0.026	69 (0.03)	132 (-0.06)	6 (148)	10.00	117	17 (12)	34 (6)	21	0	9	1	10	0	0	0	-0.02
	MASH	-0.040	78 (0.03)	129 (-0.08)	0 (128)	13.34	101	50 (17)	0 (0)	39	1	0	9	10	0	0	0	-0.06
	ACMANT2	-0.037	46 (0.04)	160 (0.06)	1 (124)	11.46	102	36 (20)	18 (1)	30	0	2	8	10	0	0	-0.03	0.02

Table B.34. As in table 2, but for the North East scenarios 1 and 2.

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
NEW1	Released	0.694	3 (9)	(0,0.06)	(1.42, 1.80)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.191	5 (18)	(0, 0.06)	(0.06, 0.39)	51	38	2	50 (5)	0	10	0	10	0	0	72.37
	Climatol-M	0.184	3 (16)	(0, 0.06)	(0.09, 0.33)	61	49	3	30 (3)	1	9	0	10	0	0	73.50
	MASH	0.187	0 (3)	(0.03, 0.18)	(0.09, 0.31)	51	71	24	0 (0)	1	0	9	10	0	0	73.11
	ACMANT2	0.174	2 (13)	(0, 0.16)	(0.08, 0.42)	56	61	8	19 (2)	1	7	2	10	0	0	75.99
	DAP1	0.535	2 (9)	(0, 0.08)	(0.18, 1.50)	7	69	14	54 (2)	0	8	2	10	0	0	23.03
	HOM1	0.536	2 (9)	(0, 0.08)	(0.17, 1.49)	7	69	13	55 (2)	0	8	2	10	0	0	22.76
	SPLIDHOM1	0.536	2 (9)	(0, 0.08)	(0.17, 1.50)	7	68	14	55 (2)	0	8	2	10	0	0	22.90
NEW2	Released	0.631	6 (13)	(0, 0.03)	(1.32, 2.40)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.179	6 (20)	(0, 0.03)	(0.03, 0.52)	67	57	4	73 (6)	0	10	0	10	0	0	71.64
	Climatol-M	0.155	6 (16)	(0, 0.13)	(0.04, 0.18)	71	89	7	34 (6)	0	9	1	10	0	0	75.36
	MASH	0.215	0 (5)	(0, 0.13)	(0.08, 0.57)	53	121	33	0 (0)	0	0	10	10	0	0	65.88
	ACMANT2	0.164	1 (13)	(0, 0.10)	(0.09, 0.31)	66	103	19	18 (1)	0	2	8	10	0	0	74.02

Table B.35. As in table 3, but for the North East scenarios 1 and 2.

Scenario	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
NEW1	Clean	-0.088	0.366	145	0.245	1	-0.088	142	-	-	-	-	-	-	-	-	-	-	0.245	-
	Released	-0.655	-0.927	128	0.334	18	-0.239	120 (38)	-	-	-	-	-	-	-	-	-	-	0.264	-
	Climatol-D	-0.037	-0.421	145	0.246	1	-0.037	141 (114)	79	6 (7)	1	50 (3)	0	0	10	10	0	0	0.255	47.67
	Climatol-M	-0.037	0.446	145	0.246	1	-0.037	140 (128)	97	4 (9)	3	30 (3)	1	0	9	10	0	0	0.246	95.46
	MASH	-0.043	0.401	145	0.260	1	-0.043	142 (121)	89	27 (11)	19	0 (0)	3	7	0	10	0	0	0.260	22.55
	ACMANT2	-0.162	0.387	146	0.245	0	-	144 (130)	96	10 (12)	7	19 (2)	1	2	7	10	0	0	0.259	31.09
	DAP1	-0.405	0.690	135	0.289	11	-0.133	119 (56)	18	57 (3)	12	54 (2)	0	2	8	10	0	0	0.257	40.71
	HOM1	-0.405	0.708	135	0.288	11	-0.137	119 (57)	19	57 (3)	10	55 (2)	0	2	8	10	0	0	0.257	37.95
	SPLIDHOM1	-0.405	0.747	135	0.287	11	-0.137	119 (55)	19	54 (3)	13	55 (2)	0	2	8	10	0	0	0.256	41.85
NEW2	Clean	0.086	0.371	207	0.238	0	-	202	-	-	-	-	-	-	-	-	-	-	0.239	-
	Released	-0.466	1.268	188	0.317	19	-0.180	166 (74)	-	-	-	-	-	-	-	-	-	-	0.272	-
	Climatol-D	0.048	0.388	207	0.245	0	-	201 (167)	99	10 (12)	7	73 (6)	0	0	10	10	0	0	0.245	80.38
	Climatol-M	0.032	0.394	207	0.245	0	-	203 (197)	127	8 (18)	14	34 (6)	0	1	9	10	0	0	0.246	79.07
	MASH	0.079	0.412	207	0.248	0	-	205 (191)	124	37 (16)	30	0 (0)	1	9	0	10	0	0	0.249	70.47
	ACMANT2	0.138	0.385	207	0.255	0	-	206 (191)	119	27 (15)	27	18 (1)	0	8	2	10	0	0	0.255	52.69

Table B.36. As in table 4, but for the North East scenarios 1 and 2.

Scenario	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
NEW1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.706	0.876
	54	31	6	50 (5)	0	10	0	10	0	0	79	14	0	50 (3)	0	10	0	10	0	0	0.063 (111)	NA	0.978	0.991
	80	24	9	30 (3)	1	9	0	9	0	1	91	21	1	30 (3)	1	9	0	10	0	0	0.066 (139)	0.001 (126)	0.987	0.993
	84	25	16	19 (2)	1	7	2	9	0	1	94	25	6	19 (2)	1	7	2	10	0	0	0.115 (56)	0.013 (110)	0.990	0.995
	79	44	23	0 (0)	2	0	8	10	0	0	96	36	14	0 (0)	2	0	8	10	0	0	0.032 (106)	0.013 (41)	0.989	0.994
	14	64	12	54 (2)	0	8	2	10	0	0	18	64	8	54 (2)	0	8	2	10	0	0	1.265 (101)	0.322 (101)	0.823	0.926
	13	64	12	55 (2)	0	8	2	10	0	0	19	62	8	55 (2)	0	8	2	10	0	0	1.268 (101)	0.333 (101)	0.821	0.935
	14	64	11	55 (2)	0	8	2	10	0	0	18	62	9	55 (2)	0	8	2	10	0	0	1.266 (101)	0.326 (101)	0.821	0.935
NEW2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.764	0.899
	85	30	13	73 (6)	0	10	0	8	0	2	106	21	1	73 (6)	0	10	0	10	0	0	0.401 (109)	0 (91)	0.977	0.992
	118	36	13	34 (6)	0	9	1	9	0	1	136	27	4	34 (6)	0	9	1	10	0	0	0.038 (175)	0.004 (111)	0.990	0.995
	123	53	31	0 (0)	0	0	10	9	0	1	144	37	26	0 (0)	0	0	10	10	0	0	0.098 (199)	0.012 (56)	0.991	0.995
	133	30	25	18 (1)	1	2	7	9	0	1	133	35	20	18 (1)	0	2	8	10	0	0	0.070 (126)	0.024 (145)	0.990	0.995

Table B.37. As in table 5, but for the North East scenarios 1 and 2.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
NEW1	Released	91	52	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	69	72	2	14	8	24	27	18	53 (3)	0	10	0	7	0	3
	Climatol-M	67	76	5	19	16	23	24	26	33 (3)	0	9	1	7	0	3
	MASH	95	51	5	20	51	19	28	23	0 (0)	0	0	10	7	0	3
	ACMANT2	76	68	2	18	19	29	30	27	21 (2)	0	7	3	7	0	3
	DAP1	80	64	0	14	33	5	9	29	56 (2)	0	8	2	3	0	7
	HOM1	84	60	2	11	30	3	15	28	57	0	8	2	3	0	7
	SPLIDHOM1	81	63	1	13	32	2	13	28	57	0	8	2	4	0	6
NEW2	Released	116	85	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	95	106	3	24	16	28	32	25	79 (6)	0	10	0	9	0	1
	Climatol-M	84	117	12	23	19	30	35	48	40 (6)	1	9	0	6	0	4
	MASH	110	97	17	27	50	34	44	35	0 (0)	1	0	9	7	0	3
	ACMANT2	100	106	3	26	38	21	44	56	19	0	1	9	7	0	3

Table B.38. As in table 6, but for the North East scenarios 1 and 2.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
NEW1	Released	68 (80)	-	-	-	-	-	-	38 (56)	-	-	-	-	-	-
	Climatol-D	77 (101)	5	7	1	10	20	2	63 (104)	3	6	1	13	18	1
	Climatol-M	84 (113)	4	7	2	10	8	2	64 (116)	4	4	1	10	10	1
	MASH	75 (114)	6	1	2	16	4	3	52 (100)	3	1	3	23	9	7
	ACMANT2	77 (116)	6	2	0	12	7	3	58 (117)	3	1	1	12	8	4
	DAP1	75 (96)	4	13	2	7	23	1	44 (70)	13	12	7	12	28	4
	HOM1	77 (96)	6	13	1	6	24	0	45 (73)	10	12	5	13	27	6
	SPLIDHOM1	76 (93)	6	13	1	10	23	0	43 (70)	11	11	7	13	28	6
NEW2	Released	106 (123)	-	-	-	-	-	-	59 (98)	-	-	-	-	-	-
	Climatol-D	122 (159)	1	8	3	10	26	0	77 (149)	2	12	2	17	22	3
	Climatol-M	122 (172)	5	3	2	10	14	1	90 (160)	7	6	2	16	10	6
	MASH	100 (168)	6	2	5	15	4	7	64 (138)	11	6	4	22	6	20
	ACMANT2	119 (174)	6	1	2	14	8	2	74 (156)	0	4	3	20	6	18

Table B.39. As in table 7, but for the North East scenarios 1 and 2.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
NEW1	94 (94)	28 (29)	357	564	0.208	0.047	0.383	0.196 (0.196)	43.7%	10.4%	83.7%	39.4%	1.09%	18.8%	26.2%	15.2%
	94 (94)	88 (93)	357	504	0.208	0.149	0.402	0.173 (0.173)	35.9%	13.9%	69.4%	42.5%	2.18%	17.3%	26.2%	18.5%
	75 (76)	188 (215)	376	404	0.166	0.318	0.624	0.113 (0.114)	24.6%	12.9%	38.8%	33.1%	5.09%	13.6%	12.3%	17.4%
	1 (2)	94 (113)	450	498	0.002	0.159	0.248	0.002 (0.004)	0%	0.32%	0%	0.79%	0%	0%	0.60%	0%
NEW2	153 (153)	40 (42)	431	743	0.262	0.051	0.324	0.244 (0.244)	55.4%	12.8%	90.6%	52.1%	2.04%	18.2%	41.0%	17.2%
	139 (139)	143 (152)	445	640	0.238	0.183	0.487	0.189 (0.189)	44.0%	14.5%	71.7%	42.6%	6.12%	19.4%	33.3%	16.4%
	143 (147)	244 (274)	441	539	0.245	0.312	0.703	0.167 (0.171)	37.0%	18.8%	45.3%	43.1%	11.1%	22.5%	31.4%	16.4%

Table B.40. As in table 8, but for the North East scenarios 1 and 2.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
NEW1	106 (106)	19 (20)	345	560	0.235	0.033	0.266	0.225 (0.225)	47.2%	12.6%	91.8%	42.5%	2.55%	20.9%	29.2%	18.5%
	129 (129)	58 (62)	322	521	0.286	0.100	0.402	0.251 (0.251)	48.6%	19.4%	89.8%	55.1%	5.45%	23.6%	35.7%	26.1%
	152 (153)	126 (141)	299	453	0.337	0.218	0.624	0.257 (0.258)	46.5%	27.8%	75.6%	65.4%	10.9%	24.6%	42.9%	11.3%
	5 (7)	90 (108)	446	489	0.011	0.155	0.248	0.009 (0.012)	1.41%	0.97%	2.04%	1.57%	0.73%	1.05%	1.19%	1.09%
NEW2	176 (176)	19 (21)	408	748	0.301	0.025	0.324	0.291 (0.291)	60.9%	16.0%	96.2%	58.5%	4.37%	21.3%	46.2%	20.7%
	199 (200)	87 (93)	385	680	0.341	0.113	0.487	0.294 (0.295)	57.6%	23.3%	81.1%	62.8%	11.1%	27.9%	46.2%	25.9%
	250 (254)	151 (170)	334	616	0.428	0.197	0.703	0.332 (0.335)	59.2%	35.3%	83.0%	93.1%	22.2%	38.0%	52.9%	35.3%

Table B.41. As in table 1, but for the North East scenario 3 and South West scenario 1.

Scenario	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
NEW3	Released	-0.024	96 (0.38)	99 (-0.42)	12 (54)	78.13	-	-	-	-	-	-	-	-	-	-	0	-0.11
	Climatol-D	-0.001	87 (0.03)	107 (-0.03)	13 (134)	12.15	104	12 (8)	64 (12)	7	0	10	0	10	0	0	0	-0.01
	Climatol-M	-0.007	109 (0.04)	87 (-0.06)	11 (157)	9.53	127	12 (8)	34 (11)	15	0	9	1	10	0	0	-0.01	0.01
	MASH	0.001	118 (0.05)	89 (-0.06)	0 (140)	11.35	124	28 (16)	1 (0)	38	0	0	10	10	0	0	0.02	-0.05
	ACMANT2	-0.004	91 (0.04)	106 (-0.04)	10 (162)	7.98	128	22 (13)	11 (10)	23	0	8	2	10	0	0	-0.01	0
SWW1	Released	0.015	71 (0.31)	70 (-0.28)	10 (53)	41.64	-	-	-	-	-	-	-	-	-	-	0	0.25
	Climatol-D	-0.018	64 (0.05)	77 (-0.08)	10 (98)	9.27	69	6 (3)	55 (10)	8	0	10	0	10	0	0	0	-0.01
	Climatol-M	0.004	68 (0.14)	76 (-0.12)	7 (72)	18.34	58	11 (12)	33 (7)	30	0	7	3	8	0	2	0.04	-0.04
	MASH	-0.017	65 (0.08)	86 (-0.09)	0 (87)	13.33	70	20 (14)	1 (0)	46	0	0	10	8	0	2	-0.02	-0.07
	ACMANT2	-0.050	38 (0.06)	110 (-0.09)	3 (69)	11.70	62	16 (13)	17 (3)	40	0	4	6	10	0	0	-0.02	-0.07
	DAP1	0.011	70 (0.24)	71 (-0.21)	10 (60)	32.62	10	37 (3)	78 (10)	13	0	10	0	6	3	1	0	.29
	HOM1	0.008	69 (0.24)	72 (-0.22)	10 (60)	32.55	9	38 (2)	78 (10)	14	0	10	0	6	3	1	0	0.28
	SPLIDHOM1	0.010	71 (0.24)	70 (-0.22)	10 (60)	32.59	10	36 (3)	78 (10)	14	0	10	0	6	3	1	0	0.29

Table B.42. As in table 2, but for the North East scenario 3 and South West scenario 1.

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
NEW3	Released	0.733	12 (20)	(0, 0)	(1.61, 2.17)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.169	13 (41)	(0, 0)	(0.06, 0.34)	88	42	1	64 (12)	0	10	0	10	0	0	77.01
	Climatol-M	0.190	11 (31)	(0, 0.09)	(0.09, 0.34)	91	62	9	34 (11)	0	9	1	10	0	0	74.10
	MASH	0.168	0 (8)	(0.01, 0.17)	(0.11, 0.68)	75	94	37	1 (0)	0	0	10	10	0	0	77.03
	ACMANT2	0.165	23 (10)	(0, 0.09)	(0.07, 0.71)	86	82	18	11 (10)	0	8	2	10	0	0	77.50
SWW1	Released	0.576	10 (18)	(0, 0.02)	(1.13, 1.92)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.196	10 (20)	(0, 0.02)	(0.07, 0.28)	39	42	5	55 (10)	0	10	0	10	0	0	66.04
	Climatol-M	0.310	7 (13)	(0, 0.25)	(0.09, 1.66)	29	56	26	33 (7)	0	7	3	8	0	2	46.09
	MASH	0.276	0 (8)	(0.04, 0.15)	(0.08, 1.73)	38	72	40	1 (0)	0	0	10	8	0	2	52.06
	ACMANT2	0.195	3 (8)	(0, 0.24)	(0.07, 0.81)	36	60	35	17 (3)	0	4	6	10	0	0	66.21
	DAP1	0.491	10 (16)	(0, 0.02)	(0.44, 1.92)	1	45	17	78 (10)	0	10	0	6	3	1	14.79
	HOM1	0.492	10 (16)	(0, 0.02)	(0.46, 1.92)	1	45	17	78 (10)	0	10	0	6	3	1	14.57
	SPLIDHOM1	0.491	10 (16)	(0, 0.02)	(0.44, 1.92)	1	45	17	78 (10)	0	10	0	6	3	1	14.80

Table B.43. As in table 3, but for the North East scenario 3 and South West scenario 1.

Scen.	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
NEW3	Clean	0.141	0.340	207	0.236	0	-	203	-	-	-	-	-	-	-	-	-	-	0.236	-
	Rel	-0.738	1.145	174	0.314	33	-0.207	163 (73)	-	-	-	-	-	-	-	-	-	-	0.231	-
	CD	0.020	0.345	207	0.239	0	-	201 (188)	113	4 (9)	5	64 (12)	0	0	10	10	0	0	0.239	152.20
	CM	0.131	0.340	207	0.235	0	-	201 (199)	131	6 (13)	12	34 (11)	0	1	9	10	0	0	0.235	78.56
	MASH	0.131	0.404	207	0.237	0	-	203 (196)	136	24 (15)	31	1 (0)	0	10	0	10	0	0	0.236	103.14
	A2	0.164	0.419	207	0.242	0	-	204 (198)	134	16 (14)	22	11 (10)	0	2	8	10	0	0	0.241	183.79
SWW1	Clean	-0.107	0.272	123	0.135	28	-0.051	68	-	-	-	-	-	-	-	-	-	-	0.098	-
	Rel	-0.759	0.800	101	0.221	50	-0.191	94 (27)	-	-	-	-	-	-	-	-	-	-	0.083	-
	CD	-0.156	0.405	124	0.141	27	-0.047	73 (59)	69	4 (7)	6	55 (10)	0	0	10	10	0	0	0.105	144.11
	CM	-0.118	0.405	136	0.121	15	-0.037	66 (48)	65	6 (20)	20	33 (7)	0	3	7	9	1	0	0.103	128.70
	MASH	-0.103	0.290	122	0.139	29	-0.043	68 (61)	83	23 (8)	36	1 (0)	0	10	0	10	0	0	0.103	129.04
	A2	-0.064	0.297	131	0.143	20	-0.022	81 (56)	73	14 (17)	27	17 (3)	0	6	4	10	0	0	0.119	234.56
	D1	-0.678	0.655	109	0.185	42	-0.154	87 (30)	12	41 (1)	9	78 (10)	0	0	10	7	0	3	0.091	49.32
	H1	-0.678	0.655	109	0.187	42	-0.157	88 (30)	13	41 (1)	8	78 (10)	0	0	10	7	0	3	0.093	61.83
	S1	-0.678	0.655	109	0.187	43	-0.150	88 (30)	12	41 (1)	9	78 (10)	039.94	0	10	7	0	3	0.091	49.94

Table B.44. As in table 4, but for the North East scenario 3 and South West scenario 1.

Scenario	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
NEW3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.664	0.867
	110	17	3	64 (13)	0	10	0	10	0	0	126	5	0	64 (12)	0	10	0	10	0	0	0.026 (140)	-	0.983	0.994
	123	28	11	34 (11)	0	9	1	10	0	0	139	18	5	34 (11)	0	9	1	10	0	0	0.029 (140)	0.031 (192)	0.992	0.996
	137	47	22	1 (0)	0	0	10	10	0	0	146	38	22	1 (0)	0	0	10	10	0	0	0.072 (191)	0.005 (21)	0.991	0.995
	139	22	25	11 (10)	0	8	2	9	0	1	139	30	17	11 (10)	0	8	2	10	0	0	0.299 (30)	0.005 (21)	0.990	0.993
SWW1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.698	0.858
	53	21	12	55 (10)	0	10	0	9	0	1	69	11	6	55 (10)	0	10	0	10	0	0	0.101 (61)	0.131 (32)	0.967	0.834
	48	29	34	33 (7)	0	7	3	7	0	3	63	26	22	33 (7)	0	7	3	9	0	1	0.695 (99)	0.075 (94)	0.947	0.976
	80	31	39	1 (0)	0	0	10	10	0	0	90	26	34	1 (0)	0	0	10	10	0	0	0.844 (128)	0.043 (128)	0.979	0.991
	66	30	35	17 (3)	0	4	6	9	0	1	80	20	31	17 (3)	0	4	6	10	0	0	0.686 (99)	0.049 (60)	0.969	0.986
	13	38	12	78 (10)	0	10	0	7	0	3	15	41	7	78 (10)	0	10	0	6	3	1	1.070 (148)	0.194 (148)	0.794	0.908
	13	38	12	78 (10)	0	10	0	7	3	0	15	40	8	78 (10)	0	10	0	6	3	1	1.093 (148)	0.203 (148)	0.795	0.908
	13	38	12	78 (10)	0	10	0	7	3	0	15	40	8	78 (10)	0	10	0	6	3	1	1.049 (148)	0.185 (148)	0.794	0.908

Table B.45. As in table 5, but for the North East scenario 3 and South West scenario 1.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
NEW3	Released	117	78	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	97	97	4	18	17	37	30	24	76 (12)	0	10	0	9	0	1
	Climatol-M	91	105	8	22	23	32	33	44	45 (11)	0	9	1	8	0	2
	MASH	110	97	15	29	48	24	49	41	1 (0)	0	0	10	8	0	2
	ACMANT2	103	94	2	31	38	32	44	39	21 (10)	0	8	2	8	0	2
SWW1	Released	83	58	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	61	80	0	10	3	16	32	25	65 (10)	0	10	0	9	0	1
	Climatol-M	55	89	1	13	14	15	32	36	40 (7)	0	7	3	9	0	1
	MASH	62	89	5	25	24	21	30	45	1 (0)	0	0	10	9	0	1
	ACMANT2	69	79	0	17	20	18	42	34	20 (3)	0	4	6	7	0	3
	DAP1	74	67	1	6	21	4	12	19	88 (10)	0	10	0	4	3	3
	HOM	78	63	1	7	23	5	11	16	88 (10)	0	10	0	4	3	3
	SPLIDHOM1	75	66	2	6	22	5	10	18	88 (10)	0	10	0	5	3	2

Table B.46. As in table 6, but for the North East scenario 3 and South West scenario 1.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
NEW3	Released	101 (119)	-	-	-	-	-	-	70 (95)	-	-	-	-	-	-
	Climatol-D	122 (151)	8	15	3	10	19	1	106 (167)	8	9	2	7	11	3
	Climatol-M	122 (167)	7	9	4	9	8	2	94 (168)	14	5	1	10	4	5
	MASH	105 (168)	12	5	6	9	5	2	72 (147)	24	2	5	22	3	4
	ACMANT2	125 (169)	9	6	4	12	5	2	94 (175)	8	3	3	10	4	4
SWW1	Released	103 (127)	-	-	-	-	-	-	70 (94)	-	-	-	-	-	-
	Climatol-D	105 (134)	0	4	4	2	7	0	81 (116)	6	5	5	5	11	3
	Climatol-M	105 (131)	1	6	2	2	7	2	72 (109)	5	5	15	5	8	4
	MASH	97 (119)	1	5	1	9	4	12	68 (120)	6	2	6	7	3	7
	ACMANT2	96 (123)	1	2	5	6	4	10	68 (108)	6	1	3	17	3	13
	DAP1	104 (128)	0	10	1	1	11	0	72 (99)	3	19	0	3	25	2
	HOM1	103 (128)	0	10	1	0	11	1	74 (100)	1	19	0	4	25	2
	SPLIDHOM1	103 (128)	0	10	1	0	11	1	72 (99)	1	19	0	4	26	2

Table B.47. As in table 7, but for the North East scenario 3 and South West scenario 1.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
NEW3	188 (188)	20 (20)	353	723	0.348	0.027	0.383	0.335 (0.335)	74.2%	12.7%	93.4%	59.6%	2.84%	27.2%	42.7%	NA
	155 (155)	119 (120)	386	624	0.287	0.160	0.506	0.234 (0.234)	54.1%	14.4%	69.7%	49.9%	6.38%	21.9%	35.9%	NA
	131 (132)	214 (242)	410	528	0.242	0.288	0.693	0.167 (0.168)	43.8%	13.3%	40.8%	39.9%	9.57%	20.4%	28.2%	NA
SWW1	78 (78)	50 (89)	287	461	0.214	0.098	0.448	0.172 (0.172)	40.4%	10.0%	75.6%	37.0%	0%	22.5%	26.8%	9.64%
	62 (62)	102 (108)	303	408	0.170	0.200	0.555	0.131 (0.131)	31.6%	8.30%	56.1%	27.6%	2.03%	17.8%	20.9%	8.43%
	77 (77)	170 (200)	288	336	0.211	0.336	0.757	0.136 (0.136)	28.7%	16.6%	46.3%	36.2%	6.09%	24.0%	22.9%	13.3%
	18 (18)	63 (67)	347	449	0.049	0.123	0.224	0.042 (0.042)	8.08%	3.06%	19.5%	7.09%	0.51%	4.65%	4.58%	6.02%

Table B.48. As in table 8, but for the North East scenario 3 and South West scenario 1.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
NEW3	198 (198)	10 (10)	343	718	0.366	0.014	0.383	0.359 (0.359)	75.8%	14.7%	93.4%	63.4%	3.90%	28.7%	45.0%	NA
	211 (211)	64 (64)	330	664	0.390	0.088	0.506	0.349 (0.349)	72.2%	20.5%	85.5%	61.7%	11.7%	32.6%	45.8%	NA
	256 (258)	107 (118)	285	620	0.473	0.147	0.693	0.388 (0.390)	79.4%	29.4%	82.9%	71.0%	22.3%	43.7%	51.1%	NA
SWW1	100 (100)	31 (68)	265	465	0.274	0.063	0.448	0.231 (0.231)	50.7%	13.5%	90.2%	47.2%	1.52%	29.5%	32.7%	14.5%
	95 (95)	69 (77)	270	425	0.260	0.140	0.456	0.215 (0.215)	45.6%	14.4%	63.4%	48.0%	4.06%	25.6%	32.7%	14.5%
	145 (145)	110 (133)	220	379	0.397	0.225	0.757	0.291 (0.291)	60.3%	27.5%	85.4%	71.7%	9.64%	38.0%	49.0%	25.3%
	39 (39)	42 (45)	326	455	0.107	0.085	0.224	0.095 (0.095)	19.9%	5.24%	36.6%	15.7%	2.03%	10.1%	11.8%	9.64%

Table B.49. As in table 1, but for the South West scenarios 2 and 3.

Scen.	Algorithm	Region bias	No. positively biased (mean bias)	No. negatively biased (mean bias)	No. non-biased (to measurement precision)	Sum absolute biases	Percentage recovery				PR best stats			PR worst stats			Best stats mean bias	Worst stats mean bias
							GI	I	U	MW	I	U	MW	I	U	MW		
SWW2	Released	-0.061	86 (0.33)	121 (-0.35)	15 (65)	71.03	-	-	-	-	-	-	-	-	-	-	0	-0.05
	Climatol-D	-0.027	86 (0.04)	120 (-0.08)	16 (149)	12.32	104	14 (11)	71 (15)	7	0	10	0	10	0	0	0	-0.05
	Climatol-M	-0.057	70 (0.10)	146 (-0.13)	6 (99)	26.13	80	23 (22)	45 (6)	46	0	4	6	9	0	1	0.01	-0.30
	MASH	-0.053	86 (0.05)	136 (-0.12)	0 (133)	19.81	97	33 (21)	0 (0)	71	0	0	10	9	0	1	-0.01	-0.20
	ACMANT2	-0.061	39 (0.07)	175 (-0.09)	8 (98)	19.14	101	28 (14)	17 (8)	54	0	6	4	10	0	0	-0.04	-0.06
	DAP1	-0.058	92 (0.19)	117 (-0.26)	13 (82)	47.36	29	59 (5)	98 (13)	18	0	8	2	9	1	0	0.01	-0.22
	HOM1	-0.063	89 (0.19)	120 (-0.26)	13 (82)	48.23	27	60 (6)	98 (13)	18	0	8	2	9	1	0	0.01	-0.23
	SPLIDHOM1	-0.057	92 (0.18)	117 (-0.26)	13 (82)	47.49	29	60 (5)	98 (13)	17	0	8	2	9	1	0	0.01	-0.22
SWW3	Released	-0.003	111 (0.39)	102 (-0.43)	9 (65)	87.72	-	-	-	-	-	-	-	-	-	-	0	0.22
	Climatol-D	0.027	145 (0.06)	66 (-0.05)	11 (155)	12.66	116	10 (11)	69 (9)	7	0	10	0	10	0	0	0	0.01
	Climatol-M	-0.007	121 (0.09)	94 (-0.14)	7 (112)	24.44	100	21 (16)	41 (7)	37	0	9	1	10	0	0	-0.01	-0.03
	MASH	0.016	123 (0.07)	98 (-0.05)	1 (140)	14.40	119	36 (19)	0 (1)	47	0	1	9	10	0	0	-0.01	0.04
	ACMANT2	-0.020	104 (0.07)	116 (-0.10)	2 (102)	19.03	110	24 (14)	17 (2)	55	0	2	8	10	0	0	-0.05	-0.03
	DAP1	0.012	114 (0.28)	99 (-0.30)	9 (77)	61.87	27	59 (5)	105 (9)	17	0	10	0	8	1	1	0	0.57
	HOM1	0.007	112 (0.29)	101 (-0.30)	9 (78)	62.71	25	61 (4)	106 (9)	17	0	10	0	8	1	1	0	0.49
	SPLIDHOM1	0.012	114 (0.28)	99 (-0.30)	9 (78)	61.97	26	61 (4)	105 (9)	17	0	0.56	0	10	0	8	1	1

Table B.50. As in table 2, but for the South West scenarios 2 and 3.

Scenario	Algorithm	Region RMSE	No. perfect RMSEs (to measurement precision)	Range of RMSEs in best stats	Range of RMSEs in worst stats	Percentage recovery				PR best stats			PR worst stats			PR value for region
						GI	I	MW	U	I	U	MW	I	U	MW	
SWW2	Released	0.658	20 (13)	(0, 0)	(1.36, 3.62)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.164	28 (16)	(0, 0)	(0.05, 0.31)	80	58	0	71 (13)	0	10	0	10	0	0	75.15
	Climatol-M	0.289	6 (12)	(0, 0.22)	(0.11, 1.77)	60	73	38	45 (6)	0	4	6	9	0	1	56.56
	MASH	0.254	0 (6)	(0.03, 0.12)	(0.08, 1.47)	66	108	48	0 (0)	0	0	10	10	0	0	61.43
	ACMANT2	0.207	8 (11)	(0, 0.25)	(0.11, 0.64)	65	90	42	17 (8)	0	6	4	10	0	0	68.49
	DAP1	0.526	13 (18)	(0, 0.21)	(0.16, 3.02)	5	86	20	98 (13)	0	8	2	9	1	0	20.07
	HOM1	0.529	13 (18)	(0, 0.16)	(0.17, 2.95)	4	87	20	98 (13)	0	8	2	9	1	0	19.61
	SPLIDHOM1	0.525	13 (18)	(0, 0.19)	(0.16, 3.01)	5	86	20	98 (13)	0	8	2	9	1	0	20.14
SWW3	Released	0.796	9 (22)	(0, 0.01)	(1.57, 3.88)	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	0.182	11 (37)	(0, 0.01)	(0.07, 0.38)	95	47	2	69 (9)	0	10	0	10	0	0	77.16
	Climatol-M	0.259	7 (28)	(0, 0.09)	(0.09, 0.65)	76	73	25	41 (7)	0	9	1	10	0	0	67.47
	MASH	0.204	0 (7)	(0, 0.17)	(0.10, 0.29)	85	84	53	0 (0)	0	0	10	10	0	0	74.42
	ACMANT2	0.251	2 (13)	(0, 0.33)	(0.14, 0.39)	82	74	47	17 (2)	0	2	8	10	0	0	68.47
	DAP1	0.610	9 (20)	(0, 0.01)	(0.69, 3.12)	9	84	15	105 (9)	0	10	0	8	1	1	23.35
	HOM1	0.629	9 (20)	(0, 0.01)	(0.77, 3.12)	6	84	17	106 (9)	0	10	0	8	1	1	20.99
	SPLIDHOM1	0.610	9 (20)	(0, 0.01)	(0.73, 3.13)	9	85	14	105 (9)	0	10	0	8	1	1	23.32

Table B.51. As in table 3, but for the South West scenarios 2 and 3.

Scenario	Algorithm	Decadal trends (°C)		Positive trends		Negative trends		No. of signif. trends (pre-served)	Percentage recovery				PR best stats			PR worst stats			Regional average trend	PR for regional average trend
		Min	Max	No.	Mean	No.	Mean		GI	I	MW	U	I	MW	U	I	MW	U		
SWW2	Clean	-0.137	0.291	180	0.136	42	-0.047	102	-	-	-	-	-	-	-	-	-	-	0.099	-
	Released	-0.790	1.107	160	0.251	62	-0.184	153 (35)	-	-	-	-	-	-	-	-	-	-	0.126	-
	Climatol-D	-0.196	0.319	182	0.142	40	-0.048	117 (92)	118	7 (10)	3	71 (13)	0	0	10	10	0	0	0.106	75.30
	Climatol-M	-0.075	0.678	215	0.131	7	-0.030	117 (87)	100	16 (20)	35	45 (6)	0	6	4	10	0	0	0.123	12.89
	MASH	-0.097	0.300	180	0.145	42	-0.038	115 (95)	119	33 (20)	50	0 (0)	0	10	0	10	0	0	0.108	67.89
	ACMANT2	-0.248	0.424	204	0.140	18	-0.031	130 (87)	112	18 (20)	47	17 (8)	0	4	6	10	0	0	0.125	5.79
	DAP1	-0.435	1.107	170	0.202	52	-0.121	135 (44)	28	57 (7)	19	98 (13)	0	2	8	8	1	1	0.124	10.32
	HOM1	-0.438	1.107	171	0.205	51	-0.122	136 (43)	25	61 (7)	18	98 (13)	0	2	8	8	1	1	0.127	-1.89
	SPLIDHOM1	-0.442	1.107	169	0.204	53	-0.120	135 (44)	27	60 (7)	17	98 (13)	0	2	8	8	1	1	0.124	9.54
SWW3	Clean	-0.128	0.283	179	0.135	43	-0.047	104	-	-	-	-	-	-	-	-	-	-	0.098	-
	Released	-1.271	0.967	143	0.265	79	-0.223	146 (45)	-	-	-	-	-	-	-	-	-	-	0.090	-
	Climatol-D	-0.251	0.314	175	0.127	47	-0.053	93 (98)	122	8 (6)	6	69 (11)	0	10	0	10	0	0	0.087	-30.00
	Climatol-M	-0.097	0.380	199	0.114	23	-0.034	88 (76)	112	15 (20)	27	41 (7)	0	9	1	10	0	0	0.096	77.72
	MASH	-0.095	0.321	175	0.129	47	-0.039	90 (90)	137	30 (11)	44	0 (0)	0	0	10	10	0	0	0.092	25.95
	ACMANT2	-0.055	0.552	184	0.131	38	-0.018	92 (85)	116	22 (13)	52	17 (2)	0	2	8	10	0	0	0.104	168.02
	DAP1	-1.046	0.828	149	0.203	73	-0.141	121 (43)	32	58 (7)	11	105 (9)	0	10	0	9	1	0	0.088	-17.37
	HOM1	-1.045	0.835	149	0.208	73	-0.144	123 (41)	31	58 (7)	11	106 (9)	0	10	0	9	1	0	0.090	9.20
	SPLIDHOM1	-1.048	0.827	149	0.204	26	-0.142	122 (42)	29	62 (6)	11	105 (9)	0	10	0	9	1	0	0.088	-17.96

Table B.52. As in table 4, but for the South West scenarios 2 and 3.

Scenario	PR for inter-decadal smooths				PR for inter-decadal best			PR for inter-decadal worst			PR for inter-annual smooths				PR for inter-annual best			PR for inter-annual worst			Biggest correlation decrease (station ID)		Overall correlation for region	
	GI	I	MW	U	I	U	MW	I	U	MW	GI	I	MW	U	I	U	MW	I	U	MW	D	Y	D	Y
SWW2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.694	0.842
	88	36	11	71 (16)	0	10	0	7	0	3	123	13	0	71 (15)	0	10	0	10	0	0	0.037 (169)	-	0.975	0.989
	81	50	40	45 (6)	0	4	6	6	0	4	108	36	27	45 (6)	0	4	6	10	0	0	0.932 (153)	0.064 (153)	0.958	0.981
	126	47	49	0 (0)	0	0	10	9	0	1	148	28	47	0 (0)	0	0	10	10	0	0	0.301 (153)	0.015 (222)	0.984	0.992
	102	47	48	17 (8)	0	6	4	9	0	1	124	35	38	17 (8)	0	6	4	10	0	0	0.704 (153)	0.074 (222)	0.967	0.985
	27	64	20	98 (13)	0	8	2	5	1	4	33	65	13	98 (13)	0	8	2	8	1	1	0.470 (16)	0.024 (21)	0.820	0.912
	26	67	18	98 (13)	0	8	2	5	1	4	30	69	12	98 (13)	0	8	2	8	1	1	0.488 (16)	0.017 (14)	0.818	0.910
	27	65	19	98 (13)	0	8	2	5	1	4	33	65	13	98 (13)	0	8	2	8	1	1	0.486 (16)	0.021 (21)	0.820	0.912
SWW3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.607	0.795
	103	24	15	69 (11)	0	10	0	8	0	2	131	11	2	69 (9)	0	10	0	10	0	0	0.271 (222)	0.007 (197)	0.981	0.990
	95	26	53	41 (7)	0	9	1	7	0	3	118	27	29	41 (7)	0	9	1	9	0	1	0.897 (111)	0.065 (152)	0.957	0.981
	127	42	53	0 (0)	0	0	10	10	0	0	144	39	39	0 (0)	0	0	10	10	0	0	0.340 (95)	0.029 (95)	0.983	0.991
	108	32	63	17 (2)	0	2	8	8	0	2	123	33	47	17 (2)	0	2	8	10	0	0	0.679 (111)	0.060 (173)	0.970	0.984
	14	70	24	105 (9)	0	10	0	6	1	3	29	70	9	105 (9)	0	10	0	9	1	0	1.130 (39)	0.369 (193)	0.751	0.883
	14	70	23	106 (9)	0	10	0	8	1	1	27	72	8	106 (9)	0	10	0	9	1	0	1.135 (39)	0.430 (193)	0.744	0.879
	14	72	22	105 (9)	0	10	0	6	1	3	29	71	8	105 (9)	0	10	0	9	1	0	1.138 (39)	0.317 (193)	0.750	0.882

Table B.53. As in table 5, but for the South West scenarios 2 and 3.

Scenario	Algorithm	No. stats more variable than clean	No. stats less variable than clean	Variability increases			Variability decreases			Variability unchanged (because of perfection)	PR best stats			PR worst stats		
				GI	I	MW	GI	I	MW		I	U	MW	I	U	MW
SWW2	Released	132	75	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	97	109	1	10	7	25	52	41	86 (15)	0	10	0	9	0	1
	Climatol-M	101	115	12	13	28	22	44	52	51 (6)	0	4	6	7	0	3
	MASH	99	123	4	29	29	26	66	68	0 (0)	0	0	10	8	0	2
	ACMANT2	103	111	0	24	28	26	62	57	25 (8)	0	6	4	9	0	1
	DAP1	123	86	2	4	44	10	18	33	111 (13)	0	8	2	3	1	6
	HOM1	125	84	1	8	43	7	22	30	111 (13)	0	8	2	5	1	4
	SPLIDHOM1	124	85	3	5	43	9	20	31	111 (13)	0	8	2	5	1	4
SWW3	Released	151	62	-	-	-	-	-	-	-	-	-	-	-	-	-
	Climatol-D	104	107	3	5	7	44	52	33	78 (9)	0	10	0	9	0	1
	Climatol-M	102	113	3	22	24	33	41	48	48 (7)	0	9	1	7	0	3
	MASH	130	92	9	21	59	27	61	45	0 (0)	0	0	10	8	0	2
	ACMANT2	103	117	1	16	24	33	63	49	19 (2)	1	2	7	9	0	1
	DAP1	137	76	0	5	43	9	33	18	114 (9)	0	10	0	4	1	5
	HOM	137	76	0	5	41	8	30	23	115 (9)	0	10	0	3	1	6
	SPLIDHOM1	138	75	0	5	45	8	33	17	114 (9)	0	10	0	2	1	7

Table B.54. As in table 6, but for the South West scenarios 2 and 3.

Scenario	Algorithm	Warm extremes							Cold extremes						
		Exact (± 0.14)	Too warm in returned			Too cool in returned			Exact (± 0.14)	Too warm in returned			Too cool in returned		
			I	U	MW	I	U	MW		I	U	MW	I	U	MW
SWW2	Released	166 (187)	-	-	-	-	-	-	95 (120)	-	-	-	-	-	-
	Climatol-D	170 (196)	1	9	1	4	10	1	122 (176)	6	11	0	11	18	0
	Climatol-M	157 (181)	0	8	2	9	11	11	95 (144)	14	10	8	14	14	18
	MASH	145 (181)	2	3	4	8	10	14	111 (171)	12	5	3	1	10	6
	ACMANT2	156 (184)	1	5	4	9	5	14	105 (164)	6	5	0	21	9	17
	DAP1	169 (192)	0	12	0	4	13	1	95 (128)	9	30	1	5	48	1
	HOM1	166 (191)	0	12	0	5	13	1	93 (127)	8	31	2	5	48	1
	SPLIDHOM1	168 (194)	0	12	0	2	13	1	94 (127)	10	30	1	5	48	1
SWW3	Released	146 (160)	-	-	-	-	-	-	94 (117)	-	-	-	-	-	-
	Climatol-D	156 (180)	5	16	3	5	13	0	130 (181)	16	11	1	5	8	0
	Climatol-M	152 (171)	8	15	2	7	16	3	111 (162)	20	9	6	7	15	3
	MASH	132 (170)	12	8	7	6	7	12	102 (170)	14	4	8	13	9	4
	ACMANT2	147 (179)	4	9	5	7	9	9	111 (172)	10	3	6	7	12	5
	DAP1	146 (162)	5	27	1	1	26	0	98 (128)	6	35	2	7	43	1
	HOM1	147 (163)	3	27	1	1	27	0	98 (129)	6	34	2	6	45	0
	SPLIDHOM1	146 (160)	6	27	1	2	26	0	98 (124)	10	34	2	7	44	1

Table B.55. As in table 7, but for the South West scenarios 2 and 3.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
SWW2	143 (143)	54 (58)	388	689	0.269	0.073	0.366	0.243 (0.243)	45.7%	16.8%	85.5%	46.9%	0.70%	20.4%	38.3%	16.4%
	107 (107)	176 (195)	424	566	0.202	0.237	0.555	0.147 (0.147)	37.6%	10.7%	56.5%	37.7%	0.70%	20.4%	23.4%	14.1%
	108 (108)	263 (313)	423	476	0.203	0.356	0.778	0.128 (0.128)	29.6%	15.4%	42.0%	38.3%	4.18%	16.6%	24.3%	18.8%
	32 (32)	113 (116)	499	630	0.060	0.152	0.272	0.049 (0.049)	11.3%	3.19%	18.8%	8.57%	1.39%	5.52%	7.66%	3.91%
SWW3	176 (176)	44 (56)	342	692	0.340	0.060	0.449	0.307 (0.307)	62.1%	15.9%	89.1%	50.6%	1.61%	28.8%	38.1%	NA
	141 (141)	144 (149)	377	592	0.272	0.196	0.560	0.211 (0.211)	51.2%	11.7%	66.3%	41.0%	2.82%	23.6%	30.1%	NA
	102 (102)	283 (327)	416	450	0.197	0.386	0.834	0.121 (0.121)	35.0%	9.84%	30.4%	32.0%	6.85%	18.3%	20.8%	NA
	25 (25)	124 (131)	493	612	0.048	0.168	0.301	0.039 (0.039)	8.37%	2.54%	17.4%	5.06%	0%	3.06%	6.23%	NA

Table B.56. As in table 8, but for the South West scenarios 2 and 3.

Scenario	Hits	FAs	Misses	CRs	HR	FAR	Freq. bias	Critical Success Index	Prop. CO IHs found	Prop. EV IHs found	Prop. large IHs found	Prop. medium IHs found	Prop. small IHs found	Prop. SCs found	Prop. SRs found	Prop. urbani-sation IHs found
SWW2	168 (168)	33 (36)	363	695	0.316	0.045	0.366	0.296 (0.296)	51.1%	21.2%	94.2%	54.9%	2.44%	24.3%	43.2%	21.9%
	150 (150)	138 (154)	381	588	0.282	0.190	0.555	0.219 (0.219)	46.2%	18.6%	69.6%	52.6%	3.48%	24.3%	33.8%	24.2%
	215 (215)	171 (204)	316	547	0.405	0.238	0.778	0.293 (0.293)	58.6%	30.7%	84.1%	71.4%	11.1%	33.7%	48.6%	35.9%
	72 (72)	76 (77)	459	652	0.136	0.104	0.272	0.118 (0.118)	24.2%	7.83%	42.0%	21.7%	1.74%	11.0%	18.5%	8.59%
SWW3	198 (198)	22 (33)	320	708	0.382	0.030	0.448	0.359 (0.359)	70.4%	17.5%	94.6%	59.6%	2.02%	34.1%	41.5%	NA
	196 (196)	90 (94)	322	640	0.378	0.123	0.560	0.320 (0.320)	69.0%	17.8%	78.3%	60.1%	6.85%	34.5%	44.7%	NA
	236 (236)	165 (192)	282	561	0.456	0.227	0.834	0.332 (0.332)	76.8%	25.4%	78.3%	73.0%	13.7%	39.3%	50.5%	NA
	71 (71)	83 (85)	447	648	0.137	0.114	0.301	0.118 (0.118)	25.1%	6.35%	41.3%	18.0%	0.40%	10.5%	16.3%	NA

Bibliography

Ahrens, C. D. (2000). *Meteorology Today: An Introduction to Weather, Climate and the Environment*. Brooks/Cole, Thomson Learning, sixth edition.

Akima, H. (1978). A method of bivariate interpolation and smooth surface fitting for irregularly distributed data points. *ACM Transactions on Mathematical Software*, 4:148–164.

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *Journal of Climatology*, 6:661–675.

Allen, R., editor (2003). *Penguin English Dictionary*. Penguin Books, second edition.

Auchmann, R. and Bronnimann, S. (2012). A physics-based correction model for homogenising sub-daily temperature series. *Journal of Geophysical Research*, 117:1–13.

Barnes, L. R., Schultz, D. M., Gruntfest, E. C., Hayden, M. H., and Benight, C. C. (2009). Corrigendum: False alarm rate or false alarm ratio? *Weather and Forecasting*, 24:1452–1454.

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*, 111:1–21.

Caussinus, H. and Lyazhri, F. (1997). Choosing a linear model with a random number of change-points and outliers. *Annals of the Insitute 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 - C - Applied Statistics*, 53:405–425.

Compo, G. P., Sardeshmukh, P. D., Whitaker, J. S., Brohan, P., Jones, P. D., and McColl, C. (2013). Independent confirmation of global land warming without the use of station temperatures. *Geophysical Research Letters*, 40:3170–3174.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, Jr., B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Br¨onnimann, S., Brunet, M., Crouthame, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, O., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and J.Worleyu, S. (2011). The twentieth century reanalysis project. *Quarterly Journal of the Royal Meteorological Society*, 137:1–28.

Conrad, V. (1946). *Methods in Climatology*. Harvard University Press, second edition.

Corcoran, W. T. and Johnson, E. (2005). North america, climate of. In Oliver, J. E., editor, *Encyclopedia of World Climatology*, pages 525–534. Springer, 1 edition.

Costa, A. C. and Soares, A. (2009). Homogenisation of climate data: Review and new perspectives using geostatistics. *Mathematical Geosciences*, 41:291–305.

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. John Wiley and Sons Inc.

DeGaetano, A. T. (2006). Attributes of several methods for detecting discontinuities in mean temperature series. *Journal of Climate*, 19:838–853.

Della-Marta, P. M. and Wanner, H. (2006). A method for homogenising the extremes and mean of daily temperature measurements. *Journal of Climate*, 19:4179–4197.

Deque, M. (2007). Frequency of precipitation and temperature extremes over france in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global and Planetary Change*, 57:16–26.

Domonkos, P. (2008a). Quantifying efficiency of homogenisation methods. In *Proceedings of the Sixth Seminar for Homogenisation and Quality Control in Climatological Databases*, pages 20–32.

Domonkos, P. (2008b). Testing of homogenisation methods purposes, tools, and problems of implementation. In *Proceedings of the 5th Seminar for Homogenisation and Quality Control in Climatological Databases*, pages 131–148.

Domonkos, P. (2011). Efficiency evaluation for detecting inhomogeneities by objective homogenisation methods. *Theoretical and Applied Climatology*, 105:455–467.

Domonkos, P. (2013). Measuring performances of homogenisation methods. *Quarterly Journal of the Hungarian Meteorological Service*, 117:91–112.

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*, pages 46–72.

Ducre-Robitaille, J., Vincent, L., and Boulet, G. (2003). Comparison of techniques for detection of discontinuities in temperature series. *International Journal of Climatology*, 23:1087–1101.

Durre, I., Menne, M. J., Gleason, B. E., Houston, T. G., and Vose, R. S. (2010). Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49:1615–1633.

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.

Ferguson, C. R. and Villarini, G. (2012). Detecting inhomogeneities in the twentieth century reanalysis over the central united states. *Journal of Geophysical Research*, 117:1–11.

Gsech, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., and Tyler, D. (2002). The national elevation dataset: Photogrammetric engineering and remote sensing. *Journal of the American Society for Photogrammetry and Remote Sensing*, 68:5–11.

Gsech, D. B. (2007). The national elevation dataset. In Maune, D., editor, *Digital Elevation Model Technologies and Applications: The DEM Users Manual*, pages 99–118. Bethesda, Maryland, American Society for Photogrammetry and Remote Sensing, 2 edition.

Harrison, R. G. (2010). Natural ventilation effects on temperatures within stevenson screens. *Quarterly Journal of the Royal Meteorological Society*, 136:253–259.

Hausfather, Z., Menne, M. J., Williams, C. N., Masters, T., Broberg, R., and Jones, D. (2013). Quantifying the effect of urbanisation on u.s. historical climatology network temperature records. *Journal of Geophyscial Research: Atmospheres*, 118:481–494.

Hogan, R. J. and Mason, I. B. (2012). Deterministic forecasts of binary events. In Jolliffe, I. T. and Stephenson, D. B., editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, chapter 3, pages 31–59. Wiley-Blackwell, 2 edition.

Hubbard, K. G. and Lin, X. (2006). Reexamination of instrument change effects in the us historical climatology network. *Geophysical Research Letters*, 33.

IPCC (2014). Climate change 2014: Synthesis report.contribution of working groups i, ii and iii to the fifth assessment report of the intergovernmental panel on climate change. Technical report, IPCC, Geneva, Switzerland. Core writing team: R.K. Pachauri and L.A. Meyer [eds], 151 pp.

Karl, T. R., Arguez, A., Huang, B., Lawrimore, J. H., McMahon, J. R., Menne, M. J., Peterson, T. C., Vose, R. S., and Zhang, H. (2015). Possible artefacts of data biases in the recent global surface warming hiatus. *Science*, 348:1469–1472.

Karl, T. R. and Williams JR., C. N. (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities. *Journal of Climate and Applied Meteorology*, 26:1744–1763.

Lopardo, G., F.Bertiglia, Curci, S., Roggero, G., and Merlone, A. (2014). Comparative analysis of the influence of solar radiation screen ageing on temperature measurements by means of weather stations. *International Journal of Climatology*, 34:1297–1310.

Lund, R., Wang, X. L., Li, Q., Reeves, J., Gallagher, C., and Feng, Y. (2007). Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, 20:5178–5190.

Marshall, J. and Plumb, R. A. (2008). *Atmosphere, Ocean and Climate Dynamics - An Introductory Text*. Elsevier Academic Press.

McCarthy, M. P., Titchner, H. A., Thorne, P. W., Tett, S. F. B., Haimberger, L., and Parker, D. E. (2008). Assessing bias and uncertainty in the Had-AT-adjusted radiosonde climate record. *Journal of Climate*, 21:817–832.

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. (2012a). Global historical climatology network - daily (ghcn-daily), version 3.00. Database - version 3.00 `http://doi.org/10.7289/V5D21VHZ` [Access date: 19-11-2012] - NOAA National Climatic Data Center.

Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G. (2012b). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology*, 29:897–910.

Menne, M. J. and Williams JR., C. N. (2005). Detection of undocumented changepoints using multiple test statistics and composite reference series. *Journal of Climate*, 18:4271–4286.

Menne, M. J. and Williams JR., C. N. (2008). Homogenization of temperature series via pairwise comparisons. *Journal of Climate*, 22:1700–1717.

Menne, M. J. and Williams JR., C. N. (2009). Homogenisation of temperature series via pairwise comparisons. *Journal of Climate*, 22:1700–1717.

Menne, M. J., Williams JR., C. N., and Vose, R. S. (2009). The u.s. historical climatology network monthly temperature data, version 2. *Bulletin of the American Meteorological Society*, July: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.

NASA (1991). In Starr, D. O. and Melfi, S., editors, *The role of water vapor in climate: A stategic research plan for the proposed GEWEX water vapor project (GVaP)*, number NASA Conference Publication 3120.

NOAA (2014). Twentieth century reanalysis. Database - version 2 - Provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA from their web site at `http://www.esrl.noaa.gov/psd/` [Access dates: June 2013 - January 2014].

Parker, D. E. (1994). Effects of changing exposure of thermometers at land stations. *International Journal of Climatology*, 14:1–31.

Parker, D. E. (2010). Urban heat island effects on estimates of observed climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 1:123–133.

Pepin, N. C. and Lundquist, J. D. (2008). Temperature trends at high elevations: Patterns across the globe. *Geophysical Research Letters*, 35:1–6.

Pepin, N. C. and Norris, J. R. (2005). An examination of the differences between surface and free-air temperature trend at high-elevation sites: Relationships with cloud cover, snow cover and wind. *Journal of Geophysical Research*, 110:1–19.

Pepin, N. C. and Siedel, D. J. (2005). A global comparison of surface and free-air temperatures at high elevations. *Journal of Geophysical Research*, 110:1–15.

Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E. J., Hassen-Bauer, I., Alexandersson, H., Jones, P., and Parker, D. (1998). Homogeneity adjustments of in situ atmospheric climate data: A review. *International Journal of Climatology*, 18:1493–1517.

Potter, K. W. (1981). Illustration of a new test for detecting a shift in mean precipitation series. *Monhtly weather review*, 109:2040–2045.

Quayle, R. G., Easterling, D. R., Karl, T. R., and Hughes, P. Y. (1991). Effects of recent thermometer changes in the cooperative station network. *Bulletin of the American Meteorological Society*, 72:1718–1723.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., and Rowell, D. P. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research*, 108.

Reeves, J., Chen, J., Wang, X., Lund, R., and Lu, Q. (2007). A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46:900–915.

Ren, G. and Zhou, Y. (2014). Urbanisation effect on trends of extreme temperature indices of national stations over mainland china, 1961-2008. *Journal of Climate*, 27:2340–2360.

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.

Ruckstuhl, C., Philipona, R., Morland, J., and Ohmura, A. (2007). Observed relationship between surface specific humidity, integrated water vapour, and longwave downward radiation at different altitudes. *Journal of Geophysical Research*, 112:1–7.

Shein, K. A., Todey, D. P., Akyuz, F. A., Angel, J. R., Kearns, T. M., and Zdrojewski, J. L. (2013). Revisiting the statewide climate extremes for the united states: Evaluating existing extremes, archived data and new observations. *Bulletin of the American Meteorological Society*, March, 2013:393–402. Values were taken from the database discussed in this paper with can be found at `www.ncdc.noaa.gov/extremes/scec/`.

Stepanek, P. (2004). Homogenisation of air temperature series in the czech republic during a period of instrumental measurements. In *Proceedings of the 4th Seminar for Homogenisation and Quality Control in Climatological Databases*, pages 117–133.

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. *Idojaras, Quarterly Journal of the Hungarian Meteorological Service*, 117:123–141.

Szentimrey, T. (1999). Multiple analysis of series for homogenisation. In *Proceedings of the Second Seminar for Homogenisation of Surface Climatological Data*, pages 27–46.

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*, pages 116–125.

Thorne, P. W., Brohan, P., Titchner, H. A., McCarthy, M. P., Sherwood, S. C., Peterson, T. C., Haimberger, L., Parker, D. E., Tett, S. F. B., Santer, B. D., Fereday, D. R., and Kennedy, J. J. (2011a). A quantification of uncertainties in historical tropical tropospheric temperature trends from radiosondes. *Journal of Geophysical Research*, 116:1–19.

Thorne, P. W., Willett, K., Allan, R. J., Bojinski, S., Christy, J. R., Fox, N., Gilbert, S., Jolliffe, I., Kennedy, J. J., Kent, E., Klein Tank, A., Lawrimore, J., Parker, D. E., Rayner, N., Simmons, A., Song, L., Stott, P. A., and Trewin, B. (2011b). Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bulletein of the American Meteorological Society*, November:40–47.

Titchner, H. A., Thorne, P. W., McCarthy, M. P., Tett, S. F. B., Haimberger, L., and Parker, D. E. (2009). Critically reassessing tropospheric temperature trends from radiosondes using realistic validation experiments. *Journal of Climate*, 22:465–485.

Trenberth, K. E. (1983). What are the seasons? *Bulletein of the American Meteorological Society*, 64:1276–1282.

Trenberth, K. E., Fasullo, J., and Smith, L. (2005). Trends and variability in column-integrated atmospheric water vapour. *Climate Dynamics*, 24:741–758.

Trewin, B. (2010). Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdisciplinary Reviews: Climate Change*, 1:490–506.

Trewin, B. (2013). A daily homogenized temperature dataset for australia. *International Journal of Climatology*, 33:1510–1529.

Venema, V. K. C., Bachner, S., Rust, H. W., and Simmer, C. (2006). Statistical characteristics of surrogate data based on geophysical measurements. *Nonlinear Processes in Geophysics*, 13:449–466.

Venema, V. K. C., Mestre, O., Aguilar, E., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., D.Rasol, 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.

Venema, V. K. C., Mestre, O., Aguilar, E., Guijarro, J. A., Domonkos, P., Vertacnik, G., Szentimrey, T., Stepanek, P., Zahradnicek, P., Viarre, J., Müller-Westermeier, G., Lakatos, M., Williams, C. N., Menne, M. J., Lindau, R., D.Rasol, 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. (2011). Description of the cost-home monthly benchmark dataset and the submitted homogenised contributions. Technical report, Metrological institute of the University of Bonn.

von Storch, H. and Zwiers, F. W. (2001). *Statistical Analysis in Climate Research*. Cambridge University Press.

Wang, F., Liu, Z., and Notaro, M. (2013). Extracting the dominant sst modes impacting north america's observed climate. *Journal of Climate*, 26:5434–5452.

Warrens, M. J. (2008). *Similarity coefficients for binary data*. PhD thesis, Leiden University, Netherlands.

Willett, K., Williams, C., Jolliffe, I. T., Lund, R., Alexander, L. V., Brönnimann, S., Vincent, L. 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.

WMO (1989). *Calculation of Monthly and Annual 30-year standard normals*, wcdp-no. 10, wmo-td/no. 341 edition.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Taylor and Francis Group.

Xu, W., Li, Q., Wang, X. L., Yang, S., Cao, L., and Feng, Y. (2013). Homogenisation of chinese daily surface air temperatures and analysis of trends in the extreme temperature indices. *Journal of Geophysical Research and Atmospheres*, 118:9708–9720.

Yazici, C., Yozgatligil, C., and Batmaz, I. (2012). A simulation study on the performances of homogeneity tests applied in meteorological studies. In ICAMC: International Conference on Applied and Computational Mathematics, Ankara, Turkey, Book of Abstracts.

Young, K. C. (1993). Detecting and removing inhomogeneities from long-term monthly sea level pressure time series. *Journal of Climate*, 6:1205–1220.

Yozgatligil, C., Purutcuoglu, V., Yazici, C., and Batmaz, I. (2011). Validity of homogeneity tests for meteorological time series data: A simulation study. In *Proceedings of the 58th World Statistics Congress (ISI2011)*, Dublin, Ireland.

Yozgatligil, C. and Yazici, C. (2015). Comparison of homogeneity tests for temperature using a simulation study. *International Journal of Climatology*, -:–.

Zhao, W. and Khalil, M. A. K. (1993). The relationship between precipitation and temperature over the contiguous united states. *Journal of Climate*, 6:1232–1236.